

Research Paper

Is the connectionist-logicist clash one of AI's wonderful red herrings?¹

SELMER BRINGSJORD

Abstract. A careful adjudication of the connectionist-logicist clash in AI and cognitive science seems to disclose that it is a mirage.

Keywords: logicism, connectionism, symbol systems, neural networks

Received 20 March 1991; revised 10 July 1991

1. Introduction

An often unfriendly debate continues to rage in AI and cognitive science between 'logicists' and 'connectionists.' Many connectionists (e.g. Smolensky 1988a, Churchland & Churchland 1990, Waltz 1988, Schwartz 1988, Kaplan *et al.* 1990, Horgan and Tienson 1989) hold that their doctrines ought to supplant or at least supplement the logic-based ones of traditional logicist or symbolist (or 'strong,' Searle 1980a,b, 1982; 'good old-fashioned,' Haugeland 1986; 'old hand,' Doyle 1988; 'person building,' Charniak & McDermott 1985, Nilsson & Genesereth 1988, Pollock 1989) AI. On the other hand, many logicists (e.g. Fodor & Pylyshyn 1988) hold that any successful AI model of human cognition, and *a fortiori* any sentient artificial intelligence itself, must use classical, logic-driven architecture.

Herein I will attempt to adjudicate this clash—by, in a word, showing it to be 'one of AI's wonderful red herrings.'² In order to carry out this adjudication, I will need to adopt an approach that relies heavily on the formalization of declarative English sentences within first-order logic. There is considerable irony in the fact that the only rigorous method open to one seeking adjudication of the clash in question is one with which the logicist is likely to be comfortable, but one that is, if not anathema, then at least a bit foreign, to a connectionist more at home with differential equations than syllogisms. While human cognition, from where we stand at the moment, may or may not be profitably identified with the connectionist's so-called dynamical systems, one thing *is* clear, even at this stage, before embarking on our inquiry: the connectionist-logicist *clash* isn't treatable as such a system. It is, rather, treatable, if at all, as a clash of *propositions* thought by their proponents to be true and their opponents to be false.³ A logic-based adjudication of the debate in question is something connectionists, most prominently (Smolensky 1988a), have themselves attempted, but not, in my opinion, pulled off. (Later, we will see that Smolensky's 1988a 'declarativization' of the clash leaves much to be desired.) At any rate, let us begin: let \mathcal{C} (\mathcal{L}) denote connectionism (logicism) in the form of some as-yet-unarticulated set of

propositions. Let \vdash denote ordinary first-order implication; you are free to think of this implication in terms of your favourite theorem prover, or some natural deduction system with which you are familiar, or the sequent calculus, . . . whatever. Where Φ is a set of first-order formulae, we write $\text{Inc } \Phi$ iff Φ is inconsistent, in the ordinary sense that there is some formula ϕ such that $\Phi \vdash \phi$ and $\Phi \vdash \neg \phi$. If Φ is consistent, i.e. not inconsistent, we write $\text{Con } \Phi$.

'Declarativizing' the clash in question will allow us to pass beyond the impressionistic and ask what truth values the elements of \mathcal{C} (\mathcal{L}) have, what variations on \mathcal{C} (\mathcal{L}) can be carried out by tinkering with (perhaps by deleting from, or adding to) the propositions therein, what things follow from \mathcal{C} (\mathcal{L}), etc. And we can ask the 'big question,' namely:

Con ($\mathcal{C} \cup \mathcal{L}$)?

This is the big question because, among other reasons, if an affirmative response to it is correct, then connectionists (logicists) who see their approach as the *one* that will succeed will have been shown to be misguided. (Smolensky 1988a, p. 6), a prominent connectionist, has described the situation by saying that if $\text{Con} (\mathcal{C} \cup \mathcal{L})$, then 'connectionist modelling does become mere implementation.' He goes on to say

Such an outcome [i.e. $\text{Con} (\mathcal{C} \cup \mathcal{L})$] would constitute a genuine defeat of a research program that I believe many connectionists are pursuing.

Many logicists would also say that $\text{Con} (\mathcal{C} \cup \mathcal{L})$ implies a defeat of *their* research programme. I will soon show, however, that $\text{Con} (\mathcal{C} \cup \mathcal{L})$ is in fact the case—but in doing so I will put on display accounts of \mathcal{C} and \mathcal{L} which are not in the least defeated by this consistency.

It is worth noting at the outset that the opposing sides in what will hereafter be abbreviated as the \mathcal{C} - \mathcal{L} debate are far from monolithic. It seems to me that each side, \mathcal{C} and \mathcal{L} , is *itself* constituted by a good number of mutually exclusive instantiations of what might be called the connectionist or logicist 'spirit.' To put it schematically, \mathcal{C} might be constituted by $\mathcal{C}_0, \dots, \mathcal{C}_k$ independent positions^d, and \mathcal{L} the same with respect to $\mathcal{L}_0, \dots, \mathcal{L}_k$, where \mathcal{C}_0 and \mathcal{L}_0 are the most extreme versions of connectionism and logicism, respectively. Accordingly, the first thing I will do is set out, in the form of a continuum, \mathcal{C}_0 to \mathcal{L}_0 . The second this continuum is out in the open, it will become clear that adjudicating the general \mathcal{C} - \mathcal{L} debate in less than the space of an entire book is probably well nigh impossible (which in part explains why [Smolensky 1988a + Peer Commentary + Smolensky 1988b] was a protracted affair), and that we had better set ourselves the task here of adjudicating a specific *sub*-debate, say that between \mathcal{C}_3 and \mathcal{L}_3 . And yet, on the positive side, if, sticking with our pretend participants \mathcal{C}_3 and \mathcal{L}_3 , it were shown that the \mathcal{C}_3 - \mathcal{L}_3 is a red herring, the stage would perhaps be set to dissolve other \mathcal{C}_i - \mathcal{L}_i clashes. To a small but significant degree, I will be consciously occupied with setting the stage in this way.

I will adjudicate debate between what I call 'strong' connectionism (\mathcal{C}_S) and 'strong' logicism (\mathcal{L}_S), two camps that occupy determinate points in my continuum. My adjudication will consist in the dissolution of the \mathcal{C}_S - \mathcal{L}_S clash. In a nutshell, I will argue, carefully, that in light of well-known simulation proofs making, in some to-be-defined sense, cellular automata 'the same as' Turing machines, \mathcal{C}_S is in a bind. The bind is this: if \mathcal{C}_S presupposes the falsity of functionalism (roughly, for now, the view that consciousness can arise from the correct causal

interconnection of physical stuff quite different from human flesh, including, say, a silicon-based substrate),⁵ then, given that functionalism is very plausible, the only way \mathcal{C}_S can remain viable is if, in reversal, it *affirms* functionalism but embraces the view that analog computing devices are qualitatively superior to non-analog devices—a view that has little or no empirical or theoretical support. I will then attempt to show that the dilemma is nothing for a proponent of \mathcal{L}_S to gloat over—because a parallel one threatens \mathcal{L}_S : \mathcal{L}_S 's explicit affirmation of functionalism, conjoined with elementary results from computability theory, implies that there is nothing special about a symbolic program over and above a connectionist's subsymbolic system.

Two assumptions underlie the coming argumentation, and are worth setting out before we embark.

I assume, first, that the sort of AI (or cognitive science; I will hereafter collapse both under 'AI') with which we are concerned, whether it be connectionist or logicist or hybrid in spirit, is, at bottom, *aggressive*. Someone who views AI as nothing more than the attempt to do things like model computationally the olfactory component of rat brains will find the debate with which I am concerned to be otiose. On the other hand, if one has a sanguine, rounded view of AI, my treatment should be of interest. Such a view of AI, from my perspective, is twofold in nature, namely that AI's engineering side is reflected by the aim of *building* an agent, or mind, or person (not necessarily of the human variety), while its scientific side is reflected by the fact that reaching the engineering objective requires a thorough *understanding* of mentality itself. (This twofold view of AI will be working in the background when, below, I try to define \mathcal{C}_S and \mathcal{L}_S). If all you care about is building a particular system, such as one that can carry out real-time machine translation, or one that can drive a car, land a plane, run an automated factory—if this is all that is near and dear to your heart, if you do not see the ultimate aim of AI as building a genuine robot agent, then this paper isn't for you.

My second assumption is simply that readers of what is to come are familiar with the concepts central to the debate in question. I assume here, in particular, that readers have a background, on the connectionist side, largely derivable from Rumelhart & McClelland (1986, volume I), that is that they have assimilated neural net concepts like input, output and hidden units, activation, values, weights, training with back propagation, and so-called recurrent nets.⁶ On the logicist side, I assume readers to be comfortable with n -order extensional logics, timid to full-blown intentional logics, and traditional symbolic projects in AI employing fragments of these symbolic schemes. It would be nice if readers had a formal understanding (sufficient, say, to assimilate Lindstrom's first and second theorems—see Chapter XII of Ebbinghaus *et al.* 1984) of what have come to be called 'symbol systems,' but an informal account of the sort given by Harnad (1990) suffices. Harnad's account is simply a generalization of a first-order symbol system based on the familiar ' \vdash ' and ' \vDash .' This simplest of symbol systems is based on the traditional symbol set (of n -ary relation symbols and functors) which, in accordance with the standard formation rules (e.g. if ϕ and ψ are wffs, then $\phi \wedge \psi$ is a wff), allows for the formation of 'atomic' formulae, and then more complicated 'molecular' formulae. Sets of these formulae (say Φ) may be said to proof theoretically entail individual formulae (say ϕ); such a situation is encapsulated by such familiar meta-expressions as ' $\Phi \vdash \phi$,' an expression that is already at the heart of our so-called 'big question,' cited above. A symbol system

must also include a semantic side that systematically provides meaning. In first-order systems, formulae are said to be true (or false) on an *interpretation*, often written as $I \models \phi$. For example, the formula ' $\forall x \exists y Gyx$ ' might mean, on the standard interpretation (for arithmetic) that for every natural number n , there is a natural number m such that $m > n$. Generalizing on this system is easy enough to envisage. A starting place would be to allow a symbol set to be, in Harnad's words, 'scratches on paper, holes on a tape, events in a digital computer, etc.'—the possibilities are endless. What is often not appreciated is that there are as many possibilities for fleshing out \vdash and \models in non-first-order ways. This is in fact why I insist that participants in the debate in question be in command of the sophisticated machinery that underlies \mathcal{L}_S ; it is also why I am being so wordy about logicist concepts and not connectionist ones.⁷ Logicists no doubt find it quite infuriating to have the symbol systems that are admittedly at the core of their program immutably identified with first-order symbol systems, or with unhelpful buzzwords like 'sentence-logic view of cognition' (Churchland & Churchland 1989). The fact that logicist programs are, to use the word connectionists are fond of using (Smolensky 1988a, 1989), 'brittle,' may very well be *because* these programs are first-order. Even intermediate-level books like Nilsson & Genesereth (1988) urge consideration of logicist techniques that are hardly first-order (e.g. circumscription, which in some instances involves processing on second-order formulae).

