Artificial Intelligence 47 (1991) 251–288
Elsevier

The owl and the electric encyclopedia*

Brian Cantwell Smith

Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA; and Center for the Study of Language and Information, Stanford University, Stanford, CA 94305, USA

Received January 1989
Revised February 1990

Abstract

Smith, B.C., The owl and the electric encyclopedia, Artificial Intelligence 47 (1991) 251–288.

A review of "On the thresholds of knowledge", by D.B. Lenat and E.A. Feigenbaum.

1. Introduction

At the 1978 meeting of the Society for Philosophy and Psychology,[1] somewhat to the audience's alarm, Zenon Pylyshyn introduced Terry Winograd by claiming that his pioneering work on natural language processing had represented a "breakthrough in enthusiasm". Since those heady days, AI's hubris has largely passed. Winograd himself has radically scaled back his estimate of the field's potential (see, in particular [70, 72]), and most other practitioners are at least more sober in their expectations. But not to worry. Unbridled enthusiasm is alive and well, living in points South and West.[2]

* Thanks to David Kirsh, Ron Chrisley, and an anonymous reviewer for helpful comments on an earlier draft, and to Randy Davis for slowing down its original presentation.

[1] Tufts University, Medford, MA.

[2] Or at least it is alive. The original version of Lenat and Feigenbaum's paper (the one presented at the Foundations of AI conference, in response to which this review was initially written) was considerably more optimistic than the revision published here some four years later. For one thing, their estimate of the project's scale has grown: whereas in 1987 they suggested the number of things we know to be "many hundreds of thousands—perhaps a few million", that estimate has now increased to "many millions (perhaps a few hundred million)". In addition, whereas their original paper suggested that inference was essentially a non-problem (a sentiment still discernible in their "Knowledge Is All There Is Hypothesis", p. 192), the project is now claimed to incorporate at least "two dozen separate inference engines", with more on the way. Again, not

Enthusiasm takes many forms, even in AI. Most common is the belief that a simple mechanism can accomplish extraordinary feats, if only given enough of some resource (time, information, experience, computing power). Connectionist networks are a current favourite, but the tradition is time-honoured. Feedback circuits, theorem provers, production systems, procedural representations, meta-level architectures—all have had their day. In their present paper, Lenat and Feigenbaum take up the enthusiast's cause, defending a new flavour of "great expectation". They suggest that just a million frames, massaged by already-understood control structures, could intelligently manifest the sum total of human knowledge.

The paper exhibits another kind of zeal as well—more general than precipitate faith in mechanism, and ultimately more damaging. This time the fervour is methodological: an assumption that you can move directly from broad intuition to detailed proposal, with essentially no need for intermediate conceptual results. Let's look at this one first.

General insights, even profound ones, often have the superficial air of the obvious. Suppose Newton, in an attempt to strike up a conversation at a seventeenth century Cambridge pub, opened with the line that he had made an astonishing discovery: that it takes energy to do work. It is hard to believe the remark would have won him an extra pint. Newton is famous not for enunciating glib doctrines, but for elaborating a comprehensive system of details reaching from those encompassing insights all the way through to

---

only has the sophistication of their representation scheme increased, but (as predicted here in Section 3) their representational conventions have developed from those of a simple frame system towards something much more like full predicate calculus, complete with propositions, constraints, set-theoretic models, etc. (Their words: "the need for more formality, for a more principled representation language" was one of the "surprises that actually trying to build this immense KB has engendered".) All these signs of increased sobriety are reassuring, of course, although, given their ambition and eclecticism, one wonders whether the resulting complexity will be manageable.

More seriously, a conceptual shift has overtaken the project—more ramifying than these relatively simpler issues of scale. At the 1988 CYC review meeting (in Palo Alto), Lenat claimed that whereas he and Feigenbaum had initially taken their project as one of coding up everything in the encyclopedia (hence the name "CYC"), they were now convinced that the real task was to write down the *complement* of the encyclopedia: everything we know, but have never needed to say. This is an astounding reversal. Dreyfus should feel vindicated [22], since this shift in focus certainly strengthens any doubts about the ultimate adequacy of an allegiance to explicit representation.

For all that, their optimism remains intact. They still believe that by 1994 they will approach the crossover point where a system will pass the point of needing any further design or hands-on implementation, and will from then on improve simply by reading and asking questions (implying, I suppose, that AI's theoretical preliminaries will be concluded). Furthermore, they suggest that this second "language-based learning" stage will in turn end by about the end of the decade, at which point we will have a system "with human-level breadth and depth of knowledge". They claim these things, furthermore, in spite of such telling admissions as the following, written in 1989: "much of the 1984–89 work on CYC has been to get an adequate global ontology; i.e., has been worrying about ways to represent knowledge; most of the 1990–94 work will be actually representing knowledge, entering it into CYC."

precise differential equations. It is this intermediating conceptual structure that rescues his original insight from fatuity.

Lenat and Feigenbaum (L&F) announce their own impressive generalizations: the Knowledge Principle, the Breadth Hypothesis, the Empirical Inquiry Hypothesis, etc. Each, in its own way, makes sense: that competence in a domain arises because of specific knowledge of the constitutive subject matter; that "intelligent performance often requires the problem solver to fall back on increasingly general knowledge, and/or to analogize to specific knowledge from far-flung domains"; etc. I agree; I expect most readers would agree—and so, I'd wager, would Newton's drinking partners. The problem is that L&F, with only the briefest of intervening discussion, then arrive at radically concrete claims, such as that three decades will suffice to carry out the following sweeping three-stage research program: (i) the slow hand-coding of a frame-based knowledge base, approximating "the full breadth of human knowledge" ($50 million, due to be completed by 1994), sufficient to bring the system to a point (ii) where it will be able to read and assimilate the remaining material on its own (approximately the turn of the century), followed by a stage (iii) where it is forced to carry out its own program of research and discovery, since it will have advanced "beyond the frontier of human knowledge".

One is reminded of tunnel diodes. For a moment the argument is on the plane of common sense, and then—presto!—it is suddenly at an extreme level of specificity, without ever having been anywhere in between. From the generality of human knowledge to the intricacies of slot inheritance; from the full flowering of intelligence to particular kinds of controlled search—leaps like these are taken without warning, often mid-sentence. The problem is not simply that the reader may disagree with the conclusions, but that there is no hint of the complex intellectual issues and decades of debate that lie in the middle. I.e., whereas tunneling electrons—or so we're told—genuinely switch from one place to another without ever being half-way in between, arguments don't have this luxury. Truth and reason are classical, so far as we know, constrained to follow continuous trajectories. That's why the middle ground of conceptual analysis and carefully laid-out details is the stuff and substance of AI.

So: After giving a better sense (in the next section) of the sort of argument that's missing, I will take it as the task of this review to map out at least some of the intermediate conceptual territory. The immediate goal will be to figure out what view of its structure could have led L&F to tunnel through in the way they did. As for their conclusions, I've already suggested I find them implausible, but others will want to judge for themselves. My larger aim is to convince the reader that any serious assessment of L&F's paper (or indeed of any analogous proposal) must be made against the backdrop of that hidden middle realm.

2. Conceptual tunneling

L&F start with the Knowledge Principle, cited above: that you have to know specific things about a domain to be competent at it. This insight is then used to discriminate a set of levels of expertise: rudimentary, middle-level practitioner, and expert. These levels are introduced with tautological generalization: to get started, you need to know something; the more you know, the less you need to search; once you know enough, additional knowledge will only infrequently (though still occasionally) be useful. Little more is said, unfortunately. And if the text is read closely, it shifts from the banal to the false.

Take the middle "practitioner" level. Without comment, L&F claim that "today's expert systems . . . include enough knowledge to reach the level of a typical practitioner performing the task." This claim may be true in a few limited, carefully chosen domains. In the sweeping context of the paper, on the other hand, the remark implies something different: that moderate expertise is achievable in arbitrary (if still specific) arenas. The latter claim simply isn't true; we don't yet have expert system personnel managers, nurses, or private detectives, and there are many, including some of the technology's protagonists (see, e.g., [16]), who suspect we never will. So the reader ends up caught between the plausibility of the narrow reading and the presumption of the broad one.

Similarly, consider L&F's comments about getting started. They claim that to solve a problem you need a minimum amount of knowledge in order to "state [it] in a well-formed fashion". This is a major assumption, again debatable. As students of AI are increasingly realizing (see [1, 2, 13, 21, 24, 39, 44, 48, 57–59, 67, 72] for a variety of such views), there's no reason to believe that people formulate anything like all the problems they solve, even internally.[3] Children happily charge around the world long before they acquire any conceptual apparatus (such as the notions of "route" and "destination") with which to formulate navigational problems. So too with language: fluent discourse is regularly conducted in complete absence of a single linguistic concept—including "word" or "sentence", let alone Bosworth's "prose" or the logician's "substitution *salve veritate*". Similarly, when you reach around and retrieve your coffee cup from the side table, there is no reason—especially no a priori reason—to believe that you formulate much of anything at all. Problems stated in words have to be formulated, yes; but only because to "formulate" means to state in words.

Here we see the beginning of the tunnel. If (i), in order to sidestep issues of explicit formulation, and to avoid foundering in simplistic cases, the minimalist threshold were generalized to "the solution of any complex task requires some

[3] Suchman [67], for example, argues that conceptualizing action is often a retrospective practice—useful for a variety of purposes (such as explanation), but not implicated in engendering the action in the first place, especially in routine or everyday cases.

minimum amount of knowledge"; and (ii) the notion of "knowledge", which L&F never really explain, were generalized to include perception, motor coordination, tacit expertise, explicit conceptual powers, and all the rest—then, well, yes, we would have a more tenable reading. The problem is, we would also have a vacuous reading: no one could rationally imagine anything else. On the other hand, if instead we try to put some meat on the skeletal insights, and prohibit wanton generalization, it becomes unclear how to hang on to the original intuition without running counter to fact.[4]

Such worries don't deflect these authors, however. Without breaking stride, they claim that the Knowledge Principle is "a mandate for humanity to concretize the knowledge used in solving hard problems in various fields." Three lines later this has turned into a mandate to "spend the resources necessary to construct one immense knowledge base spanning human consensus reality". But why? Even the untenably "formulated" readings of these putative principles aren't in themselves mandates to *do* anything at all. The underlying (tunneled) argument must include something like the following presumptions: we know how to write "knowledge" down (i.e., the knowledge representation problem will imminently be solved); there won't be any interaction effects; we can ride rough-shod over all ontological problems about how people conceptualize the world;[5] and so on and so forth.

What of the other principles? At the level of grand generality, the Breadth Principle is again something that no one could plausibly deny. It recommends the use of generalization and analogy when more specific things fail. Consider just analogy. Is it important? Undoubtedly. Understood? It's unlikely that its full-time students would say so.[6] Does L&F's paper illuminate its subtleties? Very little. All that is presented are a few paragraphs barely hinting at the issues involved. Take for example the postulated Analogical Method: "if *A* and *B* appear to have some unexplained similarities, then it's worth your time to hunt for additional shared properties." But it is well known that there are just too many properties, too many similarities, to be relevant. Thomas Jefferson

[4] For example, consider one possible defense: (a) that L&F are implicitly assuming intellectual competence can be separated into two categories—one relatively tacit, perceptually or experientially grounded, less dependent on explicit formulation; the other, a kind of higher-level, fully conceptual, "expertise", relying on careful articulation; and (b) that a system manifesting the second can be constructed without any roots in the first. If this is their position, it is very, very strong—needing not just admission but defense. At a minimum, they would have to argue at least two things (in opposition to Dreyfus [21], Suchman [67], Winograd [70, 72], and others): (a) that the following three distinctions align (or at least coincide on the right): amateur versus expert, tacit versus articulated, and perceptual versus cognitive; and (b) that common sense, by their own admission a necessary ingredient in expert reasoning, *can be captured solely in "knowledge" of the second kind*. But of course no such argument is forthcoming.

[5] See, e.g., Bobrow [9], Hayes [34, 35], Hobbs and Moore [36], Hobbs et al. [37], and Levy et al. [46]. It's not so much that L&F think that ontology is *already* solved, as that they propose, in a relatively modest time-period, to accomplish what others spend lives on.

[6] See for example Gentner and Gentner [29], and—to the extent that analogy ties in with metaphor—the papers in Ortony [52].

and John Adams both died (within an hour of each other) on July 4, 1826—50 years to the day after the signing of the Declaration of Independence they co-authored. It's rumoured that the price of bananas and the suicide rate in France tracked each other almost perfectly for years. The words "abstemious" and "facetious" exhibit all five vowels in alphabetic order. Do we have an explanation for these facts? No. So, should we look for additional similarities? Probably not. A proper treatment of analogy requires a notion of *relevant* similarity. Nor can their suggestion of entering "specialized versions" of analogical reasoning in an *n*-dimensional matrix (according to "task domains, . . . user-modes, . . . , analogues with various epistemological statuses", etc.) be more than a data structural encoding of hope.

Furthermore, nothing in the paper clues the reader into the fact that these issues have been investigated for years. All we get are statements like this: "we already understand deduction, induction, analogy, specialization, generalization, etc., etc., well enough to have knowledge be our bottleneck, not control strategies." Breathtaking, but simplistic. And in a disingenuous sleight of hand, the passage continues: "On the other hand, all such strategies and methods are themselves just pieces of knowledge", with the implication that it should be straightforward to have them selected and applied at the meta-level. But this is simply not a serious argument. To start with, you can't have it both ways: either we do know enough about control structure, or we don't. And if we don't, then we're probably not ready to write it down, either. Furthermore, relying on universal meta-levels is like defending Von Neumann machines as cognitive models because they would exhibit intelligent behaviour, if only they were given the right programs. It isn't false, but it isn't useful, either.[7]

There's more. We are told that:

> . . . In a sense, natural language researchers have cracked the language understanding problem. But to produce a general Turing-testable system, they would have to provide more and more semantic information, and the program's semantic component would more and more resemble the immense [knowledge base] mandated by the Breadth Hypothesis.

This time we're given neither supporting details nor motivating intuition. On the unwarranted assumption that parsing is solved, and if by "semantic

[7] Actually, it might be false. Encoding control directions at the meta-level is another instance of L&F's unswerving allegiance to explicit formulation. Unfortunately, however, as has been clear at least since the days of Lewis Carroll, not *everything* can be represented explicitly; at some point a system must ground out on a non-represented control regimen. Now L&F are presumably relying on the computational conceit that any control structure whatsoever can be *implemented* explicitly, by representing it in a program to be run by another, non-represented, underlying control regimen. Proofs of such possibility, however, ignore resource bounds, real-time response, and the like. It is not clear that we should blithely assume that our conceit will still hold under these more restrictive constraints, especially in as pragmatic a setting as L&F imagine.

information" one includes *everything* else—pragmatic assumption, concept formation, inference, induction over experience, formations of judgment, theory change, discourse understanding, etc., coupled with everything that anyone could ever need to know or be in order to be a competent participant in a dialogue, including what L&F call "consensus reality"—then, well, yes, that's all we need to do.

The authors take that "consensus reality" seriously: it is intended to include the entire fabric of assumptions and common sense underlying all of human knowledge. One of the paper's most spectacular assertions is the claim that all people know can be captured in a million frames—a statement reinforced by citing three independent estimates, two based on sheer guesses of how many frames are needed to understand an article or word (guesses because we as yet have no real assurance that any computer has ever really understood a single word, let alone a sentence or longer text), another on an estimate of four entries into long-term memory per hour. No room is made for such commonplace phenomena as the recognition, many years later, of a face once glimpsed for just a few seconds—an ability still well beyond computational emulation. Or the empathetic stance necessary in order to understand allusions and insinuations in any piece of serious writing. Or even simple acts of speculation. Imagine, for example, a toboggan careening down an ice-clad winter hill, increasingly out of control, with the initial look of terrified glee steadily draining out of the face of the 13-year old at the helm, being replaced by an anguished expression of sheer panic. Now quick: how many "pieces" of knowledge did you just use in picturing this scene?

And so it goes. The paper accuses others of premature formalization, without even entertaining the thought that setting out to code up human knowledge in a million frames might be an instance of the very phenomenon. Empirical inquiry is endorsed, but seems only to involve the investigation of computer programs, not the phenomena they are being used to model (and even that seems confused: L&F claim we should use computers "as a tool", the way astronomers use telescopes, an injunction that I would have thought applied to physics but exactly not to AI[8]). The issues are so complex it is hard

[8] For astronomers, telescopes are *tools*, not *subject matters*; the theoretical notions in terms of which we understand telescopes aren't the constitutive notions in terms of which we understand *what is seen through telescopes*. AI, in contrast, is different: we exactly *do* claim that computational notions, such as formal symbol manipulation, *are* applicable to the emergent intelligence we computationally model.

Note in passing that although this reminiscent of Searle's [60] notions of *strong* and *weak* AI, there is a crucial difference. In making such distinctions, Searle is distinguishing the *relation* between a computational system and the mind: whether only their *surface behaviours* are claimed similar (weak), or whether the way in which the computational process works is claimed to be the way in which the mind works (strong). L&F, on the other hand, at least in this proposal, are making no psychological claims; hence Searle's terms, strictly speaking, don't apply (although L&F, if pressed, would presumably opt for the weak option). In contrast—and in complete independence of psychology—they propose to build a *computer system*, and computer systems

to tell what they think; at best they seem to have in mind what would normally be called *hypothesis testing*, not *empirical inquiry*. There's no admission that there are external data and practices to be studied—that ours isn't an entirely internalist, constructed game (they do say that "intelligence is still so poorly understood that Nature still holds most of the important surprises", but shortly thereafter dismiss all of deduction, induction, and so on as essentially solved). In a similar vein, it's striking that genuine semantics isn't even mentioned—not the question of "semantic representation" (i.e., how concepts and meanings and the like are stored in the head), but the tougher question of how symbols and representations relate to the world.

Alas, it looks as if what discouraged Winograd hasn't even been imagined by the present authors.

3. The structure of the middle realm

Perhaps someone will object. L&F march to the pragmatist's drum, after all. So is it unfair to hold them to clear theoretical standards? I think not. For one thing, in a volume on the foundations of AI, explicating premises should be the order of the day. Second, there is the matter of scale. This is a large project they propose—all of consensus reality, 50 million dollars for the first stage, etc. Untutored pragmatism loses force in the face of a task of this magnitude (you can bridge a creek without a theory, but you won't put a satellite into orbit around Neptune). Furthermore, citing the modesty of human accomplishment ("people aren't perfect at these things") won't let L&F off the hook, especially when what is particularly modest is people's understanding of their own intellectual prowess. Fortunately, we humans don't have to known much about reasoning to be good at it—cf. the discussion of formulation, above. But L&F can't piggy-back off our native competence, in creating a computational version. Given that they're both starting from scratch, and committed to an explicit-representation stance, they must understand what they're doing.

---

necessarily work in computational ways. I.e., they have to be "strong" about their own project: otherwise they would be in the odd position of having no idea how to go about developing it. And it is clear, in this sense, that they are "strong"; why else would they be discussing slots, frames, and meta-rules?

So what of empiricism? As L&F suggest (this is their primary brief), the computational models they recommend building should of course be *tested*. But as I suggest in the text, to claim that isn't to claim that computers are the paradigmatic object of *study*. On the contrary, I would have thought an appropriate "empirical" stance for computational AI would go something as follows: one would (a) study intelligent behaviour, independent of form (biological, artifactual, whatever), but known in advance (i.e., pre-theoretically) to be intelligent behaviour; (b) construct (strong) computational models that manifest the essential principles that are presumed or hypothesized to underlie that intelligence; and then (c) conduct experiments to determine those models' adequacy. The point is that it is the first stage, not the third, that would normally be called "empirical".

Table 1
A dozen foundational questions (Boxes indicate agreement).

	Logic	L&F	EC
1. Primary focus on explicit representation?	yes	yes	no
2. Contextual (situated) content?	no	no	yes
3. Meaning dependent on use?	no	no	yes
4. Consistency mandated?	yes	no	no
5. Single representational scheme?	yes	yes	no
6. Entirely discrete (no continuity, images, . . .)?	yes	yes	no
7. Representation captures all that matters?	yes	yes	no
8. Reasoning and inference central?	yes	yes	yes
9. Participation and action crucial?	no	no	yes
10. Physical embodiment important?	no	no	yes
11. Support for "original" semantics?	no	no	yes
12. Distinguish theorist's and agent's conceptual schemes?	no	no	yes

So we're brought right back to where we started: with that hidden middle realm. Let's dig deeper, therefore, and uncover some of its inner structure. I'll do this by locating L&F's position with respect to twelve foundational questions—questions that could be asked of any proposed reasoning or inference system. Given that we lack a general theory of representation (not only those of us in AI, but the wider intellectual community as well—a sobering fact, since our systems rest on it so fundamentally), posing such questions is as good an analytic strategy as any. Furthermore, these twelve will help reveal L&F's representational assumptions.

The answers are summarized in Table 1. To convey a better sense of the structure of the territory, I've flanked L&F's position with two other replies. On the left is the position of traditional formal logic (the system studied by philosophers and logicians, not "logic-based" theorem provers or logic programming languages—both too ill-defined to be of much help here). On the right is my own assessment of the minimum an AI system will require in order to achieve anything like genuine intelligence. For discussion, I'll call it a notion of "embedded computation" (EC).

One point needs emphasizing, before turning to specifics. Embedded computation is still an emerging perspective, not yet a technical proposal. That doesn't make it sheer speculation, however, nor is it purely idiosyncratic. A growing number of researchers are rallying around similar views—so many, in fact, that one wonders whether something like it won't be the next AI stage, beyond the "explicit knowledge" phase that L&F represent.[9] Nonetheless, I

[9] In part, but not solely, because of its potential compatibility with connectionism. For specfic discussion and results see, e.g., [1, 2, 12–15, 39, 48, 51, 55, 57–59, 66, 67, 72].

would be the first to admit that details remain to be worked out. But that's exactly my point. I'm contrasting it with L&F's position exactly in order to highlight how far I believe we are from achieving their stated goals. For purposes of the present argument, in other words, any claim that we don't yet understand some aspect of the speculative EC view—what nondiscrete computation would be like, say—counts *for* my position, and *against* L&F.[10] All that matters is that there is some reason to believe that the issue or phenomenon in question is at least partially constitutive of intelligence. L&F are the ones with the short-term timetable, after all, not I.

Question 1. Does the system focus primarily on explicit representation?

Logic	L&F	EC
yes	yes	no

In the current design of computer systems, there is probably no more time-worn a technique than that of "explicit representation". And there is no difficulty in discerning L&F's views on the subject, either. They line up directly with tradition. In fact that representation be explicit is the only specific coding requirement they lay down (it is mandated in their "Explicit Knowledge Principle"). Similarly, the CYC project takes explicit representation as its fundamental goal.

Unfortunately, however, it is unclear what L&F (or anyone else, for that matter) mean by this term—what, that is, the implicit/explicit distinction comes to (see [42] for a recent paper on the notion). This is not to say that the notion doesn't matter. Many programmers (and I count myself as one of them) would stubbornly insist that choices about explicit representation impinge on effectiveness, control flow, and overall system architecture. The question is what that insistence is about.

When pressed for examples of explicit representation, people typically point to such cases as the grammarian's "$S \rightarrow NP\ VP$", logical formulae such as "$P(a) \supset Q(b)$", frames in such systems as KRL, or nodes in semantic nets. The examples are almost always taken from language-like representational schemes, suggesting that some combination is required of conceptual categorization, recursive method of combination, and relative autonomy of representational element[11] (images and continuous representations are rarely, though not never, cited as paradigmatically explicit). Explicitness is also relational, hold-

[10] In fact, as it happens, it doesn't even matter whether you think the EC view *is computational at all*. What's at stake here are the requisite underpinnings for *intelligence*; it is a secondary issue as to whether those underpinnings can be computationally realized. As is happens, I believe that the (real) notion of computation is so much wider than L&F's construal that I don't take the discrepancy between genuine intelligence and their proposal as arguing against the very possibility of a computational reconstruction. But that's a secondary point.