I assume, to continue, that the reader is also in command of the basic concepts and proofs of elementary computability theory, e.g. finite state automata, Turing machines, k -tape Turing machines, cellular automata, and simulation proofs (e.g.) of the fact that, qualitatively speaking, bestowing non-determinism upon an automata gives it no new power, that a k -tape Turing machine is no more powerful than a standard one, that a cellular automaton can be viewed as just a k -tape Turing machine, and that a neural net can be recast as, among other things, a probabilistic cellular automaton. I would, in addition, like to assume that readers are familiar with analog devices, but this is perhaps unreasonable, since not only are physical analog computers in short supply, but also there is no satisfactory logico-mathematical or philosophic definition of ' x is an analog computer.' There is of course fundamental agreement that analog devices, at least considered from the theoretical standpoint, employ a 'non-symbolic' form of representation. If, for example, you want to program an ordinary digital computer to sort n numbers, you will have to use some *symbols* to stand for the numbers you seek to sort. On the other hand, there are analog computers which represent numbers directly. One such device is the famous spaghetti computer. As is well known, roughly $n \log n$ comparisons are required to sort n numbers; hence computation time grows faster than a linear function of n . But this is a complexity bound that can *apparently* be broken by the spaghetti computer.⁸ (For more on the spaghetti computer, and other analog devices, see Dewdney 1984.) I have found, much to my surprise, that some thinkers are unaware of the fact that though analog computers can apparently break complexity bounds for sequential symbolic machines, no analog computer of any sort, whether theoretical, homey (as in the spaghetti computer), or electronic, has ever cracked an *NP*-complete problem (though *perhaps* a decent argument can be mustered for saying that they come close; see Courant & Robbins 1941, nor has one ever solved a problem in principle unsolvable (e.g. the halting problem). These facts are related to the discussion to come.

Now, my plan is as follows. In section 2, I present the aforementioned continuum which gives an overview of the \mathcal{C} – \mathcal{L} debate, and grounds the sub-debate into which I descend and propose to dissolve. This continuum will contain, among other things, a *rough* characterization of \mathcal{C}_S (again: strong connectionism) and \mathcal{L}_S (again: strong logicism), the two camps of special concern to us. Jumping off from the continuum, I will fine-tune the characterizations of \mathcal{C}_S and \mathcal{L}_S within it, and then summarize the typical ‘Fodorian’ case for \mathcal{L}_S , and find it wanting. In section 3 I attempt to go beyond the continuum and set out, in detail, some of the propositions underlying it. In section 4 I characterize \mathcal{L}_S , and take a stab at a workable account of \mathcal{C}_S , in both cases by organizing the propositions isolated in section 3. In section 5, \mathcal{C}_S is gradually refined. This gradual refinement is the side-effect of a dialectic which eventuates in a contraction of \mathcal{C}_S implying Con ($\mathcal{C}_S \cup \mathcal{L}_S$). This result, as I point out at the end of section 5, in no way marks a victory for \mathcal{L} . Rather, the upshot of Con ($\mathcal{C}_S \cup \mathcal{L}_S$) is a vindication of an affirmative answer to the title of this paper: the \mathcal{C} – \mathcal{L} clash *is* a red herring. I end section 5 with a critical look at Smolensky’s (1988) well-known attempt to establish something like Inc ($\mathcal{C}_S \cup \mathcal{L}_S$). Since some will argue baldly that ‘so many clever AIniks couldn’t be so upset about a red herring,’ I will end, in section 6, with some sociological speculation about *why* we have the \mathcal{C} – \mathcal{L} clash, and why it is so strident.

2. The continuum

Figure 1 shows the continuum on the general \mathcal{C} – \mathcal{L} debate, as well as the position, advocated in this paper, that the debate in question is a red herring—a position I call ‘Ecumenical AI.’ This continuum is not meant to be completely detailed: there are doubtless many participants in the debate who deserve to be mentioned within it, but who are not.⁹ (There are also some camps sometimes included in the \mathcal{C} – \mathcal{L} debate absent by design from Figure 1, e.g. a camp Smolensky (1988a) terms ‘eliminativist neural.’ This camp does not make it into the continuum of Figure 1 for the simple reason that its members refuse to abstract from the ‘neural level,’ neither in the direction of connectionism, nor in the direction of logicism.) Moreover, this continuum is, of necessity, *compressed*. It is preferable, but in the space we have here, impossible, to spell out in detail all points along the continuum. It may be, after further analysis, that there will arise positions not captured in Figure 1. (One such uncaptured position might be Bechtel 1988.) At any rate, as noted, of special interest to us in this paper are the two camps strong connectionism, \mathcal{C}_S , and strong logicism, \mathcal{L}_S ; and so, taking off from the continuum, I will provide, in a little preview, explication of these camps.

But before this preview a word about hyper-logicism, \mathcal{L}_H , and hyper-connectionism, \mathcal{C}_H . I know of no one who has championed these views in the literature (with the possible exceptions, with respect to \mathcal{C}_H , of Hillis 1989, and, with respect to \mathcal{L}_H of Bringsjord and Zenzen 1991). As the continuum indicates, however, there is a construal of McCarthy’s famous ‘two horses in the race’ comment according to which this attitude coincides, depending on one’s prior prejudices, either with \mathcal{C}_H or \mathcal{L}_H . In a picturesque word, \mathcal{L}_H is the view that AI should proceed in the hope of building robot agents like those in Putnam’s (1981) brain-in-a-vat thought-experiment. These agents would have no sensors and no effectors, and would have no need (so the story would go) of subsymbolic processing so well-suited (as the connectionists have shown) to handling the

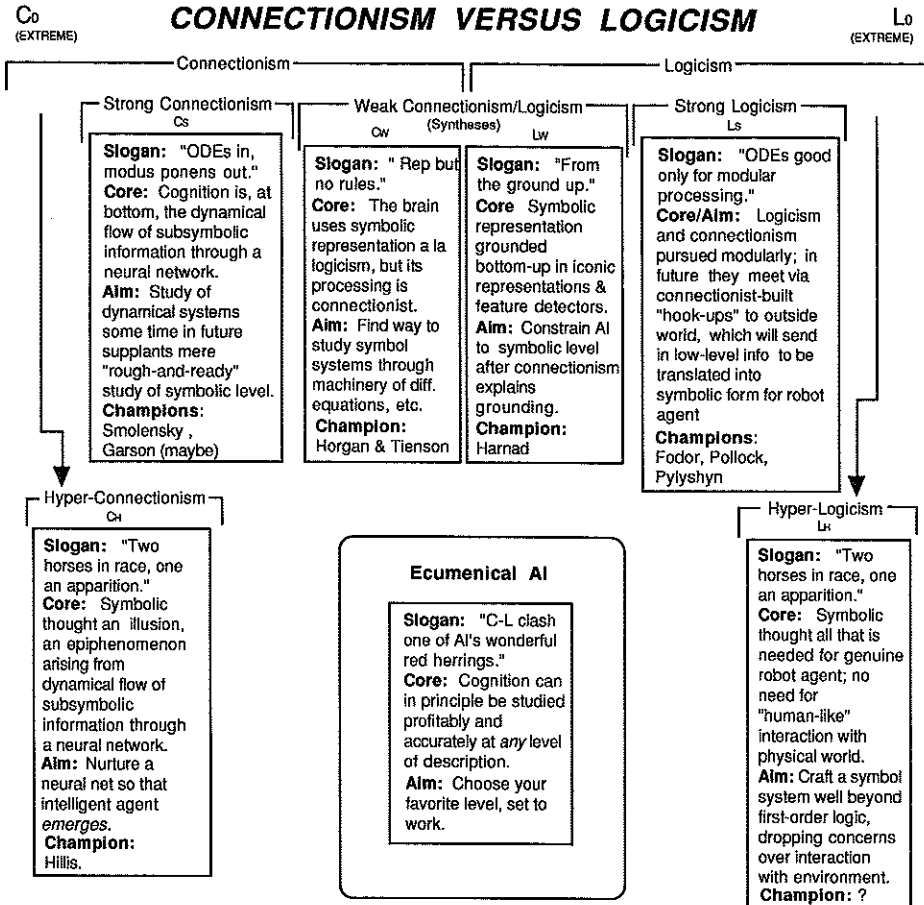


Figure 1.

relation between an agent and the world through which it navigates. I will, to a point, explicate \mathcal{L}_H below. What would hyper-connectionism, in a word, be like? Presumably the basic idea behind this view would be that a neural net could be 'dropped down in the physical environment' and, without *any* training from AIniks, develop into a genuine agent (see Hillis 1989 for an impressionistic vision of such a scenario). One might say, in a suggestive gloss, that \mathcal{C}_H is the view that learning is everything and installation nothing, while \mathcal{L}_H is the view that installation (of symbol-based knowledge) is everything and learning nothing.¹⁰ When I reach, later, the point at which I begin to specify \mathcal{L}_S and \mathcal{C}_S , I will, by modifying these specifications, provide at least rough-and-ready accounts of \mathcal{L}_H and \mathcal{C}_H .