[11] "Explicit" fragments of a representational scheme are usually the sort of thing one can imagine removing—surgically, as it were—without disturbing the structural integrity or representational content of the remainder.

ing between something (a representation) and something else (what it represents). This provides some freedom: a given structure can be implicit, explicit, neither (if, like a bread-basket, it doesn't represent anything), or both (if it represents severally). Logical axioms, for example, are often viewed as explicit representations of their own particular contents, but (in ways that Levesque [45], Halpern [26], and others have tried to make precise) as implicit representations of what they imply.

So what does explicitness come to? Though it's currently impossible to say, it seems to require a roughly determinate object (sentence, frame, whatever), of a readily discriminable type, that simultaneously plays two rather direct roles: straightforwardly representing some content or other (John's having brown hair, say), and, again quite directly, playing a causal role in the course of the system's life that constitutes that system's knowing or believing the corresponding content (or would at least lead an observer to say that the system knows or believes it).[12] I.e., explicitness seems to require (a) a degree of modularity or autonomy, (b) a coincidence of semantic and causal role, and (c) a relative directness or immediacy of the latter.

In contrast, people would label as *implicit* the representation of the letter "A" in a run-length encoded bitmap representation of a page of text, or the representation of the approach velocity of an oncoming car in the frequency difference between the outgoing and incoming radar signals in a police speed trap, or (as suggested above) the representation of a fact by the explicit representation of a different set of facts, when the first is a distant entailment of the latter set. In each case the representational element is either itself relationally encoded, or else one of its two "consequent" relations, instead of being direct, is in turn complex and relational; between the structure and its content, or between the structure and the inferential role relevant to that content.

Assuming this reconstruction points in even roughly the right direction, let's go back to L&F. To start with, it makes sense of why L&F contrast explicit with "compiled" representations (since compilation often removes the structural autonomy of distinct source elements), and of their assumption that facts can be represented in relative independence: simple content and simple causal consequence, neither depending much on what else is represented, or how anything else is used. As will become clearer in a moment, this theme of modularity, directness, and relative independence characterizes L&F's proposal at a variety of levels. (I'm prepared to argue that L&F's proposal won't work, but I'm not claiming it doesn't have a degree of integrity.)

What about the flanking views? At the level of whole sytems, formal logic is paradigmatically explicit (in spite of the "implicit" treatment of entailment mentioned above—what matters is that the explicit representations are the

[12] See the discussion of the "Knowledge Representation Hypothesis" in [62].

ones that are theoretically analyzed). If forced at theoretical gun-point to produce an "explicit representation" of the structure of Abelian groups, for example, it's hard to imagine a better place to start than with first-order axiomatization. And yet, in part as indicated by their repeated desire for a relatively minimal role for deduction and complex reasoning (see Question 8, below), L&F are even more committed to explicit representation than adherents of logic. That is to endorse a very serious amount of explicitness indeed.

The embedded view? It would be hard to argue that explicit representation isn't powerful, but, as discussions of the next questions will suggest, it carries a price of potentially unwarranted definiteness, premature categorization, and resistance to some sorts of recognition. My main dispute, however, isn't over its utility. Rather, I question whether, if explicit representation is indeed an identifiable subspecies (the only construal on which it could matter at all), it is the only sort that's required. That is something I wouldn't want to admit without a lot more evidence. In particular, I worry that a system comprised only of explicit representations would be fatally disconnected from the world its representations are about.[13]

Question 2. Is representational content contextual (situated)?

Logic	L&F	EC
no	no	yes

Under the general rubric of the term "situated" ("situated language" [8], "situated action" [67], "situated automata" [58]) a variety of people have recently argued that adequate theory cannot ignore the crucial role that context plays in determining the reference and semantic import of virtually all linguistic and other intentional phenomena. Context is obviously important in interpreting "now", "tomorrow", and "it's raining"; and in determining the temporal implications of tense. In its full glory, however, the situated claim goes much deeper: that you can't ultimately understand anything except as located in the circumstances in which it occurs. Linguistic evidence is impressive. In spite of the assumption that is sometimes made that proper names function essentially as logical constants, it's common sense that "Tom", "Dick" and "Harry" in fact refer to *whatever people in the appropriate context have those names*. Even

[13] Some of the reasons will emerge in discussions of later questions, and are argued in [65]. For analogous views, again see the exploratory systems of Rosenschein and Kaelbling [58], Brooks [12], and Chapman and Agre [13], and the writings of Suchman [67], Cussins [15], Dreyfus [21], and Smolensky [66].

L&F may of course reply that they do embrace implicit representation, in the form of compiled code, neural nets, unparsed images. But this isn't strictly fair. By "the L&F position" I don't mean the CYC system *per se*, in inevitably idiosyncratic detail, but rather the general organizing principles they propose, the foundational position they occupy, the theoretical contributions they make. I.e., it isn't sufficient to claim that the actual CYC software does involve this or that embedded aspect, as, in many cases, I believe it *must*, in order to work at all—see, e.g., footnotes 16 and 29. Rather, my plaint is with overarching intellectual stance.

"1989" isn't absolute; when it appears in the *New York Times*, it usually refers to the Gregorian calendar, not the Julian or Islamic one.

But language has no patent on contextual dependence. Computational examples are equally common. When you button "QUIT" on the Macintosh file menu, for example, the process that quits is *the one that is running*. The simple e-mail address "JOHN", without an appended "@HOST" suffix, identifies the account of whoever has that username *on the machine from which the original message is sent*. If I set the alarm to ring at 5:00 p.m., it will ring at 5:00 p.m. *today*. The machine language instruction "RETURN" returns control from the *current stack frame*. If you button "EJECT", it ejects the floppy *that is currently in the drive*.

Some quick comments on what contextual dependence isn't. First, none of the cited examples should be read as implying that terms like "now", proper names (or their internal analogues), machine instructions, and the like are ambiguous. There's no reason (other than a stubborn retention of prior theory) to treat the contextual dependence of reference as a matter of ambiguity. Second, though related, the present issue of contextuality cross-cuts the explicit/implicit distinction of question 1 ("here" and "now" are explicit representations of contextually determined states, for example, whereas QUIT and RETURN represent their contextually determined arguments implicitly, if at all). Third, as with many semantical phenomena, representations typically have (contextually dependent) contents; it's a category error to assume that those contents have to be computed. Fourth—and even more important—contents not only don't have to be, but typically can't be, determined solely by inspecting the surrounding *representational* context. In the "QUIT" case, for example, the process to be killed is instantiated on the machine, but that doesn't imply that it is represented. Similarly, in the e-mail case, the host machine plays a role in determining the relevant addressee, but the egocentricity obtains in virtue of the machine's existence, not in virtue of any self-reference. And in the use of Gregorian dates, or in the fact that "1:27 p.m." (on my word processor, today) refers to 1:27 p.m. Pacific Standard Time, not only is the relevant context not *represented* by the machine, *it is not a fact within the machine at all*, having instead to do with where and when the machine is located in the world.[14]

Here's a way to say it: the sum total of facts relevant to the semantical valuation of a system's representational structures (i.e., the relevant context) will always outstrip the sum total of facts that that system represents (i.e., its content).

[14] I am intentionally ignoring scads of important distinctions—for example, between the indexicality of representational content (of which "here" and "now" are paradigmatic exemplars), and the even more complex relation between what's in fact the case and how it's represented as being (the latter is more Suchman's [67] concern). Sorting any of these things out would take us far afield, but I hope just this much will show how rich a territory isn't explored by L&F's proposal.

What, then, of the three proposals under review? Traditional logic, again paradigmatically, ignores context.[15] The logical viewpoint, to use a phrase of Nagel's [50], embodies the historical imagination's closest approximation yet to a "view from nowhere". Contextual influence isn't completely gone, of course—it still plays a role in assigning properties and relations to predicates, for example, in selecting the "intended interpretation". But as far as possible logical theories ignore that ineliminable residue.

L&F are like the logicians; they ignore context too. *And they have to.* Context isn't a simple thing—something they don't happen to talk about much, but could add in, using their touted mechanism for coping with representational inadequacy: namely, adding another slot. On the contrary, their insistence that their "knowledge base" project can proceed without concern as to time, place, or even kind of use, is essentially an endorsement of a-contextual representation.

For my part (i.e., from the embedded perspective), I think the situated school is on to something. Something important. Even at its most objective, intelligence should be viewed as a "view from somewhere" [65]. Take an almost limiting case: suppose you were to ask L&F's system how many years it would be before the world's population reached 7 billion people? Without a contextual grounding for the present tense, it would have no way to answer, because it wouldn't know what time it was.[16]

Question 3. Does meaning depend on use?

Logic	L&F	EC
no	no	yes

This question gets at a much more radical claim than the last. The idea is not only that content or final interpretation of a representational structure (sentence, frame, whatever) depends on the situation in which it is used, but that what the structure means can't be separated from the whole complex of inferential, conversational, social, and other purposes to which it is put.[17]

[15] Except the limiting case of intrasentential linguistic context necessary to determine by which quantifier a variable is bound.

[16] L&F might reply by claiming they could easily add the "current date" to their system, and tie in arithmetic procedures to accommodate "within 10 years". My responses are three: (i) that to treat the particular case in this ad hoc way won't generalize; (ii) that this repair practice falls outside the very foundational assumptions on which the integrity of the rest of their representational project is founded; and (iii) that the problem it attempts to solve absolutely permeates the entire scope of human knowledge and intelligence.

[17] Careful distinctions between meaning and content aren't particularly common in AI, and I don't mean to use the terms technically here, but the situation-theoretic use is instructive: the *content* of a term or sentence is taken to be what a use of it refers to or is about (and may differ from use to use), whereas the *meaning* is taken, at least approximately, to be a function from context to content, and (therefore) to remain relatively constant. So the content of "I", if you use it, would be you; whereas it's meaning would (roughly) be λSPEAKER.SPEAKER. (This is approximate in part because no assumption is made in situation theory that the relationship is *functional*. See [5].)

It's one thing to say that the word "now", for example, or the state of an internal clock, refers to the time of its use; that doesn't bring purpose or function into the picture. But if you go on to say that the question of whether such a use refers to a particular recent event can't be determined except in light of the whole social pattern of activity in which it plays a role (which, as I'll admit in a moment, I believe), then, from the point of view of developing a (middle-realm) theory, you are taking on a much larger task.

To see, this, consider a series of examples. First, assume that the term "bank" is ambiguous, as between financial institutions and edges of rivers. Although neither L&F nor I have talked about ambiguity, that shouldn't be read as implying that it is trivial. Still, let's assume it can somehow be handled. Second, the word "today", as noted above, is also referentially plural—in the sense of being usable to refer to many different things, depending (typically) on the time of utterance. But "today" is indexical, not ambiguous (here's a discriminating rule of thumb: ambiguity, but not indexicality, leads to different dictionary entries[18]). As a consequence, its referential plurality (unlike that of a truly ambiguous term) can't be resolved at the parsing or internalization stage—so the indexicality will be inherited by the corresponding internal data structure. Third, and different from both, is Winograd's example of "water" [72, pp. 55–56], as used for example in the question "Is there any water in the refrigerator?". It is this last kind of example I mean to describe as having *use-dependent meaning*. In particular, depending on a whole variety of things, the word in context could mean any of a million things: Is there literally any H_2O present in the metal-contained volume (such as in the cells of the eggplant)? Is there any potable liquid? Has any condensation formed on the walls? . . . . The point is that there is no reason to suppose these variations in meaning could (or should) be systematically catalogued as properties of the *word* (as was suggested for the referent of "today"). Instead, Winograd suggests (and I agree) something more like this: the meaning of "water" is as much determined by the meaning of the discourse as the meaning of the discourse is determined by the meaning of "water".