At any rate, on to a preview of \mathcal{L}_S . The classic account of strong logicism is given by Fodor & Pylyshyn (1988) (see also Fodor 1980), who argue that connectionism may provide theories of the *implementation* of cognition, but not theories of psychology, i.e. not theories of what is *apparently* (emphasis to preclude begging any questions) symbolic in nature—things for example like the ability of human person P to produce/understand a sentence S if there is another

related sentence S' which P produces/understands. (E.g. set S = 'John loves Mary,' set S' = 'Mary loves John.')

Strong connectionists like Smolensky seem to hold the reverse, that the proper level of psychological description is necessarily sub-symbolic, and symbolicism emerges parasitically from sub-symbolic processing.

The Fodorian argument, at least *this* Fodorian argument, is unsuccessful, for the simple reason that, pointed out to a large degree by Garson (1990), recent advances on the connectionist front (compare Servan-Schreiber *et al.* 1988, Elman 1989, Kaplan *et al.* 1990) have resulted in systems that model the abilities thought by Fodor and company to be symbolic in nature. This development would seem to be thoroughly unsurprising, because it would appear to be just what the formal results ensure. If one puts no artificial limit on type or complexity of a neural net, then you quickly have the μ -recursive functions available, and therefore you have Turing computability. The converse holds too: every Turing machine can be matched by a neural net. Why anyone would have denied a proposition like 'Neural nets can do X ' while affirming some such thing as 'Turing machines can do X ' is beyond me. The mathematics of the situation, specifically the ultimate equivalence of neural nets and Turing machines, would seem to doom forever the Fodorian tack. We should before long have connectionist systems on the scene that are very good at handling those aspects of human language Fodorians hold to be the special province of logicist approaches.

I will, as promised above, further blur, with help from computability theory and mathematical logic, the \mathcal{C}_S – \mathcal{L}_S divide. The main argument, expressed in section 5, trades on the fact that strong logicists, *if they take what I called above the 'aggressive, rounded' view of AI*, must commit to an ontology in which robot agents are to be identified with symbolic paradigms, e.g. Turing machines, while strong connectionists of the same aggressive flavour must commit to an ontology in which robot agents are to be identified with connectionist paradigms, e.g. neural nets.

3. Propositions

Let's now move without further ado to a list of key propositions in the \mathcal{C} – \mathcal{L} debate. I will do my best to give these propositions mnemonic labels. We will begin with three theses which represent sanguinity about the engineering side of AI. Here is our first of these three:

(PBP)¹¹ ANiks will succeed in building a robotic person S^* .

It should be clear by this time that this thesis, given the two camps we are concerned with, gives rise to two theses, one tied to \mathcal{C}_S and one to \mathcal{L}_S , so that we have

(PBP_C) Connectionist ANiks will succeed in building a robotic person S_C^* .

(PBP_L) Logicist ANiks will succeed in building a robotic person S_L^* .

The idea is that the robots (androids?) to arrive in the future which verify (PBP_C) or (PBP_L) will mark a certain specific meeting place between symbolic and sub-symbolic processing, but the how of the meeting will differ depending on which perspective, \mathcal{C}_S or \mathcal{L}_S , we're affirming. I have sketched the two different views on this meeting in the continuum of Figure 1. Proponents of (PBP_L), according to this continuum, might hold that the meeting will be 'top-down,' one in which symbolic processing dominates, at least at the level of what will be the counterpart

to what is apparently symbolic reasoning in human persons, or what Smolensky (1988a) calls the 'conscious rule interpreter'. Strong logicians will *allow*, however, that connectionist systems will prosper when they are used to capture 'intuitive processing' (compare Smolensky 1988a):

[The intuitive processor] is presumably responsible for all of animal behavior and a large portion of human behavior: Perception, practiced motor behavior, fluent linguistic behavior, intuition in problem solving and game-playing—in short, practically all skilled performance. (Smolensky 1988a, p. 5)

Proponents of (PBP_C), on the other hand, might hold that this meeting will be 'bottom-up,' one in which sub-symbolic processing dominates, leaving perhaps only an epiphenomenal bit of symbolic processing.

It is important to note that the 'meeting' of which I have just spoken marks a concession on the part of \mathcal{L}_S to \mathcal{C}_S (viz, that 'skilled performance' may be captured best by \mathcal{C}_S , not by \mathcal{L}_S). This is a concession which does appear to come through loud and clear in the work of strong logicians. For example, Pollock's (1989) 'Q&I modules' are presumably connectionist in nature; in fact they no doubt correspond to that behaviour at which Smolensky's 'intuitive processor' excels. Pollock does in fact appear to take precisely this position:

Let me ... acknowledge that ... building a robot will ... involve vision, or more generally, perception, and also motor control, language, and myriad special-purpose modules that shortcut reasoning to make information processing more efficient. These are the Q&I systems (quick and inflexible) ... In some ways [symbolic] reasoning is the least important element of such a robot. A robot incorporating sophisticated subsystems for perception, motor control, and a variety of Q&I systems, could function very well in a cooperative environment without [symbolic] reasoning. It would be a mechanical analogue of a lower animal, some of which are extraordinarily good at what they do. But such a robot would lack any kind of self-awareness and would not be a person. On the other hand, a robot endowed with [symbolic] reasoning but lacking any Q&I systems would probably be too slow to function in real time. (Pollock 1989, p. 13)

It should be admitted, however, by both sides, that we don't now know, precisely, *how* the building of these androids will take place. We can only at present note the shared engineering goal of \mathcal{C}_S and \mathcal{L}_S , and discuss competing *strategies* for reaching this goal—a competition I've only to this point roughly sketched.

Here are the remaining key propositions. We begin with a few more useful possible 'pair-offs' of agents and computing creatures, which will enter into the discussion below.

- (PER_{AUT}) Persons are automata.
- (PER_{TUR}) Persons are Turing machines.
- (PER_{NN}) Persons are neural networks.
- (PER_{CA}) Persons are cellular automata.

Readers who find the 'agent-oriented' account of AI underlying this paper to be alien, may find the (PER₇) theses a bit odd. They should not, however—for the following reasons: first, there is reason to think that quantification over agents *of some sort* cannot be kept out of discussions about the logico-mathematical and philosophical foundations of AI. Thus Smolensky (1988a), while not (often, anyway) quantifying over agents (cognizers, persons, ...) *does* quantify over virtual machines which, for all intents and purposes, operate as agents in his discussion. Indeed, we saw Smolensky's technique used above: he speaks of the intuitive

processor and the conscious rule processor. The second thing that should be said about discomfiture over the (PER₇) theses is that there are formidable arguments for the claim that so-called *de se* beliefs (e.g. 'I believe that AI is moribund.') must be present in *any* robot-of-the-future intended by AIniks to match or exceed the symbolic reasoning, perceptual, motor, ... powers of human persons (Pollock 1989). And it is hard to see how AIniks could 'build-in' to a robot *de se* beliefs, whether these beliefs are symbolist or connectionist in character, without treating this robot, and without having the robot treat itself, as an *agent*. Third, and finally, I should hasten to point out that the (PER₇) terminology is intended to be a form of shorthand.¹² After all, a standard first-order symbolization of (PER_{TUR}) (where Tx iff x is a Turing machine, Px if x is a person) as

$$\forall x(Px \rightarrow \exists y(Ty \wedge x = y))$$

would, on the assumption that I am a person, imply that

$$\exists y(Ty \wedge selmer = y).$$

But this is hard to swallow, since Turing machines are (at bottom; see any formal account) *sets*, which, given both that (i) sets are (usually regarded to be, anyway) non-physical, and (ii) a corollary of Leibniz's Law, viz.

$$\forall x \forall y [x = y \rightarrow (Fx \rightarrow Fy)],$$

implies that *I* am non-physical. But one surely does not want to define \mathcal{L} (\mathcal{C}) in such a way that it *entails* agent dualism. Of course, each of the (PER₇) theses presented above can be slightly modified to reflect a physicalist orientation,¹³ as in for example

(PER_{PHYS-TUR}) Persons are *physical* Turing machines.

And in fact it will turn out below that the proponent of \mathcal{C}_S , in order to dodge one leg of my argument for the \mathcal{C}_S - \mathcal{L}_S clash being a red herring, will affirm (PER_{PHYS-TUR}). In light of this coming move by the strong connectionist, it is worth noting here that it does not follow from (PER_{PHYS-TUR}) that there is some standard way to *manufacture* Turing machines. This thesis simply does not take a stand on the manufacturing process. There are, of course, innumerable ways to physically implement Turing machines; and these implementations, if you will, can be ranked in terms of speed, reliability, mobility, and so on. It would furthermore seem, intuitively, that (PER_{PHYS-TUR}) is referring only to those Turing machines which fall into a specific sub-spectrum in this ranking. From this point on, unless otherwise noted, the machines alluded to in the (PER₇) theses are those that conform to the parameters delimited by this sub-spectrum. Below, when considering Smolensky's attempts to demonstrate Inc ($\mathcal{C}_S \cup \mathcal{L}_S$), we will return to a discussion of these parameters.

There are other 'silent' aspects of the (PER₇) theses. Consider, for example,

(PER_{AUTS}) Persons are automata *with sensors*.

It may be plausibly said that this thesis comes closer than (PER_{AUT}) and its 'sensorless' relatives to capturing the heart of \mathcal{L}_S 's view of future robot agents. There is no doubt that sensors are crucial to \mathcal{C}_S and \mathcal{L}_S : machine vision, for example, is an indispensable part of both movements. However, I do not think we need bother to add the 'sensor clause' to our (PER₇) theses, because,

generally speaking, sensors do not increase baseline computing power, and baseline computing power is what I focus on in the coming argumentation.¹⁴ Since the focus in the present paper is on baseline computer power, and since we lack the space to treat all the issues that arise once one allows sensors and effectors into the picture, and since both sides of the debate agree that sensors are a necessary component of a robot agent, the 'sensorized' versions of our (PER₇) theses will be ignored in what is to come.¹⁵

Now let us attempt to make explicit the remaining propositions involved in the \mathcal{C}_S - \mathcal{L}_S clash. Here, first, is a stab at articulating the sort of functionalism traditionally part of \mathcal{L}_S ; it is called 'AI-Functionalism' by Rey (1986):

- (AI-F) For every two 'brains' x and y , possibly constituted by radically different physical stuff: if the overall flow of information in x and y , represented as a pair of flow charts (or a pair of Turing machines, or a pair of Turing machine diagrams, ...), is the same, then if 'associated'¹⁶ with x there is an agent s in mental state S , there is an agent s' 'associated' with y which is *also* in S .

More—quite a *bit* more, in fact—about this thesis (its plausibility, roots, etc.) later.

Now here a few familiar general 'Church-Turing-related' principles hard to keep out of any debate like the one presently occupying us:

- (CTT) Whatever can be rendered as an algorithm can be rendered as a suitable programmed Turing machine, and vice versa.
- (CTT*) Whatever can be accomplished by a computing machine of any sort, can be accomplished by a suitably programmed Turing machine.
- (ANA) A true analog, neural net can compute things which no Turing machine can.
- (ANA*) A true analog, neural net can compute things that cannot be expressed as algorithms.

And finally the encapsulation I prefer of what is near and dear to the heart of strong logicians.

- (SYM) If (PBP), i.e. if a robotic person S^* will be eventually produced by AIniks, then S^* must be such that some¹⁷ of the propositions ϕ_0, ϕ_1, \dots which are objects of S^* 's occurrent deliberations (and hopes, fears, etc.—the objects of her *propositional attitudes*) are represented by formulae $\langle\phi_0\rangle, \langle\phi_1\rangle, \dots$ of some symbol system \mathcal{L}^T , where they can be processed according to the reasoning mechanism that is part of \mathcal{L}^T .

Note that (SYM)'s \mathcal{L}^T is a *symbol system*; \mathcal{L}^T is *not* to be cursorily considered, to use Fodor's famous phrase, a 'language of thought.' (SYM)'s \mathcal{L}^T is also quite in line with the production system approach underlying, say, SOAR (Laird *et al.* 1987). Symbol systems *subsume* particular logicist architectures.

There are many interesting logical relations among the propositions just enumerated. Many of these relations are controversial, but some are quite obvious, as in

$$(R1) \quad (ANA) \rightarrow \neg (CTT^*)$$

and

$$(R2) \quad ((ANA) \wedge (CTT)) \rightarrow (ANA^*)$$

Controversial relations among the propositions cited above are those involving (PER_{TUR}) and (SYM). It may be thought, for example, that

$$(R3) \quad (PBP) \rightarrow [(SYM) \rightarrow (PER_{TUR})].$$

What formal rationale might incline one to affirm (R3) \approx ‘If AI-niks will build persons, then given that robot-persons will carry out their thought in symbol systems, persons of any variety *are* Turing machines’? Well, consider the following apparent proof of (R3): in preparation for two applications of conditional proof, assume (PBP) and (SYM). By *modus ponens* we immediately have (SYM)’s consequent. Accordingly, let S^* be the robotic agent-of-the-future produced by AI under the assumption that (PBP); and, in addition, let ϕ_1, ϕ_2, \dots be the objects of S^* ’s propositional attitudes.

We now appeal to the standard concepts of Turing machine configurations and of movement from configuration to configuration in a Turing machine: Suppose, encapsulating these standard accounts, that $c_n \vdash_M c_{n+1}$ iff Turing machine M is permitted to go in one step from configuration c_n to c_{n+1} ; and write $c_n \vdash^*_M c_{n+k}$ iff M is permitted to go from configuration c_n to c_{n+k} in a *number* of steps, whatever they might be. Fix some method of representing Φ (and first-order formulae in general) as input on a Turing machine’s tape (and there are many: recall, for example, the one at the heart of the common proof of the undecidability of first-order logic, where the fact that a given formula ϕ is valid iff a corresponding Turing machine halts is employed); call this encoding $\mathcal{E}[\Phi]$. Suppose also that we render in computational terms all the deductive rules underlying the ordinary concept of \vdash . Then we have a situation where $\Phi \vdash \phi$ iff $c_{\mathcal{E}[\Phi]} \vdash^*_M c_{\mathcal{E}[\phi]}$, where the configurations here are (obviously) ones in which the indicated formula(e) are, in encoded form, on the Turing machine’s tape.

The purported proof concludes as follows: what the machinery we have introduced in the previous paragraph allows us to do, overall, is to translate agent S^* into a Turing machine, say machine M^* . Since S^* is arbitrary, we have derived that agents (or persons) are Turing machines, i.e. that (PER_{TUR}). By successive applications of conditional proof, as planned, we arrive at (R3).

But this ‘proof’ is mistaken. All the standard machinery for interchanging Turing machines and proof-theoretic schemes shows us, in the context of the assumptions that get the ‘proof’ here going, and where $\mathcal{P}\text{--}\mathcal{L}^T$ denotes the proof-theoretic component of symbol system \mathcal{L}^T , is that

$$(R4) \quad (PBP) \rightarrow [(SYM) \rightarrow \exists x(Tx \wedge x = \mathcal{P}\text{--}\mathcal{L}^T)]$$

which is of course mightily unimpressive, since we knew at the outset that the proof-theoretic component of a symbol system can be identified with some Turing machine. What’s needed, of course, is the proposition that *persons* are Turing machines, not merely that *part* of the *symbol systems* associated with persons are Turing machines. In general, then, despite the close connection between symbol systems and Turing machines, (R3) is false; and the moral of the story is that those bent on collapsing together the agent-oriented (PER_{AG}) theses, with theses about symbol systems associated with agents, ought to tread warily in their

attempt to do so. From this point on, in fact, I will assume that it makes good sense to assume autonomous ontological categories for agents, automata, and symbol systems. The general issue here will not, however, go away; it will crop up again when we look below at Smolensky's (1988) attempt to show $\text{Inc} (\mathcal{C}_S \cup \mathcal{L}_S)$.

4. \mathcal{C}_S Versus \mathcal{L}_S

How, given the propositions we have allowed ourselves, can we put together determinate accounts of \mathcal{C}_S and \mathcal{L}_S ? Well, as I see it, strong logicism (of an aggressive, rounded nature, recall) consists of the following propositions:

\mathcal{L}_S
(PBP _L)
(PER _{TUR})
(AI-F)
(CTT)
(CTT*)
(SYM)

This account of \mathcal{L}_S is one which Fodor, Pylyshyn, Pollock and company would embrace. But not only that: the account here coincides with strong logicism as it is viewed by its detractors. Churchland & Churchland 1990, p. 32 tell us, for example, that \mathcal{L} (subscript absent: the Churchland's discussion isn't fine-grained enough to warrant imbedding it into the continuum of Figure 1) is based on (i) 'the enormous power of symbols that undergo rule-governed transformations' ((SYM)), (ii) 'Church's thesis, that every effectively computable function is recursively computable,' ((CTT*), (CTT)), (iii) the fact 'that any recursively computable function can be computed in infinite time ... by a universal Turing machine' ((PER_{TUR})), and, in light of the following quote, (iv) (AI-F):

There were a few puzzles, of course [concerning the \mathcal{L} program]. For one thing, symbol-manipulating machines were admittedly not very brainlike. Even here, however the classical approach [= \mathcal{L}] had a convincing answer. First, the physical-material of any symbol-manipulating machine had nothing essential to do with what function it computes.

(SYM), as I have mentioned, encapsulates the declarative orientation of the logicists.¹⁸ These thinkers hold that AI's flashy robots of the future must have some means of representing part of the external world internally in a declarative fashion, and the means must be some symbol system, at minimum a first-order one, almost certainly a modal or higher-order one, and, I would say, probably a robust intensional logic of a sort not familiar to the detractors of strong logicism.¹⁹ Logicists with technical backgrounds in logic know, beyond a shadow of a doubt, that first-order logic is only the first, laughably primitive layer of the gargantuan to-be-devised formalism that is \mathcal{L}^T .²⁰

It is interesting to note that modifying (SYM) might give rise to part of the core of hyper-logicism (\mathcal{L}_H), as in

(SYM!) If (PBP), i.e. if a robot person S^* will be eventually produced by AIniks, then S^* must be such that *all* of the propositions ϕ_0, ϕ_1, \dots which are objects of S^* 's propositional attitudes are represented by formulas $\langle\phi_0\rangle, \langle\phi_1\rangle, \dots$ of some symbol system \mathcal{L}^T , where they can be processed according to the reasoning mechanism that is part of \mathcal{L}^T .

More about \mathcal{L}_H later. It's time now to attempt a definition of \mathcal{C}_S .

As many readers no doubt know, the AI world has of late been greatly energized by \mathcal{C} . (A superficial but engaging and readable introduction to \mathcal{C} , and some of the debate surrounding it, can be found in Graubard (1988). The classic, comprehensive study of \mathcal{C} in the context of the clash with which we are herein concerned, is Smolensky (1988a). For a more technical/theoretical view of connectionism and neural computing see Aleksander (1989). For a lively, logicist rallying cry, see Pinker and Mehler (1988). For a broad view of \mathcal{C} see Nadel *et al.* (1989). It is not easy to define strong connectionism (compare along these lines Hunter 1989)). (A moment ago, defining strong logicism *looked* easy enough, but the one guarantee about the game we have entered is that no definition will be attractive to everyone.) Defining \mathcal{C}_S will be quite a bit trickier than defining strong logicism: it will turn out, in fact, that 12 distinct versions of strong connectionism will be obtained as we pass through the promised dialectic. Here are the first two versions, both of which, for reasons to be given later, lay claim to being the starting place in an attempt to define \mathcal{C}_S :

\mathcal{C}_{S1}

(PBP _C)
(PER _{NN}) (OR (PER _{CA}), ...)
\neg (AI-F)
(CTT)
(ANA)
$\therefore \neg$ (CTT*)
\neg (SYM)

\mathcal{C}_{S2}

(PBP _C)
(PER _{NN}) (OR (PER _{CA}), ...)
\neg (AI-F)
(CTT)
(CTT*)
\neg (SYM)

To allay fears that I have either garbled the strong connectionist's position or have set up, in \mathcal{C}_{S1} and \mathcal{C}_{S2} , straw men, let me now justify these starting accounts with reference to statements made by proponents of strong connectionism.