Nothing in this view is incoherent, or even (at least necessarily) repellent to systematic analysis: imagine that semantical interpretation (including the non-effective semantical relations to the world) works in the cycle of a relaxation algorithm, influenced by a variety of forces, including the actual participatory involvement of the agent in the subject matter. Still, use-dependent meaning does pose problems for a theorist. Take just two examples. First, it undermines the very coherence of the notion of sound (or complete) inference; those concepts make sense only if the semantic values of representational formulae are conceptually independent of their role in reasoning. The problem isn't just

[18] Imagine the dictionary entry if "today" were taken to be ambiguous: . . . $\text{today}_{1,236,781}$: June 24, 1887; $\text{today}_{1,236,782}$: June 25, 1887; $\text{today}_{1,236,783}$: June 26, 1887; . . . !

that there is no obvious model-theoretic analysis, since it is unclear what model-theoretic structure would be assigned to the term "water". Or even, setting model theory aside, that it is unclear what a well-defined semantical value for such a term could be. More seriously, soundness is fundamentally a claim that the use of a term or predicate has respected its independently given semantical value. Making interpretation dependent on use, at least at first blush, therefore gives one every reason to suppose that the notion of soundness is rendered circular, hence vacuous.[19]

Second, it is a likely consequence of this view that the meaning or significance of a complex representational structure won't be able to be derived, systematically, from the "bottom up", but will instead have to be arrived at in some more holistic way. It challenges, in other words, the traditional view that semantics can be "compositionally" defined on top of a base set of atomic values.[20] I.e., the point isn't just that the interpretation of a sentence (its propositional value) is sometimes determined by mutually interlocking constraints established by various sentential constituents (as suggested in indexical cases, such as for the pronoun structure in "though Jim didn't like her, Mary was perfectly happy with him"), say by some sort of relaxation method. Rather, a deeper claim is being made: that the very meaning of the parts of a discourse can depend on the interpretation of the whole. For example, suppose the clouds clear, and you make a comment about the relentless sun. It is easy to imagine that I understand the meaning of "relentless"[21] in virtue of knowing what you're talking about, rather than the other way around. And if it is whole sentences that connect with situations, this may have to be done not bottom-up in terms of the representational constituents, but if anything top-down.

None of this suggests that representation, or interpretation, is impossible.

[19] See the discussion of *coordination conditions* in [65] for one suggestion as to how to retain the integrity of intentional analysis (better: integrity to the notion of intentionality) in the face of this radical a theoretical revision.

[20] To make this precise, you have to rule out cheats of encoding or implementation, of the following sort: Suppose there is some holistic regularity $\mathcal{H}$, a function of all kinds of contextual aspects $\mathcal{C}_i$, whereby complete intentional situations take on a meaning or significance $\mathcal{M}$, and suppose that $\mathcal{H}$ is in some way parameterized on the constituent words w_1, w_2, etc. (which of course it will be—on even the most situated account it still matters what words you use). By a kind of inverted currying process, this can be turned into a "bottom-up" analysis, based on a meaning of the form $\lambda\mathcal{C}_1, \mathcal{C}_2, \ldots, f_k(\mathcal{H})$ for each word w_k, so that when it is all put together $\mathcal{M}$ results, rather in the way in which control irregularities in programming languages (like QUIT, THROW, and ERROR) are handled in denotational semantics of programming languages by treating the continuation as a component of the context. The problem with such deviousness is that it essentially reduces compositionality to mean no more than that there exists *some* systematic overall story.

[21] Or, again, the meaning of the internal data structure or mental representation to which the word "relentless" corresponds. Nothing I am saying here (or anywhere else in this review) hinges on *external* properties of language. It's just simpler, pedagogically, to use familiar examples from natural language than to construct what must inevitably be hypothetical internal cases. As pointed out a few paragraphs back, of all the sorts of referential indefiniteness under review, only genuine ambiguity can be resolved during the parsing phase.

What it does bring into question are the assumptions on which such a system should be built, including for example the inferential viability of a system without any access to the interpretation of its representational structures—without, that is to say, *participating* in the subject matters about which it *reasons* (one way in which to resolve the obvious difficulty raised by the statement just made: that an agent know what is being said other than through the vehicle of the saying). But I'll leave some of these speculations until a later question.

For the time being, note merely that logic avoids this "meaning-depends-on-use" possibility like the plague. In fact the "use = representation + inference" aphorism reflects exactly the opposite theoretical bias: that representation (hence meaning) is an independent module in the intentional whole.

Once again, L&F's position is similar: nothing in their paper suggests they are prepared to make this radical a move. At one point they do acknowledge a tremendous richness in lexical significance, but after claiming this is all metaphor (which typically implies there is a firm "base case"), they go on to assert, without argument, that "these layers of analogy and metaphor eventually 'bottom out' at physical—somatic—primitives: up, down, forward, back, pain, cold, inside, seeing, sleeping, tasting, growing, containing, moving, making noise, hearing, birth, death, strain, exhaustion, ..." It's not a list I would want to have responsibility for completing.

More seriously, the integrity of L&F's project *depends* on avoiding use-dependent meaning, for the simple reason that they don't intend to consider use (their words: "you can never be sure in advance how the knowledge already in the system is going to be used, or added to, in the future", which they take as leading directly to the claim that it must be represented explicitly). If we were to take the meaning-depends-on-use stance seriously, we would be forced to conclude that *nothing in their knowledge base means anything*, since no one has yet developed a theory of its use.

I.e., L&F *can't* say yes to this one; it would pull the rug out from under their entire project.

In contrast (and as expected), the embedded view embraces the possibility. Perhaps the best way to describe the tension is in terms of method. A liberal logicist might admit that, in natural language, meaning is sometimes use-dependent in the ways described, but he or she would go on to claim that proper scientific method requires idealizing away from such recalcitrant messiness. My response? That such idealization throws the baby out with the bathwater. Scientific idealization is worth nothing if in the process it obliterates the essential texture of what one hopes to understand. And it is simply my experience that much of the structure of argument and discourse—even, the *raison d'être* of rationality—involves negotiating in an intentional space where meanings are left fluid by our linguistic and conceptual schemes, ready to be grounded in experience.

Question 4. Is consistency mandated?

Logic	L&F	EC
yes	no	no

L&F are quite explicit in rejecting an absolute dependence on consistency, to which traditional logical systems are so famously vulnerable. As indicated in the table, this is the first of the dozen questions where they and the embedded view align. That much said, however, it's not clear how deep the similarity goes. In particular, I'm unsure how much solace can be found in their recommendation that one carve the "knowledge base" into separate "buttes", and require each to be locally consistent, with neighbouring buttes maximally coherent. At least it's not clear, once again, without a much better intermediate theory.[22]

Fundamentally, the problem is that consistency is a relational property—the consistency of a set of sentences stands or falls on the set as a whole, not on an individual basis. This means that some relations between or among sentences (or frames) will have to be used as a basis for the partition (and to tie the resulting "buttes" together). Call these the system's *organizational principles*. Without them (on any remotely reasonable assumptions of error rates, dependence, etc.) the number of possible different configurations meeting their structural requirements would be intractably immense.

Furthermore, the organizational principles can't themselves be defined in terms of consistency; organizing a database *by* internal consistency would be crazy. Rather, I take it that what L&F really want is to be able to demonstrate (local) consistency for a database organized according to some other metric. What other metric? Surely only one makes sense: according to similarity or integrity of subject matter. X should be stored next to Y, in other words, because of the presence of (semantic) compatibility, not just the absence of (syntactic) incompatibility. Otherwise, descriptions of national politics might nestle up to lists of lemon meringue pie ingredients, but be kept separated from other statements about Washington policy making—so that things ended up together not because they agreed, but because they didn't have anything to do with one another.

So adequate organization will need to be defined in terms of a notion of subject matter. But where are we to find a theory of that? The problem is similar to that of representation in general: no one has one. The issue comes up in natural language attempts to identify topic, focus, etc. in theories of discourse (see, e.g., [30]), and in some of the semantical work in situation

[22] There's one problem we can set aside. As it happens, the very notion of consistency is vulnerable to the comments made in discussing question 3 (about use-dependent meaning). Like soundness and completeness, consistency, at least as normally formulated, is founded on some notion of semantic value *independent* of use, which an embedded view may not support (at least not in all cases). This should at least render suspicious any claims of *similarity* between the two positions. Still, since they stay well within the requisite conceptual limits, it's kosher to use consistency to assess L&F on their own (not that that will resolve them of all their troubles).

theory [3, 5]. But these are at best a start. Logic famously ducks the question. And informal attempts aren't promising: if my experience with the KRL project can be taken as illustrative [10], the dominant result of any such attempt is to be impressed with how seamlessly everything seems to relate to everything else.

When all is said and done, in other words, it is unclear how L&F plan to group, relate, and index their frames. They don't say, of course, and (in this case) no implicit principles can be inferred. But the answer is going to matter a lot—and not just in order to avoid inconsistency, but for a host of other reasons as well, including search, control strategy, and driving their "analogy" mechanism. Conclusion? That viable indexing (a daunting problem for any project remotely like L&F's), though different from consistency, is every bit as much in need as anything else of "middle-realm" analysis.

And as for consistency itself, we can summarize things as follows. Logic depends on it. L&F retain it locally, but reject it globally, without proposing a workable basis for their "partitioning" proposal. As for the embedded view (as mentioned in footnote 22) the standard notion of consistency doesn't survive its answer to question 3 (about use-dependent meaning). That doesn't mean, however, that I won't have to replace it with something analogous. In particular, I have no doubt that *some* notion of semantic viability, integrity, respect for the fact that the world (not the representation) holds the weight—something like that will be required for any palatable intentional system. Important as contextual setting may be, no amount of "use", reasoning processes, or consensual agreement can rescue a speaker from the potential of being wrong. More seriously, I believe that what is required are global *coordination conditions*—conditions that relate thinking, action, perception, the passing of the world, etc., in something of an indissoluble whole. To say more now, however—especially to assume that logic's notion can be incrementally extended, for example by being locally proscribed—would be to engage in tunneling of my own (but see [65]).

Question 5. Does the system use a single representational scheme?

Logic	L&F	EC
yes	yes	no

Tucked into a short paragraph of L&F's Section 9 is their response to the charge that one might encounter representational difficulties in trying to capture all of human knowledge. Their strategy is simple: "when something proves awkward to represent, add new kinds of slots to make it compactly representable". In fact they apparently now have over 5000 kinds. If only representation were so simple.

Several issues are involved. To start with, there is the question of the expressive adequacy of their chosen representational system—frames, slots, and values. Especially in advance, I see no reason to believe (nor argument to

convince me) that mass nouns, plurals, or images should succumb to this scheme in any straightforward way—or, to turn it upside down, to suppose that, if an adequate solution were worked out within a frame-and-slot framework, that the framework would contribute much to the essence of the solution. Frames aren't rendered adequate, after all, by encoding other representational schemes within them.[23]

Furthermore, one wonders whether any single representational framework—roughly, a representation system with a single structural grammar and interpretation scheme—will prove sufficient for all the different kinds of representation an intelligent agent will need. Issues range from the tie-in to motor and perceptual processing (early vision doesn't seem to be frame-like, for example; is late vision?) to the seeming conflict between verbal, imagistic, and other flavours of memory and imagination. You might view the difficulties of describing familiar faces in words, or of drawing pictures of plots or reductio arguments, as problems of externalizing a single, coherent, mentalese, but I suspect they really indicate that genuine intelligence depends on multiple representations, in spite of the obvious difficulties of cross-representational translation.

Certainly our experience with external representations supports this conclusion. Consider architecture: it is simply impossible not to be impressed with the maze of blueprints, written specifications, diagrams, topological maps, pictures, icons, annotations, etc., vital to any large construction project. And the prospect of reducing them all to any single representational scheme (take your choice) is daunting to the point of impossibility. Furthermore, there are reasons for the range of type: information easily captured in one (the shape of topological contours, relevant to the determination of building site, e.g.) would be horrendously inefficient if rendered in another (say, English).[24]

The same holds true of computation. It is virtually constitutive of competent programming practice to be able to select (from a wide range of possibilities) a particular representational scheme that best supports an efficient and consistent implementation of desired behaviour. Imagine how restrictive it would be if, instead of simply enumerating them in a list, a system had to record N user names in an unordered conjunction of N^2 first-order claims:

[23] As indicated in their current comments, L&F have apparently expanded their representational repertoire in recent years. Instead of relying solely on frames and slots, they now embrace, among other things: blocks of compiled code, "unparsed" digitized images, and statistical neural networks. But the remarks made in this section still largely hold, primarily because no mention is made of how these different varieties are integrated into a coherent whole. The challenge—still unmet, in my opinion—is to show how the "contents" contained in a diverse set of representational schemes are semantically commensurable, in such a way as to support a generalized, multi-modal notion of inference, perception, judgment, action. For some initial work in this direction see [6] for a general introduction, and [7] for technical details.