There would seem to be no controversy about including in \mathcal{C}_{S1} and \mathcal{C}_{S2} , the two propositions (PBP_C) and (CTT). After all, (PBP_C) merely reflects optimism about the agent-oriented engineering side of \mathcal{C}_S : it says simply that strong connectionists will, down the road, give us generally intelligent, productive, flexible, ... robots. Likewise (CTT) would seem to be indisputable (as long as 'algorithm' is taken, as it usually is, to mean a *finite* procedure), and for this reason alone perhaps worth including in \mathcal{C}_{S1} and \mathcal{C}_{S2} . (Smolensky (1988, p. 7) explicitly affirms it.) But what about the other elements of \mathcal{C}_{S1} and \mathcal{C}_{S2} ? I consider them now, in turn.

What about (PER_{NN})? Why is *this* proposition included in \mathcal{C}_{S1} and \mathcal{C}_{S2} ? Here, once again, I don't think there will be much controversy. It is more than clear from the relevant literature that \mathcal{C}_S includes the view that human agents are to

be identified not with Turing machines and the like, but with computational creatures which have come to be known as neural networks. Sometimes these nets are picked out by a more fine-grained description (as, for example, in 'quasilinear dynamical systems'), but the phrase 'neural net' is certainly regarded by connectionists to more than serviceably reflect their commitment to computing architectures that are, in some sense, 'more brainlike' than Turing machines.

What, now, about $\neg(\text{AI-F})$? Why do I include this proposition in my first two attempts to specify \mathcal{C}_S ? The reason for including it is clear enough. In the previous quote, from* the connectionists (Churchland & Churchland 1990, p. 32), we see hints that (AI-F) is thought to be one of the *problems* with \mathcal{L} . In fact, (Churchland & Churchland 1990, p. 35) make things quite explicit:

The emerging consensus on [the] failures of [logistic AI] is that the functional architecture of classical SM (Turing) machines is simply the wrong architecture.

(AI-F) , to put it a bit barbarically, is the view that the *stuff* in which computation takes place is unimportant. And this is a view with which proponents of \mathcal{C}_S are clearly uncomfortable, since one of their main points seems to be that the *stuff does matter*:

[Connectionists] argue that intelligence will emerge only from a special hardware that reproduces the massive parallelism of the human brain, in which huge numbers of interconnected cells tackle different parts of the same task at the same time. ... Hardware is the essence of intelligence, says connectionism, and not only does traditional AI miss out on this fact, but it uses the wrong hardware. (Hurlbert and Poggio 1989)

I should point out here that, to forestall a premature objection, I do consider below rationales that might be given by the strong connectionist for *keeping* (AI-F) in a fleshing out of \mathcal{C}_S .

What, now, about (ANA) ? It is this thesis which differentiates \mathcal{C}_{S1} and \mathcal{C}_{S2} : the former account includes it, the latter does not. Why the split? The explanation is straightforward: on the one hand, it is clear from some of the connectionist literature that there *are* connectionists who affirm (ANA) , hence its inclusion in \mathcal{C}_{S1} . On the other hand, this affirmation is thought by many connectionists to be remarkable, perhaps remarkably imprudent—since the mathematical and empirical evidence is generally thought to weigh heavily against (ANA) . We will look, below, at some of this evidence. At this juncture I only want to make the point that there *is* a split on (ANA) , borne out by (at least partially) conflicting statements like the following two. First, the (ANA) supporting

I believe that ... there is a reasonable chance that connectionist models will lead to the development of new somewhat-general-purpose self-programming, massively parallel analog computers, and a new theory of analog parallel computation: They may possibly even challenge the strong construal of Church's Thesis [= our (CTT*)] as the claim that the class of well-defined computations is exhausted by those of Turing machines. (Smolensky 1988a, p. 3).

And then, on the flip side, the (CTT*) supporting 'By objecting to traditional AI approaches I am not disputing the notions of universal computation or the Turing machine results, which are established mathematically beyond doubt' (Waltz 1988, pp. 196–7).

What about $\neg(\text{SYM})$, the last member of \mathcal{C}_{S1} and \mathcal{C}_{S2} ? The negation of (SYM) can apparently be extracted from the classic and comprehensive (Smolensky 1988a), and also from remarks like:

The physical symbol system hypothesis ... is that a vocabulary close to natural language ... would be sufficient to express all concepts that ever need to be expressed. My belief is that natural language-like terms are, for some concepts, hopelessly coarse and vague, and that a much finer, “subsymbolic” distinction must be made, especially for encoding sensory inputs. (Waltz 1988)

I say that \neg (SYM) may *apparently* be extracted from such quotes as this one. For notice that (SYM) is more cautious than what the author of this quote is prepared for. (SYM) does not say that *all* concepts needed for the computation to underlie a robot agent are representable in \mathcal{L}^T . (And presumably whatever cannot be represented in \mathcal{L}^T cannot be processed in accordance with its proof-theoretic side, nor can whatever must go unrepresented be interpreted by \mathcal{L}^T 's semantic side. These facts are important to keep in mind when reflecting on (SUB–SYM), a thesis to arrive in a moment.) In light of this, it is easy enough to spell out more circumspect versions of \mathcal{C}_S . We can do so by replacing \neg (SYM) with the more constructive and cautious

(SUB–SYM) If (PBP), then S^* must be such that *some* of its mental processing involves subsymbolic encodings not representable in \mathcal{L}^T .

While (SUB–SYM) constitutes a denial of (SYM!), it is consistent with (SYM).

A parenthetical remark: yet another aspect of \mathcal{L}_H arises from a modification of (SUB–SYM), namely:

(NO–NUB) If (PBP), then S^* can be such that *all* of its mental processing involves subsymbolic encodings not representable in \mathcal{L}^T .

(Recall, in connection with (NO–SUB), my earlier remarks about \mathcal{L}_H 's commitment to the possibility of agents living like brains in vats.) In the same spirit as (NO–SUB), only in reverse, \mathcal{C}_H would presumably include

(SUB!) If (PBP), then S^* can be such that *all* of its mental processing involves subsymbolic encodings not representable in \mathcal{L}^T .²¹

To return to our objective of setting out \mathcal{C}_S , we have arrived at the following two candidates:

\mathcal{C}_{S3}

\mathcal{C}_{S4}

(PBP _C)
(PER _{NN}) (or (PER _{CA}), ...)
\neg (AI–F)
(CTT)
(ANA)
$\therefore \neg$ (CTT*)
(SUB–S _S YM)

(PBP _C)
(PER _{NN}) (or (PER _{CA}), ...)
\neg (AI–F)
(CTT)
(CTT*)
(SUB–SYM)

And note, with respect to our ‘big question,’ that

$\text{Inc}(\mathcal{L}_S \cup \mathcal{C}_{S3})$ and $\text{Inc}(\mathcal{L}_S \cup \mathcal{C}_{S4})$

A proponent of \mathcal{C}_S might propose either of two additional specifications of her camp, namely:

$$\mathcal{C}_{S5} = \mathcal{C}_{S3} \cup \{\neg(\text{SYM})\}$$

$$\mathcal{C}_{S6} = \mathcal{C}_{S4} \cup \{\neg(\text{SYM})\}$$

Obviously, since (among other reasons) $\text{Inc}(\mathcal{L}_S \cup \{\neg(\text{SYM})\})$, we have

$$\text{Inc}(\mathcal{L}_S \cup \mathcal{C}_{S5}) \text{ and } \text{Inc}(\mathcal{L}_S \cup \mathcal{C}_{S6})$$

But neither \mathcal{C}_{S5} nor \mathcal{C}_{S6} can lay valid claim to being a specification of *strong* connectionism, because these two specifications collapse into *hyper*-connectionism. In order to see this we have only to note the logical structure of (SYM), which, at the appropriate level of description (!)²², is

$$\phi(S^*) \rightarrow \exists x \psi(x, S^*)$$

Negating this yields by propositional logic

$$\phi(S^*) \wedge \neg \exists x \psi(x, S^*)$$

which in turn by quantifier shift becomes

$$\phi(S^*) \wedge \forall x \neg \psi(x, S^*)$$

And $\forall x \neg \psi(x, S^*)$ is a symbolization of the proposition that 'none of S^* 's mental processing involves subsymbolic encodings representable in \mathcal{L}^T , i.e. (SUB!), which is of course the hallmark of the \mathcal{C}_H . My aim herein is to pit \mathcal{L}_S against not \mathcal{C}_H , but \mathcal{C}_S , and this is reason enough for us to drop consideration of \mathcal{C}_{S5} and \mathcal{C}_{S6} . I should mention, however that there are *general* arguments afoot against \mathcal{C}_H . I believe, in fact, that these arguments are precisely what motivates weak connectionists (advocates, sticking with our code, of \mathcal{C}_W) to take the position they do. As the continuum of Figure 1 indicates, Harnad (1990) embraces \mathcal{C}_W ; his reasons for doing so are based on a respect for arguments in favour of \mathcal{L}_S which mix introspection and empirical evidence together in an interesting brew. Here, in Harnad's own words, is a sketch of one such argument:

Our linguistic capacities are the primary examples [of behaviour that appears to be symbolic in nature], but many of the other skills we have—logical reasoning, mathematics, chess-playing, perhaps even our higher-level perceptual and motor skills—also seem to be symbolic. In any case, when we interpret our sentences, mathematical formulas, and chess moves (and perhaps some of our perceptual judgements and motor strategies) as having a systematic meaning or content, we know at first hand that that's literally true, and not just a figure of speech. Connectionism hence seems to be at a disadvantage in attempting to model these cognitive capacities. (Harnad 1990)

I am not claiming here that Harnad's argument, or even his argument tidied up, is irresistible; I'm simply making the point that there is good reason to be suspicious of \mathcal{C}_H , and therefore reason to focus on the more plausible \mathcal{C}_S . (One might say that this paper itself displays precisely the kind of symbolic cognitive capacity Harnad has in mind.) So: we're in the business of sorting out the clash between \mathcal{C}_S and \mathcal{L}_S , and it's time to start doing just that. In the dialectic to come, the specifics of the inconsistency between \mathcal{L}_S and \mathcal{C}_S (as we have specified it so far) will be important.