[24] Different representational types also differ in their informational prerequisites. Pictures and graphs, for example, *can't* depict as little information as can English text—imagine trying to draw a picture of "either two adults or half a dozen children".

$$(\exists x_1 \mid \text{user}(x_1)) \wedge (\exists x_2 \mid \text{user}(x_2)) \wedge \cdots \wedge (\exists x_n \mid \text{user}(x_n))$$
$$\wedge ((x_1 \neq x_2) \wedge (x_1 \neq x_3) \wedge \cdots \wedge (x_1 \neq x_n))$$
$$\wedge ((x_2 \neq x_3) \wedge \cdots) \wedge \cdots \wedge ((x_{n-1} \neq x_n))$$

Or how equally untenable it would be to prohibit a reasoning system from using existentials, or to limit it to domains where uniqueness of names could always be assumed. Yet one or other options would be forced by commitment to a "single scheme". Similarly, it's as unthinkable to prohibit display hardware from using bitmaps, in favour of frame-and-slot representations of each illuminated spot, as to force all representation into a bit-per-pixel mold.

Against all such considerations, however, logic and L&F are once again similar in pledging allegiance to a single representational scheme. As representative of the embedded view, I'll vote for variety.

Question 6. Are there only discrete propositions (no continuous representation, images, . . .)?	Logic	L&F	EC
	yes	yes	no

If pressed to represent continuous phenomena, L&F would presumably entertain real numbers as slot values, but that barely scratches the surface of the differences between discrete representations like formulae in a formal language, and various easily imagined forms of continuity, vagueness, indeterminacy, analogues, etc. And it is not just that we can imagine them; anything like real intelligence will have to deal with phenomena like this. We have the whole messy world to capture, not just the distilled, crystalline structure of Platonic mathematics.

In assessing the typology of representation, the distinction between discrete (digital) and continuous (analogue[25]) representations is sometimes given pride of place, as if that were the ultimate division, with all other possibilities subcategorized below it. But other just as fundamental divisions cross-cut this admittedly important one. For example, there is a question of whether a representation rests on a conception or set of formulated categories, or is in some way pre- or non-conceptual (terminology from [15]). The natural tendency, probably because of the prevalence of written language, is to assume that discrete goes with conceptual, continuous with non-conceptual, but this isn't true. The use of ocean buoys to demarcate treacherous water, for example, is presumably discrete but non-conceptual; intonation patterns to adjust the

[25] Calling continuous representations "analogue" is both unfortunate and distracting. "Analogue" should presumably be a predicate on a representation whose structure corresponds to that of which it represents: continuous representations would be analogue if they represented continuous phenomena, discrete representations analogue if they represented discrete phenomena. That continuous representations should historically have come to be called analogue presumably betrays the recognition that, at the levels at which it matters to us, the world is more foundationally continuous than it is discrete.

meanings of words ("what an *extraordinary* outfit") are at least plausibly both continuous and conceptual. Or consider another distinction: whether the base or "ur-elements" on which a representation is founded have determinate edges or boundaries. Both discrete and continuous objects of the sort studied in mathematics (the integers, the real line, and even Gaussian distributions and probability densities) are determinate, in the sense that questions about them have determinate answers. It's unclear, however, in questions about when tea-time ends, or about what adolescence is, or about exactly how many clouds there were when you poked your head out of your tent and said, with complete confidence, "there are lots of clouds today"—it's unclear in such cases whether there are determinate answers at all. The problem isn't an epistemic one, about incomplete knowledge, or a linguistic one, about the exact meanings of the words. The point is that the metaphysical facts just aren't there—nor is there any reason to suppose they should be there—to support a clean, black-and-white distinction. The competent use of the English plural, that is to say, doesn't require the existence of a denumerable base set of discrete elements. I am convinced that this distinction between phenomena that have sharp boundaries (support determinate answers) and those that don't is more profound and more consequential for AI than the distinction between discrete and continuous instances of each variety.

Modern logic, needless to say, doesn't deal with foundational indeterminacy. Nor are we given any reason to suppose that L&F want to take it on. One wonders, however, whether our lack of understanding of how relative certainty can arise on top of a foundationally vague base (no one would deny that there were lots of clouds outside that tent, after all) may not be the most important obstacle to the development of systems that aren't brittle in the way that even L&F admit we're limited to today.

Question 7. Do the representations capture all that matters?

Logic	L&F	EC
yes	yes	no

The situated view of representation cited earlier rests on the tenet that language, information, and representation "bridge the gap", in Perry's terms,[26] between the state of the user(s) of the representation, and the state of the world being referred to. It's a position that accords with a familiar view of language as dynamic action, rather than simply as static description. And it has among its more extreme consequences the realization that not all of what matters about a situation need be captured, at least in the traditional sense, in the meanings of its constituent representations.

For example, if someone simply yells "fire!", then some of what matters, including your understanding of what fire is, may be contributed by the surrounding situation, possibly even including the impinging thermal radiation. Call this totality of what matters—i.e., everything relevant to an assessment of

[26] The phrase is from various of John Perry's lectures given at CSLI during 1986–88.

whether the communication worked properly—its *full significance*. The claim, then, is that *the full significance of an intentional action can outstrip its content*. Facts of embodiment, of being there, of action, of experience, can, along with the content, influence the net or intended result.

To understand what this means, consider three things that it doesn't. First, it isn't merely a repetition of the claim made in discussing question 2: that conceptual content isn't uniquely determined by the type of representation used, but is partially determined by the context of its use. Nor, second, is it a replay of the stronger claim made in discussing question 3: that even the meanings—not just contents! (see footnote 17)—of words or internal structures may depend on their actual use. Although both of these involve use and context in a variety of ways, they remain claims about the relation between a representation and its semantic value. The current claim is stronger: that the full significance of an intentional act will outstrip even the situated semantic value of the representational ingredients constitutive of it, no matter how indexical, use-dependent, or situated a notion of content you care to come up with.

Even this last way of putting it, however, isn't strong enough, because it allows room for a third possible stance, stronger than the previous two (i.e., stronger than the embedded responses to questions 2 and 3), but still weaker than I have in mind here. In particular, someone might agree that an intentional action's full significance lies outside the content of the particular act itself, but go on to look for that additional contribution in the content of other representational structures. Thus, in determining the significance of "fire", you might look to other representations already present in the agent's head, or to conclusions that could be (quickly) drawn from things already represented. For example, you might expect to find the escape heuristic (that if someone shouts "fire!" it's good to get out of the way) represented in a previously stored internal frame.

I don't disagree that this can happen; in fact I take it as almost obvious (what else is inference for, after all?). However, I intend with this seventh question to get at a stronger position yet: that the full significance of an intentional action (not just a communicative one) can crucially involve *non-representational* phenomena, as well as representational ones. I.e., it is a claim that the millennial story about intelligence won't consist solely of a story about representation, but will inevitably weave that story together with analyses of other, non-representational aspects of an intentional agent. Some of these other ingredient stories will describe salient facts of embodiment (possibly even including adrenaline levels), but they will talk about other things as well, including genuine *participation* in represented subject matters,[27] and the internal *manifestation* (rather than *representation*) of intentionally important prop-

[27] The foundational notion underlying the view of embedded computation, in particular, is one of *partially disconnected participation*; see [65].

erties. Some modern roboticists, for example, argue that action results primarily from the dynamical properties of the body; the representational burden to be shouldered by the "mind", as it were, may consist only of adjustments or tunings to those non-representational capacities (see, e.g., [55, 56]). Rhythm may similarly as much be exhibited as encoded in the intelligent response to music. Or even take a distilled example from LISP: when a system responds with the numeral "3" to the query "(LENGTH '(A B C))", it does so by interacting with non-representational facts, since (if implemented in the ordinary way) the list '(A B C) will *have* a cardinality, but not one that is *represented*.

Distinguishing representational from non-representational in any careful way will require a better theory of representation than any we yet have.[28] Given such a story, it will become possible to inquire about the extent to which intelligence requires access to these non-formulated (non-formulable?) aspects of the subject matter. Although it's premature to take a definite stand, my initial sense is that there is every reason to suppose (at least in the human case) that it does. Introspection, common sense, and even considerations of efficient evolutionary design would all suggest that inferential mechanisms should avail themselves of any relevant available resources, whether those have arisen through representational channels, or otherwise. If this is true, then it follows that a system lacking any of those other channels—a system without the right kind of embodiment, for example—won't be able to reason in the same way we do. And so much the worse, I'd be willing to bet, for it.

How do our three players stand on this issue? I take it as obvious that L&F require what logic assumes: that representation has to capture all that matters, for the simple reason that there isn't anything else around. For L&F, in other words, facts that can't be described might as well not be true, whether about fire, sleep, internal thrashing, or the trials of committee work. They are forced to operate under a maxim of "inexpressible → irrelevant".

In contrast, as I've already indicated, I take seriously the fact that we are beaten up by the world—and not only in intentional ways. I see no reason to assume that the net result of our structural coupling to our environment—even that part of that coupling salient to intelligent deliberation—is exhausted by its representational record. And if that is so, then it seems overwhelmingly likely that the full structure of intelligence will rely on that residue of maturation and embodiment. So I'll claim no less for an embedded computer.

Here's a way to put it. L&F believe that intelligence can rest entirely on the meaning of *representations*, without any need for correlated, *non-representational experience*. On the other hand, L&F also imagine their system starting to read and distill things on its own. What will happen, however, if the writers

[28] Though some requirements can be laid down: such as that any such theory have enough teeth so that not *everything* is representational. That would be vacuous.

tacitly rely on non-representational actions on the part of the reader? The imagined system wouldn't be able to understand what it was reading. For example, there is no way in which L&F's system would ever be able to understand the difference between right and left.[29]

Question 8. Are reasoning and inference central? (Logic: yes | L&F: yes | EC: yes)

When logicians develop axiomatic accounts of set theory, criteria of elegance and parsimony push towards a minimal number of axioms—typically on the order of a dozen—from which an infinite number of truths follow. It's a general truth: economy of statement is often a hallmark of penetrating insight.

No one, however, expects distilled scientific theories alone to sustain complete, workaday, general-purpose reasoning. It is obvious that any reasonable problem solver (like any imaginable person), rather than deriving all its conclusions from first principles, will depend on a rich stock of facts and heuristics, derived results and rules of thumb—to say nothing of a mass of a-theoretic but relevant particulars (such as who it's talking to). So we should expect general intelligence to rest on a relatively high ratio of relevant truths to foundational axioms, especially in the face of resource-bounded processing, complex or just plain messy subject matters, and other departures from theoretical purity.

Nonetheless, you can't literally know everything. No matter how knowledgeable, an agent will still have to think in order to deal with the world *specifically*—to conclude that if today is Tuesday then tomorrow must be Wednesday, for example (derived from the general fact that Wednesdays follow Tuesdays), or to figure out whether your friend can walk from Boston to Cambridge, not otherwise having heard of your friend. Universal instantiation and modus ponens may not be all there is to thought, but without some such faculty a system would be certifiably CPU-dead.[30] And instantiating universals

[29] All the remarks made in footnote 16 apply here: it won't do to reply that L&F could build a model of right and left inside the system, or even attach a camera, since that would fall outside their stated program for representing the world. I too (i.e., on the embedded view) would attach a camera, but I want a *theory* of what it is to attach a camera, and of some other things as well—such as how to integrate the resulting images with conceptual representations, and how envisionment works, and how this all relates to the existence of "internal" sensors and effectors, and how it ties to action, and so on and so forth—until I get a theory that, as opposed to slots-and-frames, really does do justice to full-scale participation in the world. Cameras, in short, are just the tip of a very large iceberg.

[30] To imagine the converse, furthermore, would be approximately equivalent to the proposal that programming languages do away with procedures and procedure calls, in favour of the advance storage of the sum total of all potentially relevant stack frames, so that any desired answer could merely be "read off", without having to do any work. This is no more plausible a route to intelligence than to satisfactory computation more generally. And it would raise daunting issues of indexing and retrieval—a subject for which, as discussed under question 4 (on consistency), there is no reason to suppose that L&F have any unique solution.

is only the beginning. "Inference" includes not only deduction, but induction, abduction, inference to the best explanation, concept formation, hypothesis testing—even sheer speculation and creative flights of fancy. It can hardly be argued that some such semantically coordinated processing[31] is essential to intelligence.

It shouldn't be surprising, then, that inference is the one issue on which all three positions coincide—logic, L&F, and EC. But superficial agreement doesn't imply deep uniformity. There are questions, in each case, as to what that commitment means.

To see this, note that any inference regimen must answer to at least two demands. The first is famous: though mechanically defined on the form or structure of the representational ingredients,[32] inference must make semantic sense (that's what makes it *inference*, rather than ad hoc symbol mongering). There simply must be some semantic justification, that is to say—some way to see how the "formal" symbol manipulation coordinates with semantic value or interpretation. Second, there is a question of finitude. One cannot forget, when adverting to inference as the mechanism whereby a finite stock of representations can generate an indefinite array of behaviour, that the inference mechanism iself must be compact (and hence productive). The deep insight, that is to say, is not that reasoning allows a limited stock of information to generate an unlimited supply of answers, but that a synchronously finite system can manifest diachronically indefinite semantic behaviour.

Logic, of course, supplies a clear answer to the first demand (in its notion of soundness), but responds only partially to the second (hence the dashed lines around its positive answer). A collection of inferential schemata are provided—each demonstrably truth-preserving (the first requirement), and each applicable to an indefinite set of sentences (the second). But, as AI knows so well, something is still missing: the higher-level strategies and organizational principles necessary to knit these atomic steps together into an appropriate rational pattern.[33] Being able to reason, that is to say, isn't just the ability to take the right atomic steps; it means knowing how to think in the large—how to argue, how to figure things out, how to think creatively about the world. Traditional logic, of course, doesn't address these questions. Nor—and this is the important point—is there any a priori reason to believe that that larger inferential demand can be fully met within the confines of logic's peculiar formal and semantic conventions.

[31] By "semantically coordinated" I mean only to capture what deduction, induction, reasoning, contemplation, etc., have in common: roughly, some kind of coordination between what is done to (or happens because of, or whatever) a representation and its semantic value or content. Soundness, completeness, and consistency are particularly disconnected species; I suspect much more complicated versions will ultimately be required.

[32] Or so, at least, it is traditionally argued. This is not a view I am ultimately prepared to accept.

[33] For simplicity, I'm assuming that rational belief revision will consist of a pattern of sound inference steps—almost certainly not true. See e.g. [38].

On the other hand—and this takes us to the embedded view—once one moves beyond logic's familiar representational assumptions (explicit, a-contextual representation, and so forth), no one has yet presented an inferential model that meets the first demand. To accept the embedded answers to questions 1–7 is thus to take on a substantial piece of homework: developing, from the ground up, a semantically coordinated and rationally justifiable notion of inference itself. This is just one of the reasons why the embedded perspective is still emerging.

Nonetheless, important steps are being taken in this direction. The development of a contextually sensitive model of inference (based on a semantic notion of information, rather than symbolic form) is constitutive of Barwise and Etchemendy's work on situation theory, for example [6, 7]. Similarly, in the situated automata work of Rosenschein, a similarly non-syntactic notion of inference is analyzed in terms of a machine's carrying information relative to the structure of its embedding environment[34]. In a somewhat different vein, I have argued that an embedded notion of inference will ultimately be as relevant to clocks and other transducers as to sentential transformation [64]. It is also becoming clear that even more traditional (i.e., linguistic) forms of inference will as much involve the preservation of reference across a change in context, as the more familiar preservation of truth across a change in subject matter.[35] Important as these new thrusts are, however, they are still just early steps.

What about L&F? They have two options. To the extent that they agree with the present characterization of their position, vis-à-vis questions 1–7, they would probably want to avail themselves of logic's notion of inference. For reasons discussed earlier, however, this isn't enough: they would still have to take a stand on the relationship between truth-preserving logical entailment and the appropriate structure of rational belief revision, for example (see footnote 33), to say nothing of providing a finite account of an appropriate set of high-level control strategies, in order to provide a complete answer to the second demand. On the other hand, to the extent that they feel confined by logic's stringent representational restrictions (as they admit they do, for example, at least with respect to its insistence on full consistency—see question 4), and want to embrace something more like the embedded view, then they too must answer to the much larger demand: of not simply presenting their inferential mechanism (let alone claiming to have embraced 20 different ones), but of explaining what their very notion of inference is.

[34] Where information is approximately taken as counterfactual supporting correlation, in the spirit of Dretske [19] and Barwise and Perry [8]. See also Rosenschein [57].

[35] For the application of some of these ideas to the design of an embedded programming language, see [18].

Question 9. Are participation and action crucial?

Logic	L&F	EC
no	no	yes

Reasoning is a form of action. Earlier I commented on L&F's relegation of reasoning to a secondary status by their treatment of it as search, their suggestion that the "control" problem is largely solved, and their claim that with enough "knowledge" deep reasoning will be largely unnecessary.

But reasoning isn't the only kind of action that (at least in humans) has to be coordinated with representation. If you wander around Kyoto for the first time, poking your head into small shops, stopping for tea on the Philosopher's Walk, and gradually making your way back to the ryokan by something like dead reckoning, then your emergent conceptual understanding of the layout of the city must be constantly coordinated with your on-going but non-conceptual bodily movements. For example, if you remember that the hotel is somewhere off to your right, and then turn in that direction, you need to know that it is now roughly in front of you. In a similar way, we all need to know that tomorrow today will be "yesterday". Representations that lead to action often have to be revised in light of that very action's being taken.

Coordination management, as I will call this indissoluble blend of adjustment, feedback, action, belief revision, perception, dance, etc., arises in many corners of AI, ranging from planning and robotics to systems dealing with their own internal state (reflection and meta-level reasoning). Nor is AI the first discipline to recognize its importance: philosophers of science, and theorists of so-called "practical reasoning", have always realized the importance—and difficulty—of connecting thinking and doing. Students of perception, too, and of robotics, wrestle with their own versions of the coordination problem.

Curiously enough, even L&F, although they don't embrace a participatory stance, won't entirely be able to avoid it. Though their system will clearly shun the external world as much as possible,[36] it will still have to grapple with internal participation, if they go ahead with their proposal to encode (at the meta-level) such control knowledge as turns out genuinely to be needed. For example, suppose someone adds the following rule: that if the system uses any search strategy for more than 10 seconds without making definite progress, it should abandon that approach and try an alternative. Obeying this injunction requires various kinds of participation: recognizing that you have wasted 10 seconds (perception); stopping doing so (action); registering what it was that you were doing (perception); selecting a plausible alternative (inference); setting that new goal in motion (action); "letting go" of the meta-level deliberations (action on inference). Introspection and reflection might be better

[36] One thing it won't be able to shun, presumably, will be its users. See footnote 37.

described as varieties of self-*involvement* than of self-*reference* (in spite of my "Varieties of Self-Reference" [63]; see also [65]).[37]

So we end this one with a curious tally. In virtue of its utterly disconnected stance, and of *not being a computational system*, logic is singularly able to ignore action and subject matter participation. On the embedded side, I take participatory connections with the world as not just important, but as essential. In fact the embedded view could almost be summed up in the following claim:

> *Participation in the subject matter is partially constitutive of intelligence.*

When all is said and done, in other words, I believe the term "intelligent" should be predicated of an integrated way of being that includes both thought and action, not simply an abstract species of disconnected symbol manipulation. This may contravene current theoretical assumptions, but I suspect it is consonant with ordinary common sense. Frankly, I don't see how you could believe a system could comprehend all of consensus reality without being able to understand "See you tomorrow!".[38]

Between these two, L&F occupy a somewhat unstable middle ground. I have listed them with logic, since that's where their claims go; there is no hint that they envisage tackling issues of coordination. On the other hand, they will have to confront coordination management merely in order to get their system to turn over, quite apart from whether it manifests anything I would call intelligence.

Question 10. Is physical embodiment important?

Logic	L&F	EC
no	no	yes

The authors of the mathematical theory of computability claimed as a great victory their elevation of the subject of computation from messy details of physical implementation and fallible mechanism onto a pure and abstract plane. And the prime results of recursive function theory, including the famous

[37] This paragraph makes explicit something I have otherwise tried, in this article, to sidestep: the fact that (at least on my analysis) L&F's theoretical framework is not only inadequate for understanding *intelligence*, but is also inadequate for understanding *their own system* (which, I am claiming, won't be *intelligent*, but will still *exist*). Driving a wedge between what computation is actually like and how we think of it is a primary brief of [65]; for the moment, simply assume that L&F, if they proceed with their project, will have to resort to a-theoretical programming techniques to handle this and other such issues. Control structure is only one example; another is user interaction. To the extent computers carry on conversations, after all, they actually *carry them on*, rather than merely representing them as being carried on (though they may do that as well).

[38] Again, as I said in footnote 16, it won't do to reply that they could simply add a counter to mark the passage of time. For one thing (or at least so I claim) this example, although simple, is symptomatic of a deep problem; it's not a surface nuisance to be programmed around. Furthermore, even if it were simply disposed of, for L&F to treat it in an ad hoc, procedural way would be to part company with their own analysis.

proofs of undecidability, genuinely didn't seem to rely on any such implementational details. Modern programmers don't typically traffic in recursive function theory in any very conscious way, but they still accept the legacy of a computational level of analysis separate from (and possibly not even theoretically reducible to[39]) the physical level at which one understands the underlying physical substrate.

More recently, however, especially with the increasing realization that relative computability is as important as (if not more important than) the absolute computability of the 1930s, the story is growing murkier. Though it treats its subject matter abstractly, complexity theory still deals with something called time and space; it's not entirely clear what relation those rather abstract notions bear to the space and time of everyday experience (or even to those of physics). At least with regard to time, though, real (non-abstract) temporal properties of computation are obviously important. Whether differences among algorithms are measured in minutes, milliseconds, or abstract "unit operations", the time they take when they run is the same stuff that I spend over lunch. And the true spatial arrangement of integrated circuits—not just an abstracted notion of space—plays an increasing role in determining architectures.

Although it isn't clear where this will all lead, it does allow the question to be framed of whether considerations of physical embodiment impinge on the analysis of a given computational system. For traditional logic, of course, the answer is *no*; it is as pure an exemplar as anything of the abstract view of computation and representation. And once again L&F's stance is similar: nothing suggests that they, along with most of the formal tradition, won't ignore such issues.

Again the embedded view is different. I am prepared to argue that physical constraints enter computational thinking in a variety of familiar places. For one thing, I have come to believe that what (in a positive vein[40]) we call the "formality" of computation—the claim, for example, that proof procedures rely solely on the formal properties of the expressions they manipulate—amounts in the end to neither more nor less than "whatever can be physically realized in a causally efficacious manner".[41] But this is not the only place where

[39] *Reducibility*, as the term is normally used in the philosophy of science, is a relation between *theories*; one theory is reducible to another if, very roughly, its predicates and claims can be translated into those of another. In contrast, the term *supervenience* is used to relate phenomena themselves; thus the strength of a beam would be said to supervene on the chemical bonds in the constitutive wood. The two relations are distinguished because people have realized that, somewhat contrary to untutored intuition, supervenience doesn't necessarily imply reducibility (see [27, 33, 40, 41]).

[40] As opposed to the "negative" reading: namely, that a formal computational process proceed independently of the semantics. That the two readings are *conceptually* distinct is obvious; that they get at different things is argued in [65].