5. The dialectic

I will initiate the promised dialectic with what we will call 'the starting argument,' in which the attempt is made to reduce \mathcal{C}_S to \mathcal{L}_S , and to thus dissolve (at least

one strain of) the \mathcal{C} – \mathcal{L} clash. The argument, put roughly to get us going, runs as follows.

The starting argument

If you're a proponent of strong connectionism, you are by definition an aspiring person-builder. That is why in both \mathcal{C}_{S3} and \mathcal{C}_{S4} the proposition (PER_{NN}) (or (PER_{CA}) ...) turns up. But this proposition, given that cellular automata are just k -tape Turing machines, and that neural nets are just probabilistic automata of the ordinary sort (Aleksander 1989), which are both in turn just *standard* Turing machines, amounts to (PER_{TUR}). But the proposition that persons are Turing machines is at the very heart of strong logicism. It would seem, then, that strong connectionism, to a significant degree, is reduced to strong logicism.

This is a hasty but nonetheless interesting little piece of reasoning. It is hasty because, to begin, clearly the last sentence may be somewhat hyperbolic, since even if ($\text{PER}_{\text{NN}} = \text{PER}_{\text{TUR}}$), \mathcal{C}_{S3} and \mathcal{C}_{S4} remain inconsistent. On the other hand, the starting argument is interesting because if its gist is correct, \mathcal{L}_S and \mathcal{C}_S (if $\mathcal{C}_S = \mathcal{C}_{S3}$ or $\mathcal{C}_S = \mathcal{C}_{S4}$) would be distinct only by virtue of some rather rarefied conflict—such as that involving Church-Turing theses. If nothing else, it would seem that the starting argument is something a strong connectionist would be obliged to rebut. So: is there anything fundamentally wrong with the argument? How can the strong connectionist respond cogently? In general I think there are two initially promising responses: (i) drop the proposition (PER_{NN}) (or (PER_{CA}), ...) from strong connectionism, or (ii) refine these propositions in such a way that there isn't such an intimate connection between 'neural net-agents' and Turing machine-agents.' I will consider these responses in turn.

But first we need to consider more carefully the slippery locution 'automaton x is just automaton y '—a locution at the heart of the starting argument. The instantiations of this locution in the starting argument, as when for example it is said that cellular automata are just k -tape Turing machines, are intended to be no different than capsule reports on, say, the equivalence of Register machines (for an introduction, see Ebbinghaus *et al.* 1984) and Turing machines. The idea here is that, ultimately, from the mathematical point of view, x and y , in the locution being considered, are the same creature: you could in principle specify both by the exact same set theoretic definition, starting from and never leaving the machinery of (say) ZFC.

Harnad (1990) has something to say on the issue here before us:

There is some misunderstanding of [the "fact" that neural nets fail to meet certain necessary conditions for a symbol system] because it is often conflated with a mere implementational issue: Connectionist networks can be simulated using symbol systems, and symbol systems can be implemented using a connectionist architecture, but that is independent of the question of what each can do qua symbol system or connectionist network, respectively. By way of analogy, silicon can be used to build a computer, and a computer can simulate the properties of silicon, but the functional properties of silicon are not those of computation, and the functional properties of computation are not those of silicon.

A proper analysis of this rather cryptic quote would require clarification of nothing less than the notions of functional properties, analogical arguments, and simulations. Such clarification is an impossibly tall order,²³ certainly given our space limitations. (If a connectionist were to appeal to Harnad's reasoning in support of her position, then presumably she would be obliged to provide such clarification.) What are the functional properties of pencils? Of pencils if some race who shun writing find them and use them to spin frisbees upon? Need simulation be two-way or can it just be one-way? Fortunately, such questions

need not detain us. I think it can rather easily be seen that Harnad isn't here threatening the key locution in the starting argument: suppose that we have an operator '[[]]' which when applied to a standard, 'user friendly' specification of an automaton or neural net, yields an account expressible exclusively in ZFC. Then what the starting argument appeals to is the unexceptionable²⁴ proposition that

(*) For every neural net N , there is a Turing machine M such that $[[M]] = [[N]]$

I take it to be obvious that for every x and y , if $x = y$, then x and y have precisely the same functional properties. Chicago's tallest building has the same functional properties, no matter what such properties amount to, as the Sear's Tower. This follows from Leibniz' Law. If this is correct, then Harnad's point, which, whatever else it may amount to, certainly hinges on the distinction between properties *simpliciter* and functional properties, evaporates.

So Harnad does not stop the starting argument from pestering the strong connectionist, when this argument is rendered in the more sophisticated form that appeals to '[[]]' and ZFC.²⁵ (I'm picking ZFC, what I'm saying could be expressed in terms of alternative set theories.) Let us return, then, to the two options I said the strong connectionist has in the face of this argument.

In the first response the strong connectionist simply holds that although it may be odd to refuse to take a stand on what a person, computationally speaking, is, and at the same time consider oneself a person builder, she will do it nonetheless. Moreover, she will say, she intended from the outset to drop propositions like (PER_{NN}) and (PER_{CA}). Her rationale for this is simple: in dropping (AI-F), she meant to drop the functionalist view that hardware doesn't matter; and she can't very well drop (AI-F) and *not* drop propositions like (PER_{NN}) and (PER_{CA}). And why is this? Why is it that (AI-F) is linked to propositions like (PER_{NN}) and (PER_{CA})?

Well, in a nutshell, both propositions about personhood assert, at bottom, that the structure determines mentality, not the stuff. And this slogan is really just functionalism encapsulated. Let me unpack a bit, on behalf of the strong connectionist, the claim that those who affirm theses like (PER_{NN}) and (PER_{CA}) are of necessity functionalists.

The 'flow chart functionalism' of Dennett (1978), i.e. (AI-F), says, intuitively, and by implication, that if you find a flow chart match between human brains and silicon-based Martian brains, then you can be assured that the human person and the Martian enjoy the same mentality. (See Figure 2.)

We have neither the space nor the time to discuss what most philosophers of mind and many cognitive scientists consider to be overwhelming empirical and theoretical support for functionalism.²⁶ But we can be clear about their position: John Pollock, an eminent functionalist, says

It seems preposterous to suppose that what a creature is made of can have anything to do with what mental states it is in, except indirectly, by influencing what structures can be built out of that stuff. It cannot be the stuff that a cognizer is made of that determines whether it can have mental states of a certain sort; it must instead be *how* it is made of that stuff, that is, how it is put together. (Pollock 1989)

Since this sort of attitude is what underlies (AI-F), and since, on this way of looking at things, (PER_{AUT}) and its refinements (in specifying the 'how put together' and the 'what stuff') amount to (AI-F), and since strong connectionism (at the moment) includes the *rejection* of (AI-F), it is only natural (or so the story

DENNETT'S "FLOW CHART" AI-FUNCTIONALISM

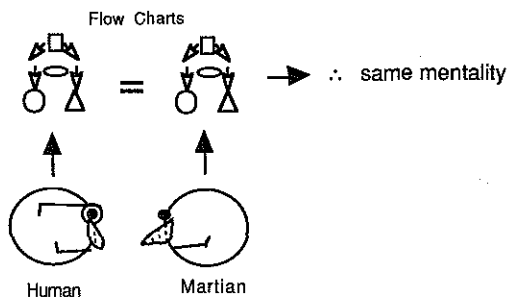
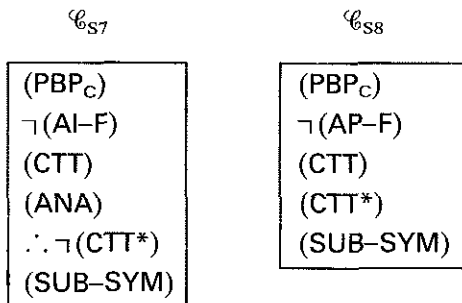


Figure 2.

under consideration goes, anyway) that \mathcal{C}_S not include (PER_{NN}). This gives rise to our next two versions of \mathcal{C}_S :



This concludes the first response to the starting argument. Is this response a solid one? Are either of \mathcal{C}_{S7} , \mathcal{C}_S tenable?

I do not think so. I don't think person builders can reject functionalism so easily, whether it is in the form of something like (AI-F), or (PER_{NN}) and (PER_{CA}). Here is why. Person, or agent, builders in AI, whether of the connectionist or logicist fold, are committed to certain techniques which, though hard to make precise, certainly include programming (or training) a high-speed computer-with-sensors. If someone managed to build a person by stirring up some fertile biological soup in the right way, that wouldn't spell success for AI person builders. To affirm (PBP) and/or its specific connectionist and logicist variants, is to say that certain 'computer' techniques will produce people. And, furthermore, the idea is not that by using these techniques you'll get lucky and bring a person into existence by a *side-effect* of what you have done.²⁷

But of course, for the person builders involved here, what the success of their techniques would show is not that people are, essentially, and in general, the *particular* computers these researchers are working on. There is no reason to think that the material, the stuff, the physical substrate—the particular computer and particular sensors, and the particular stuff out of which they are made—which compose the robot agents of the future is essential. Human persons, after all, are made of flesh, not silicon. So *some* sort of functionalist intuition is assuredly affirmed by proponents of (PBP) and its variants—some intuition

according to which people are *idealized* computers. This intuition, I submit, is, as (Haugeland 1986) recently suggests, captured generally and even elegantly by (PER_{AUT}), or at least, depending upon which side of the debate we are talking about, by refinement of it, i.e. (PER_{TUR}) for our logicians, and (PER_{NN}) and the like for our connectionists.