[41] I am not asking the reader to agree with this statement, without more explanation—just to admit that it is conceptually coherent.

physical realization casts its shadow. Consider one other example: the notion of locality that separates doubly-linked lists from more common singly-linked ones, or that distinguishes object-oriented from function-based programming languages. Locality, fundamentally, is a physical notion, having to do with genuine metric proximity. The question is whether the computational use is just a metaphor, or whether the "local access" that a pointer can provide into an array is metaphysically dependent on the locality of the underlying physics. As won't surprise anyone, the embedded viewpoint endorses the latter possibility.

Question 11. Does the system support "original" semantics?

Logic	L&F	EC
no	no	yes

It has often been pointed out that books and encyclopedias derive their semantics or connection to what they're about from the people that use them. The analogous question can be asked about computers: whether the interpretations of the symbol structures they use are in any sense "authentic" or "original" to the computers themselves, or whether computational states have their significance only through human attribution (see, e.g., [17; 31, pp. 32ff; 60].

The question is widely accepted, but no one has proposed a really good theory of what is required for semantical originality, so not a whole lot more can be said. Still, some of the themes working their way through this whole set of questions suggest that this issue of originality may be relevant not only for philosophical reasons but also for purposes of adequate inference and reasoning. In particular, if the only full-blooded connection to subject matter is through external users, then it follows that a system won't be able to avail itself of that connection in carrying out its processes of symbol manipulation, reasoning, or inference. If, on the other hand, the semantic connection is autonomous (as one can at least imagine it is, for example, for a network mail system that not only represents facts about network traffic, but also sends and receives real mail), then the chances of legitimate inference may go up.[42]

So the question should be read as one of whether the way of looking at the system, in each case, points towards a future in which systems begin to "own" their semantic interpretations—if still in a clunky and limited way, then at least with a kind of proto-originality.

Even that vague a formulation is sufficient to corral the votes—and to

[42] I am not suggesting that physical involvement with the subject matter is sufficient for original intentionality; that's obviously not true. And I don't mean, either, to imply the strict converse: that anything like simple physical connection is *necessary*, since we can obviously genuinely refer to things from which we are physically disconnected in a variety of ways—by distance, from other galaxies; by fact, from Santa Claus; by possibility, from a round square; by type, from the number 2. Still, I am hardly alone in thinking that *some kind of causal connectivity* is at least a constituent part of the proper referential story. See e.g. Kripke [43], Dretske [19], and Fodor [28].

produce another instance of what is emerging as the recurring pattern. Like logic, L&F neither address nor imagine their system possessing anything like the wherewithal to give its frames and slots autonomous referential connection with the world. In fact something quite else suggests itself. Given the paucity of inference they imagine, the heavy demands on indexing schemes, and the apparent restriction of interaction to console events, L&F's system is liable to resemble nothing so much as an electric encyclopedia. No wonder its semantics will be derivative.

Now it's possible, of course, that we might actually want an electric encyclopedia. In fact it might be a project worth pursuing—though it would require a major and revealing revision of both goals and procedure. Note that L&F, on the current design, retain only the formal data structures they generate, discarding the natural language articles, digests, etc., used in its preparation. Suppose, instead, they were to retain all those English entries, thick with connotation and ineffable significance, *and use their data structures and inference engines as an active indexing scheme*. Forget intelligence completely, in other words; take the project as one of constructing the world's largest hypertext system, with CYC functioning as a radically improved (and active) counterpart for the Dewey decimal system. Such a system might facilitate what numerous projects are struggling to implement: reliable, content-based searching and indexing schemes for massive textual databases. CYC's inference schemes would facilitate the retrieval of articles on related topics, or on the target subject matter using different vocabulary. And note, too, that it would exploit many current AI techniques, especially those of the "explicit representation" school.

But L&F wouldn't be satisfied; they want their system itself to know what those articles mean, not simply to aid us humans. And it is against that original intention that the embedded view stands out in such stark contrast. With respect to owls, for example, an embedded system is more likely to resemble the creatures themselves than the *Britannica* article describing them. And this, I submit, to return to the question we started with, is the direction in which semantical originality lies.

Question 12. Is room made for a divergence between the representational capacities of theorist and agent?

Logic	L&F	EC
no	no	yes

The final question has to do with the relation between the representational capacities of a system under investigation, and the typically much more sophisticated capacities of its designer or theorist. I'll get at this somewhat indirectly, through what I'll call the *aspectual* nature of representation.

It is generally true that if X represents Y, then there is a question of *how* it represents it—or, to put it another way, of how it represents it *as being*. The

two phrases "The Big Apple" and "the hub of the universe" can both be used to represent New York, but the latter represents it as something that the former does not. Similarly, "the MX missile" and Reagan's "the Peacemaker".

The "represent *as*" idiom is telling. If we hear that someone knew her brother was a scoundrel, but in public *represented him as* a model citizen, then it is safe for us to assume that she possessed the representational capacity to represent him in at least these two ways. More seriously—this is where things can get tricky—we, *qua* theorists, who characterize her, *qua* subject, know what it is to say "as a scoundrel", or "as a citizen". We know because we too can represent things as scoundrels, as citizens, and as a myriad other things as well. And we assume, in this example, that our conceptual scheme and her conceptual scheme overlap, so that we can get at the world in the way that she does. So long as they overlap, trouble won't arise.[43]

Computers, however, generally don't possess anything remotely like our discriminatory capacities,[44] and as a result, it is a very substantial question for us to know how (from their point of view) they are representing the world as being. For example (and this partly explains McDermott's [49] worries about the wishful use of names), the fact that we use English words to name a computer system's representational structures doesn't imply that the resulting structure represents the world for the computer in the same way as that name represents it for us. Even if you could argue that a KRYPTON node labeled $DETENTE genuinely represented detente, it doesn't follow that it represents it as what we would call detente. It is hard to know how it does represent it as being (for the computer), of course, especially without knowing more about the rest of its representational structures.[45] But one thing seems likely: $DETENTE will mean less for the computer than "detente" means for us.

I suspect that the lure of L&F's project depends in part on their ignoring "as" questions, and failing to distinguish theorists' and agents' conceptual schemes. Or at least this can be said: that they are explicitly committed to not making a distinction between the two. In fact quite the opposite is presumably their aim: what they want, of the system they propose to build, is something

[43] In logic, this required overlap of registration scheme turns up in the famous mandate that a metalanguage used to express a truth theory must *contain* the predicate of the (object) language under investigation (Tarski's convention T). Overlap of registration scheme, however, is at least potentially a much more complex issue than one of simple language subsumption.

[44] Obviously they are simpler, but the differences are probably more interesting than that. The individuation criteria for computational processes are wildly different from those for people, and, even if AI were to succeed up to if not beyond its wildest dreams, notions like "death" will probably mean something rather different to machines than to us. Murder, for example, might only be a misdemeanor in a society with reliable daily backups.

[45] It would also be hard (impossible, in fact) for us to say, exactly, what representing something as detente would mean for *us*—but for a very different reason. At least on a view such as that of Cussins [15], with which I am sympathetic, our *understanding* of the concept "detente" is not itself a conceptual thing, and therefore can't necessarily be captured in words (i.e., concepts aren't conceptually constituted). Cf. the discussion of formulation in Section 2.

that we can interact with, in our own language (English), in order to learn or shore up or extend our own understanding of the world. In order for such interaction to work—and it is entirely representational interaction, of course—the two conceptual schemes will have to be commensurable, on pain of foundering on miscommunication.

Here, though, is the problem. I assume (and would be prepared to argue) that an agent (human or machine) can only carry on an intelligent conversation using words that represent the world in ways that are part of that agent's representational prowess. For an example, consider the plight of a spy. No matter how carefully you try to train such a person to use a term of high-energy physics, or the language of international diplomacy, subsequent conversations with genuine experts are almost sure to be awkward and "unintelligent" (and the spy therefore caught!) unless the spy can genuinely come to register the world in the way that competent users of that word represent the world as being.

It follows, then, that L&F's project depends for its success on the consonance of its and our conceptual schemes. Given that, the natural question to ask is whether the sketch they present of its construction will give it that capacity. Personally, I doubt it, because, like Evans [25], I am convinced that most common words take their aspectual nature not only from their "hook-up" to other words, but from their direct experiential grounding in what they are about. And, as many of the earlier questions have indicated, L&F quite clearly don't intend to give their system that kind of anchoring.

So once again we end up with the standard pattern. Neither traditional logic nor L&F take up such issues, presuming instead on what may be an unwarranted belief of similarity. It is characteristic of the embedded view to take the opposite tack; I don't think we'll ever escape from surprises and charges of brittleness until we take seriously the fact that our systems represent the world differently from us.

4. The logical point of view

No twelve questions, briefly discussed, can exhaust the representational terrain. Still, the general drift is clear. The repeated overlap between L&F and traditional logic betrays L&F's conception of what it is to be an "intelligent system". They must have in mind something similar to the prototypical logic-based theorem prover or question and answer system: the user types in a question and the system types back the answer, or the user types in a statement and the system types T or F, depending on its truth—that kind of thing. The system is conceived of entirely abstractly; it would have to be physically embodied, of course, in order to be typed at, but the level at which it was analyzed (syntax of frames, values of slots, etc.) would abstract away from all

such physical considerations. Such a system would not only be analyzed as disembodied, and be entirely disconnected from any of the subject domains that it "knew" about, it would thereby achieve what humans so rarely do: the ability to look out on the world from a completely objective, detached, a-contextual, universal ("from nowhere") vantage point.

As the reader will have guessed, I don't for a minute think such an achievement is possible, for man or machine (or even desirable; at its best intelligence should prepare you for being anywhere, not for being nowhere). But that's not really my point. Here, in the end, is what is most impressive about their paper. When all is said and done, L&F's vision of an intelligent system is remarkably similar to the traditional logical one: a complete axiomatization of the world manipulated by a general purpose inference engine. *The "logicists", after all, never assumed that theorem proving was any substitute for competent axiomatization*; exactly the opposite is argued by McCarthy, Hayes, Hobbs, and others [34–37, 47]. L&F, however, have the distinction of using a much less expressive language (at least as far as we can tell, given that no semantic account seems to be in the cards), and of assuming no definite control regimen. Plus one more thing: unlike any modern logicist writer, they claim they can do the whole thing.

5. Conclusion

To take representing the world seriously (it's world representation, after all, not knowledge representation, that matters for AI) is to embrace a vast space of possibilities. You quickly realize that the intellectual tools developed over the last 100 years (primarily in aid of setting logic and meta-mathematics on a firm foundation) will be about as much preparation as a good wheel-barrow would be for a 24-hour dash across Europe. The barrow shouldn't be knocked; there are good ideas there—such as using a wheel. It's just that a little more is required.

So there you have it. L&F claim that constructed intelligence is "within our grasp". I think it's far away. They view representation as explicit—as a matter of just writing things down. I take it as an inexorably tacit, contextual, embodied faculty, that enables a participatory system to stand in relation to what is distal, in a way that it must constantly coordinate with its underlying physical actions. L&F think you can tunnel directly from generic insight to system specification. I feel we're like medieval astrologers, gropings towards our (collective?) Newton, in a stumbling attempt to flesh out the theoretical middle realm. There is, though, one thing on which we do agree: we're both enthusiastic. It's just that I'm enthusiastic about the work that lies ahead; L&F seem enthusiastic that it won't be needed.

Why?—why this difference? Of many reasons, one goes deep. From my

point of view, knowledge and intelligence require participation in the world. Lenat and Feigenbaum, apparently, think not. I can only conclude that they would not agree with Yeats, who I think said it well:

> I have found what I wanted—to put it all in a phrase, I say, "Man can embody the truth, but cannot know it."[46]

References

[1] P.E. Agre, Routines, AI Memo 828, MIT, Cambridge, MA (1985).

[2] P.E. Agre, The dynamic structure of everyday life, Ph.D. Thesis, Tech. Rept., MIT, Cambridge, MA (1989).

[3] J. Barwise, The situation in logic II: conditionals and conditional information, in: E.C. Traugott, C.A. Ferguson and J.S. Reilly, eds., *On Conditionals* (Cambridge University Press, Cambridge, 1986); also: Rept. No. CLSI-85-21, Stanford, CA (1985); reprinted in: J. Barwise, *The Situation of Logic*, CLSI Lecture Notes **17** (University of Chicago Press, Chicago, IL, 1989) Chapter 5.

[4] J. Barwise, *The Situation of Logic*, CLSI Lecture Notes **17** (University of Chicago Press, Chicago, IL, 1989).

[5] J. Barwise and J. Etchemendy, Model-theoretic semantics, in: M. Posner, ed., *Foundations of Cognitive Science* (MIT Press, Cambridge, MA, 1989).

[6] J. Barwise and J. Etchemendy, Visual information and valid reasoning, in: W. Zimmermann, ed., *Visualization in Mathematics* (Mathematical Association of America, to appear).

[7] J. Barwise and J. Etchemendy, Information, infons, and inference, in: R. Cooper, K. Mukai and J. Perry, eds., *Situation Theory and Its Applications* 1, CLSI Lecture Notes (University of Chicago Press, Chicago, IL, 1990) 33–78.

[8] J. Barwise and J. Perry, *Situations and Attitudes* (MIT Press, Cambridge, MA, 1983).

[9] D.G. Bobrow, ed., *Qualitative Reasoning about Physical Systems* (North-Holland, Amsterdam, 1984).

[10] D.G. Bobrow, T. Winograd et al., Experience with KRL-0: one cycle of a knowledge representation language, in: *Proceedings IJCAI-77*, Cambridge, MA (1977) 213–222.

[11] R. Boyd, Metaphor and theory change: what is "metaphor" a metaphor for?, in: A. Ortony, ed., *Metaphor and Thought* (Cambridge University Press, Cambridge, 1979).

[12] R.A. Brooks, A robust layered control system for a mobile robot, *IEEE J. Rob. Autom.* **2** (1986) 14–23.

[13] D. Chapman and P.E. Agre, Abstract reasoning as emergent from concrete activity, in: M.P. Georgeff and A.L. Lansky, eds., *Reasoning about Action and Plans: Proceedings of the 1986 Workshop* (Morgan Kaufmann, Los Altos, CA, 1987) 411–424.

[14] W.J. Clancey, The frame of reference problem in the design of intelligent machines, in: K. VanLehn, ed., *Architectures for Intelligence* (Erlbaum, Hillsdale, NJ, to appear).

[15] A. Cussins, The connectionist construction of concepts, in: M. Boden, ed., *The Philosophy of Artificial Intelligence*, Oxford Readings in Philosophy Series (Oxford University Press, Oxford, 1990) 368–440.

[16] R. Davis, ed., *Expert Systems: How Far Can They Go? AI Mag.* **10** (1–2) (1989).

[17] D.C. Dennett, *The Intentional Stance* (MIT Press, Cambridge, MA, 1987).

[18] M.A. Dixon, Open semantics and programming language design (working title), Doctoral Dissertation, Computer Science Department, Stanford University, Stanford, CA (to appear).

[46] Taken from a letter Yeats wrote to a friend shortly before his death. Dreyfus cites the passage at the conclusion of the introduction to the revised edition of his *What Computers Can't Do* [21, p. 66]; it has also been popularized on a poster available from Cody's Books in Berkeley.

[19] F. Dretske, *Knowledge and the Flow of Information* (MIT Press, Cambridge, MA, 1981).
[20] F. Dretske, *Explaining Behavior: Reasons in a World of Causes* (MIT Press/Bradford Books, Cambridge, MA, 1988).
[21] H.L. Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence* (Harper Row, New York, rev. ed., 1979).
[22] H.L. Dreyfus, From micro-worlds to knowledge representation: AI at an impasse, in: J. Haugeland, ed., *Mind Design: Philosophy, Psychology, Artificial Intelligence* (MIT Press, Cambridge, MA, 1981) 161–205.
[23] H.L. Dreyfus, ed., *Husserl, Intentionality, and Cognitive Science* (MIT Press, Cambridge, MA, 1982).
[24] H.L. Dreyfus and S.E. Dreyfus, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* (Macmillan/Free Press, New York, 1985).
[25] G. Evans, *The Varieties of Reference* (Oxford University Press, Oxford, 1982).
[26] R. Fagin and J.Y. Halpern, Belief, awareness, and limited reasoning, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985) 491–501.
[27] J.A. Fodor, Special sciences (or: the disunity of science as a working hypothesis), *Synthese* **28** (1974) 97–115; reprinted in: N. Block, ed., *Readings in the Philosophy of Psychology* (Harvard University Press, Cambridge, MA, 1980) 120–133.
[28] J.A. Fodor, *Psychosemantics* (MIT Press/Bradford Books, Cambridge, MA, 1987).
[29] D. Gentner and D. Gentner, Flowing waters or teeming crowds: Mental models of electricity, in: D. Gentner and A. Stevens, eds., *Mental Models* (Erlbaum, Hillsdale, NJ, 1983).
[30] B.J. Grosz and C.L. Sidner, Attention, intentions, and the structure of discourse, *Comput. Linguistics* **12** (3) (1986) 175–204.
[31] J. Haugeland, Semantic engines: introduction to mind design, in: J. Haugeland, ed., *Mind Design: Philosophy, Psychology, Artificial Intelligence* (MIT Press, Cambridge, MA, 1981) 1–34.
[32] J. Haugeland, ed., *Mind Design: Philosophy, Psychology, Artificial Intelligence* (MIT Press, Cambridge, MA, 1981).
[33] J. Haugeland, Weak supervenience, *Am. Philos. Q.* **19** (1) (1982) 93–103.
[34] P.J. Hayes, The second naive physics manifesto, in: J.R. Hobbs and R.C. Moore, eds., *Formal Theories of the Commonsense World* (Ablex, Norwood, NJ, 1985) 1–36.
[35] P.J. Hayes, Naive physics I: ontology for liquids, in: J.R. Hobbs and R.C. Moore, eds., *Formal Theories of the Commonsense World* (Ablex, Norwood, NJ, 1985) 71–107.
[36] J.R. Hobbs and R.C. Moore, eds., *Formal Theories of the Commonsense World* (Ablex, Norwood, NJ, 1985).
[37] J.R. Hobbs et al., Commonsense summer: final report, Tech. Rept. CSLI-85-35, Stanford University, Stanford, CA (1985).
[38] D.J. Israel, What's wrong with non-monotonic logic?, in: *Proceedings AAAI-80*, Stanford, CA (1980).
[39] L. Kaelbling, An architecture for intelligent reactive systems, in: M.P. Georgeff and A.L. Lansky, eds., *Reasoning about Action and Plans: Proceedings of the 1986 Workshop* (Morgan Kaufmann, San Mateo, CA, 1987) 395–410.
[40] J. Kim, Supervenience and nomological incommensurables, *Am. Philos. Q.* **15** (1978) 149–156.
[41] J. Kim, Causality, identity, and supervenience in the mind-body problem, *Midwest Stud. Philos.* **4** (1979) 31–49.
[42] D. Kirsh, When is information explicitly represented?, in: P. Hanson, ed., *Information, Language, and Cognition*, Vancouver Studies in Cognitive Science **1** (University of British Columbia Press, Vancouver, BC, 1990) 340–365.
[43] S.A. Kripke, *Naming and Necessity* (Harvard University Press, Cambridge, MA, 1980).
[44] J. Lave, *Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life* (Cambridge University Press, Cambridge, 1988).
[45] H.J. Levesque, A logic of implicit and explicit belief, in: *Proceedings AAAI-84*, Austin, TX (1984) 198–202.
[46] D.M. Levy, D.C. Brotsky and K.R. Olson, Formalizing the figural, in: *Proceedings ACM Conference on Document Processing Systems*, Santa Fe, NM (1988) 145–151.

[47] J. McCarthy and P.J. Hayes, Some philosophical problems from the standpoint of artificial intelligence, in: B. Meltzer and D. Michie, eds., *Machine Intelligence* **4** (American Elsevier, New York, 1969) 463–502.
[48] J.L. McClelland, D.E. Rumelhart and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* **2**: *Psychological and Biological Models* (MIT Press/Bradford Books, Cambridge, MA, 1986).
[49] D.V. McDermott, Artificial intelligence meets natural stupidity, in: J. Haugeland, ed., *Mind Design: Philosophy, Psychology, Artificial Intelligence* (MIT Press, Cambridge, MA, 1981) 143–160.
[50] T. Nagel, *The View from Nowhere* (Oxford University Press, Oxford, 1986).
[51] D.A. Norman, *The Psychology of Everyday Things* (Basic Books, New York, 1988).
[52] A. Ortony, ed., *Metaphor and Thought* (Cambridge University Press, Cambridge, 1979).
[53] J. Perry, The problem of the essential indexical, *NOUS* **13** (1979) 3–21.
[54] J. Perry and D. Israel, What is information?, in: P. Hanson, ed., *Information, Language, and Cognition*, Vancouver Studies in Cognitive Science **1** (University of British Columbia Press, Vancouver, BC, 1990) 1–19.
[55] M.H. Raibert, Legged robots, *Commun. ACM* **29** (6) (1986) 499–514.
[56] M.H. Raibert and I.E. Sutherland, Machines that walk, *Sci. Am.* **248** (1) (1983) 44–53.
[57] S. Rosenschein, Formal theories of knowledge in AI and robotics, *New Generation Comput.* **3** (4) (1985).
[58] S. Rosenschein and L. Kaelbling, The synthesis of digital machines with provable epistemic properties, in: *Proceedings Workshop on Theoretical Aspects of Reasoning about Knowledge* (Morgan Kaufmann, Los Altos, CA, 1986); also: Tech. Rept. CSLI-87-83, Stanford University, Stanford, CA (1987).
[59] D.E. Rumelhart, J.L. McClelland and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* **1**: *Foundations* (MIT Press/Bradford Books, Cambridge, MA, 1986).
[60] J.R. Searle, Minds, brains, and programs, *Behav. Brain Sci.* **3** (1980) 417–424; reprinted in: J. Haugeland, ed., *Mind Design: Philosophy, Psychology, Artificial Intelligence* (MIT Press, Cambridge, MA, 1981) 282–306.
[61] J.R. Searle, *Minds, Brains, and Science* (Harvard University Press, Cambridge, MA, 1984).
[62] B.C. Smith, Prologue to "Reflection and semantics in a procedural language", in: R.J. Brachman and H.J. Levesque, eds., *Readings in Knowledge Representation* (Morgan Kaufmann, Los Altos, CA, 1985) 31–39.
[63] B.C. Smith, Varieties of self-reference, in: J.Y. Halpern, ed., *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the 1986 Conference* (Morgan Kaufmann, Los Altos, CA, 1986).
[64] B.C. Smith, The semantics of clocks, in: J. Fetzer, ed., *Aspects of Artificial Intelligence* (Kluwer Academic Publishers, Boston, MA, 1988) 3–31.
[65] B.C. Smith, *A View from Somewhere: An Essay on the Foundations of Computation and Intentionality* (MIT Press/Bradford Books, Cambridge, MA, to appear).
[66] P. Smolensky, On the proper treatment of connectionism, *Behav. Brain Sci.* **11** (1988) 1–74.
[67] L.A. Suchman, *Plans and Situated Actions* (Cambridge University Press, Cambridge, 1986).
[68] A. Tarski, The concept of truth in formalized languages, in: A. Tarski, ed., *Logic, Semantics, Metamathematics* (Clarendon Press, Oxford, 1956) 152–197.
[69] T. Winograd, Moving the semantic fulcrum, Tech. Rept. CSLI-84-77, Stanford University, Stanford, CA (1984).
[70] T. Winograd, Thinking machines: Can there be? Are we?, Tech. Rept. CSLI-87-100, Stanford University, Stanford, CA (1987).
[71] T. Winograd, Three responses to situation theory, Tech. Rept. CSLI-87-106, Stanford University, Stanford, CA (1987).
[72] T. Winograd and F. Flores, *Understanding Computers and Cognition: A New Foundation for Design* (Ablex, Norwood, NJ, 1986).