If the first response is no good, what, then, about the *second* response to the starting argument, to which I alluded above? This response is more subtle than its predecessor; it is based on the claim that though neural nets and cellular automata and *k*-tape Turing machines are one and the same when considered through the lens of operators like our '[]', when these automata are genuine *physical* entities in the *physical* world they are quite different; and their differences could be, from the standpoint of generating mentality, significant. Indeed, the situation here is as we might have expected: proponents of \mathcal{L}_S have held all along that the symbolic level is the *mental* level, with rule-based regularities that are independent of their physical realizations. Strong connectionists can reject this view by affirming the notion that physical realization *determines* mentality. This affirmation gives rise to our fourth version of strong connectionism, which is marked by the 'physical' version of the person-automata theses, as in

(PER_{PHYS-NN}) Persons are *physical* neural networks,

which produces (note the subscripts):

\mathcal{C}_{S9}

(PBP _C)
(PER _{PHYS-NN}) (or (PER _{PHYS-CA}))
\neg (AI-F)
(CTT)
(ANA)
$\therefore \neg$ (CTT*)
(SUB-SYM)

\mathcal{C}_{S10}

(PBP _C)
(PER _{PHYS-NN}) (or (PER _{PHYS-CA}))
\neg (AI-F)
(CTT)
(CTT*)
(SUB-SYM)

But the remarkable thing about this move is that it seems to entail just the sort of picture that impels many to *embrace* functionalism. In order to see this, consider the following situation. Suppose that we have a *physical* neural net, call it '*N**', that computes a set of functions Γ ; and suppose that *N** has been built out of stuff available in the physical world to strong connectionists. Suppose that this neural net is very complex, closer by far to real human brains than to standard textbook diagrams of multi-layer nets. And now suppose that, using *N** as a 'blueprint', we build a Turing machine *M** that computes all of Γ . If we had the time, we could specify how *M** is to be built from *N**. For example, suppose *N** is a 50 layer neural net, and that input neurons are 1000 in number; then we might want to build *M** as a 50-tape machine, with 1000 squares of the first tape used to hold the input that goes into *N**. And so on. (There is, I agree, a *lot* of toil involved in the 'so on'.)

Now, here is the crucial question: is it plausible to hold that while the immaterial set-theoretic versions of *N** and *M** amount to the same thing, i.e. that $\llbracket N^* \rrbracket = \llbracket M^* \rrbracket$, the *physical* versions do not? That *N** and *M** do not give rise to the same mental states (if in fact there *are* any in the picture)? I can't think of

a rationale supporting an affirmative answer to these questions. I do not see, then, how the second response to the starting argument has any force.

But I have moved too quickly. I have not looked carefully at \mathcal{C}_{S9} . For this view includes (ANA), and thereby includes a rejection of (CTT*). And rejecting (CTT*) allows one to hold that while our physical net N^* from above computes Γ , there is no physical Turing machine that can compute Γ ; and if there can be no such Turing machine, then our little thought-experiment involving N^* and M^* is all for naught. So we need to clarify the second response: the strong connectionist does not reject functionalism; she *affirms* it, and indeed affirms the refinements of it tied to persons, but also *constructively* embraces a positive thesis about what sort of *physical stuff* is of paramount importance—stuff that cannot be matched, functionally speaking, by any Turing machine. (This may very well be precisely the view expressed in Churchland & Churchland 1989.) This gives rise to yet another version of strong connectionism:

$$\mathcal{C}_{S11}$$

(PBP _C)
(PER _{PHYS-NN})
(AI-F)
(CTT)
(ANA)
$\therefore \neg(\text{CTT}^*)$
(SUB-SYM)

\mathcal{C}_{S11} entails that it is not really true that \mathcal{C}_S 's neural networks may be viewed as cellular automata. The aim of \mathcal{C} is to build a *genuine* neural net, and that has not yet been done. While set theoretic nets may be identified with classical architectures, not so for 'real life' physical nets. To this point nearly all neural nets have in fact been emulated on general purpose parallel machines. And while these parallel machines may be viewed as cellular automata, the bona fide neural nets of the future, so the advocate of setting $\mathcal{C}_S = \mathcal{C}_{S11}$ now says, will run on analog machines, and, courtesy of (ANA), these nets will not be things you can identify with cellular automata, and therefore will not be things you can in turn identify with Turing machines.

Have we arrived, then, in \mathcal{C}_{S11} , at a satisfactory version of \mathcal{C}_S ?

Well, if nothing else, \mathcal{C}_{S11} appears to reflect the current situation. As of 1990, nearly all neural networks are implemented on general purpose parallel computers—computers whose power is specified, mathematically, by cellular automata. Cellular automata, as we have noted, when viewed from the perspective of the foundations of mathematics, are exactly equal in power to Turing machines. Hence as of 1990 neural computers can be viewed as Turing machines, and it follows that today whatever can be done by a neural net can be done by an ordinary Turing machine. This was the cornerstone of the starting argument. But in light of this result our strong connectionist, holding to \mathcal{C}_{S11} , calmly proclaims that hardware *is* all-important in reaching AI's ultimate goals, not in the sense of contravening functionalism, not solely in the sense of moving toward 'brainlike' architectures; but hardware is all-important for the simple reason that we do not *really* have a neural net as long as we are forced to implement it on a

programmable, general purpose machine. We will have a *true* neural net, the strong connectionist continues, when and *only* when we implement a neural net which is isomorphic to that underlying the human brain on a *true analog machine*.

What are we to make of \mathcal{C}_{S11} ? Well, I am inclined to view the situation here as calling for a big application of *modus tollens*. That is, since I affirm (CTT*), and since \mathcal{C}_{S11} includes the *negation* of this proposition, I think \mathcal{C}_{S11} is simply false.

Now we have not the time to consider arguments for and against (CTT*). It is, I have intimated, at the very least inductively confirmed by the fact that we have never found a computing machine, whether analog or not, that is qualitatively superior to a Turing machine.²⁸ And while in principle a counter-example to (CTT*) is possible, no one takes this prospect seriously (as is evidenced by the fact that (CTT*) operates as a premise in canonical proofs, e.g. the standard proof that there is no Turing machine which, when given a formula ϕ from standard arithmetic, decides whether ϕ is true on the standard interpretation of arithmetic). There is also the fact, only recently noted, that Church's Thesis and its relatives may, in a strict sense in use in mathematics, be provable. (Mendelson 1990) has recently argued forcefully against the assumption that a proof connecting intuitive and precise notions is impossible. He has pointed out, among other things, that the proposition that the partial-recursive functions are effectively computable does seem to be amenable to proof. This is so because, as is well-known, the so-called initial functions are effectively computable, and the operations of substitution, recursion, and the least-number operator are known to lead from effectively computable functions to effectively computable functions. While this data may seem to constitute the basis of a proof of (CTT) rather than (CTT*), there is reason to think that Mendelson may have paved the way toward a proof of the latter. At any rate my overall point here is that since (CTT*) will be compelling for nearly all, our dialectic moves us of necessity to yet another version of \mathcal{C}_S , namely the streamlined:

$$\mathcal{C}_{S12}$$

(PBP _C)
(PR _{PHYS-NN})
(AI-F)
(CTT)
(CTT*)
(SUB-SYM)

But then the answer to our 'big question' ends up being one unpalatable to strong connectionists, namely

$$\text{Con}(\mathcal{L}_S \cup \mathcal{C}_{S12})$$

The bind that the proponent of \mathcal{C}_S finds herself in is, in summary, this: naive versions of \mathcal{C}_S include a denial of functionalism, and given that functionalism is very plausible, the only way that \mathcal{C}_S can remain viable is if it *includes* functionalism but embraces also the view that analog computing devices are qualitatively superior to non-analog devices, a view that is wholly unsupported by empirical and theoretical results.

Our terminus—Con ($\mathcal{L}_S \cup \mathcal{C}_{S12}$), and, given $\mathcal{C}_{S12} = \mathcal{C}_S$, therefore Con ($\mathcal{L}_S \cup \mathcal{C}_S$)—in no way marks a victory for \mathcal{L}_S or \mathcal{L} . For this terminus is a two-edged sword: if dialectic resulting from the starting argument reduces strong connectionism to strong logicism, it also works the other way around. After all, equivalence between neural nets and Turing machines underlies the starting argument, and this equivalence does not in and of itself favour strong logicism over strong connectionism. What is distinctive about \mathcal{C}_S and \mathcal{L}_S isn't much, and is easy enough to display:

$$\mathcal{L}_S - \mathcal{C}_S = \{(\text{PBP}_L), (\text{SYM})\}$$

$$\mathcal{C}_S - \mathcal{L}_S = \{(\text{PBP}_C), (\text{SUB-SYM})\}$$

Since (PBP_L) and (PBP_C) simply represent different articles of faith, and matters neither of logic nor science; and since (SUB-SYM) and (SYM) are compatible, the \mathcal{C}_S - \mathcal{L}_S clash evaporates.

A number of loose-ends remain; only three have I the space to take up, and only briefly at that, namely

Q1 What about consistency relations in the other permutations?

Q2 What about attempts on the part of AIniks to establish Inc ($\mathcal{C}_S \cup \mathcal{L}_S$)?

Q3 How could so many clever AIniks be so upset about the \mathcal{C} - \mathcal{L} clash if it's only a red herring?

I provide brief answers to Q1 and Q2 in this section; Q3 is covered in the next, and final, section.

The answer to Q1 can be encapsulated in this table:

	\mathcal{L}_W	\mathcal{L}_S	\mathcal{L}_H
\mathcal{C}_W	Con	Con	Inc
\mathcal{C}_S	Con	Con	Inc
\mathcal{C}_H	Inc	Inc	Inc

This table, among other things, blurs the difference between weak and strong connectionism—but it doesn't, of necessity, make it disappear. As the continuum of Figure 1 indicates, \mathcal{C}_W and \mathcal{L}_W are distinguished (from \mathcal{C}_S and \mathcal{L}_S) by explicit, practicable syntheses of logicist and connectionist techniques. (For the closest thing to a specification of such a synthesis, see Harnad 1990.) It is of course also clear from the table that affirming either of \mathcal{L}_H , \mathcal{C}_H secures inconsistency. The question of whether either of these camps is *plausible* is an issue for another paper. (But see (Bringsjord and Zenzen 1991).) But I think it is worth noting that paradigmatically 'symbolic' cognition (e.g. that associated with casting the \mathcal{C} - \mathcal{L} clash in the machinery of first-order logic; recall the Harnad quote above) casts some doubt upon \mathcal{C}_H , just as paradigmatically 'subsymbolic' cognition (e.g. hitting a baseball; recall the Pollock quote above concerning Q&I modules) casts some doubt upon \mathcal{L}_H .

Now for Q2: Smolensky (1988a) purports to demonstrate that

$$\text{Inc} (\mathcal{C} \cup \mathcal{L})$$

Here is how he attempts to pull it off: He begins (p. 5)²⁹ by distinguishing between two 'virtual machines', the 'intuitive processor' (responsible for all skilled performance; the traditional object of connectionist study) and the 'conscious rule interpreter' (responsible for (say) carrying out proofs before doing so becomes a

honed skill; the traditional object of logicist study). To ease exposition, let us give proper names to these two virtual machines, M_e for the intuitive processor, and M_x for the conscious rule interpreter. Now, two propositions (pp. 6–7) are at the heart of Smolensky's case for Inc ($\mathcal{C} \cup \mathcal{L}$), namely

- (8c) Complete, formal, and precise descriptions of M_e are generally tractable not at the symbolic level but only at the subsymbolic level.³⁰
- (10) Valid connectionist models are merely implementations, for a certain kind of parallel hardware, of symbolic programs that provide exact and complete accounts of behavior at the conceptual level.

Smolensky tells us (p. 7) that '(10) contradicts hypothesis (8c)'. Unfortunately, this is not so. (No appropriate symbolization of (8c) and (10) allows one to deduce a contradiction from their conjunction. In fact, (8c) and (10), as they stand, are provably consistent.) Charity dictates that we say Smolensky has grasped an *enthymematic* argument. In order to make his case, he must have in the back of his mind that (10) amounts to, or perhaps (when conjoined with some other proposition(s)) implies

- (10') There exists a complete, formal, precise, generally tractable symbolic description Δ of M_e ,

while (8c), rephrased, is

- (8c') There exists no complete, formal, precise, generally tractable symbolic description Δ of M_e .

Obviously, Inc $\{(10'), (8c')\}$. If we grant that (and I am prepared for the sake of argument to do so)

$$\mathcal{L}_S \rightarrow (10'), \mathcal{C}_S \rightarrow (8c')$$

then Inc ($\mathcal{C}_S \cup \mathcal{L}_S$). But, for one, is (8c') true? I don't think so. And in fact I would claim that anyone even remotely familiar with John Horton Conway's Game of Life³¹ would acknowledge that (8c') is false. (From which it follows by *modus tollens* on the second of the conditionals just introduced that \mathcal{C}_S is false! My suggestion in light of this, on behalf of the connectionist, would be that this conditional ought to be supplanted with $\mathcal{C}_H \rightarrow (8c')$. This lets things fall into place in accordance with our discussion above.) Life, as no doubt most readers know, involves a 2-dimensional cellular automaton evolving in discrete time whose cells, at any time t , are either ON or OFF, and their being so is determined by the following rule:

- (R) If exactly two of a cell C 's neighbours is ON at t , C remains unchanged at $t + 1$; if exactly three of a cell C 's neighbours is ON at t , C is ON at $t + 1$; otherwise C is OFF at $t + 1$.

Many entrancing computer simulations of Life are traveling around; perhaps the reader has seen one. But my point has nothing to do with the aesthetic aspects of Life. My point is simply that Smolensky's M_e could be described (perhaps only *imprecisely* described—we will get to this in a moment) as, and indeed *viewed* as, a cellular automaton in Life. Smolensky tells us (p. 6) that the state of M_e is a 'numerical vector evolving in time according to differential evolution equations'. Very well. Let $\langle M_e \rangle_t$ denote the state of M_e at time t . For the moment, assume

that time is discrete (modelled on \mathbb{N}), and that \Rightarrow is a transition function of the standard sort that drives classical automata, which captures (R). Then the evolution of M_e from t to $t + 4$ can be pictured like

$$\dots \Rightarrow \langle M_e \rangle_t \Rightarrow \langle M_e \rangle_{t+1} \Rightarrow \langle M_e \rangle_{t+2} \Rightarrow \langle M_e \rangle_{t+3} \Rightarrow \langle M_e \rangle_{t+4} \Rightarrow \dots$$

Suppose that I have an account like this that covers the *entire* 'life' of M_e , call this account Δ . Note, first, that Δ is surely a symbolic description of M_e , since Δ can be cast in terms of a k -tape Turing machine. If Δ is complete, formal, precise, and generally tractable, then (8c') is false, and Smolensky doesn't make his case. *Does* Δ have these properties? Well, Δ is, on any reading of the term, *formal*, that much is clear. Is Δ precise? Extremely so, I should think. Δ , after all, is expressed in the rather austere language of elementary computability theory, i.e. naive set theory (and could, as we have noted above, be re-expressed in the utterly precise language of axiomatic set theory). Surely talk of k -tape Turing machines is as precise as talk, *via* differential equations, of dynamical systems. The whole issue thus seems to boil down to whether or not Δ is complete and generally tractable. Smolensky would doubtless say that Δ has neither of these properties. Is he right? I don't think so.

Smolensky may say that Δ isn't complete, because it includes only discrete 'snapshots' of the continuous entity M_e . But suppose that Δ includes snapshots $\langle M_e \rangle_t$ and $\langle M_e \rangle_{t+1}$, and that Smolensky is concerned with what is left out here, i.e. with the states of M_e between t and $t + 1$. His concern is easily handled: one can make the interval of time during which the state of M_e is ignored arbitrarily small. (This fact, and indeed many of those that undergrid my reaction to Smolensky, is of course stock stuff.) Δ , or at least a refinement of Δ (call it ' Δ^* '), would therefore seem to be complete.

We are left, then, with this question: Is Δ^* generally tractable? (Smolensky 1988b, p. 64 seems to agree that this is indeed the key question: he claims that two of his detractors, Dietrich and Fields 1988, ignore the question.) Unfortunately, it is far from clear how this question ought to be interpreted. Let M_{Δ^*} be the Turing machine corresponding to Δ^* . Then what *is* clear is that when M_e performs some interesting computation (say that corresponding to the catching of a baseball by an outfielder), it might take a primitive physical instantiation of M_{Δ^*} a zillion years to carry out this same computation. On the other hand, there is no reason to think that (e.g.) Δ^* could not be run on a connection machine, or parallelized on 3,000,000 supercomputers, or run on a descendant of the connection machine in the year 2856—in which case it may be that Δ^* is quite tractable.³² Now it may be that Smolensky's claim that Δ^* is generally intractable amounts, instead, to the claim that no human can work with (grasp? manipulate? ...) Δ^* . There is some textual evidence for thinking that this is in fact his claim, for in replying to Dietrich and Fields (1988), Smolensky says:

The question is whether such accounts [= e.g. our Δ^*] exist in sufficiently tractable form to serve the scientific needs of building models, making predictions, and providing explanations. (Smolensky 1988b, p. 64)

Smolensky's claim is apparently that Δ^* is *not* (to use his new terminology) 'sufficiently tractable' in the sense of serving 'scientific needs'. As support for his claim, he may cite the fact that not even in Life can a 'player', with but pencil and paper and a good mind, chart precisely the evolution of even a moderately complex cellular automaton. And here he is certainly right. But the counterpart

to this point will apply equally well to the connectionist studying M_c : no connectionist, with but pencil and paper and a stellar mind for differential equations (etc.), can chart precisely the evolution of even a moderately complex dynamical system. In both cases, certain 'aids' are required—calculators, high-level programming languages, training strategies, and so on. In the case of Life, some aids have turned out to be certain 'high-level' descriptions in terms of 'eaters', 'gliders', and the like (see (Dennett 1991) for an interesting discussion of this level and others in Life). Smolensky gives us no reason for thinking that aids for rendering Δ^* in a form that facilitates scientific needs are nowhere to be found. It is perhaps worth nothing that every time I click on my machine to do something scientific, I am *employing* such aids.

I conclude, then, that (8c'), if not simply false, is at least highly doubtful. More generally, I conclude that the \mathcal{C}_S - \mathcal{L}_S clash is indeed a red herring, and that therefore, if \mathcal{C}_H and \mathcal{L}_H are, as they seem to be, problematic, what might be called 'Ecumenical AI'³³ is in order.

6. Speculative sociology

I will end by briefly tackling the question

Q3 How could so many clever AI-niks be so upset about the \mathcal{C} - \mathcal{L} clash if it is only a red herring?

My compressed answer to Q3 is this: because each side of the debate reflects the nature of the thinking and expertise that the members of this side bring to bear on the problems of AI. To put it crudely, *before allegiance to one side in the \mathcal{C} - \mathcal{L} clash is declared*, logicists are at home, and in many cases have been for most of their adult lives, with cognition that *appears* to be symbolic, and they are more likely to be immersed, professionally, in the development and specification of logics than in (say) quasilinear systems. The opposite, it seems to me, holds of connectionists: *before allegiance to one side in the \mathcal{C} - \mathcal{L} clash is declared*, they are more likely have backgrounds and intellectual investments in neuroscience, 'continuous mathematics', and so on. My hypothesis here at least has the virtue of being decidable by way of questionnaire.

Connectionism, in general, seems especially suited to perceptual, motor, and, to use Jerry Fodor's term, 'modular' systems, all of which correspond to identifiable, well-defined regions of the brain, all of which appear to be in some sense subsymbolic, and all of which will be (even the logicists are inclined to agree) profitably replicated on a brainlike machine. It is no accident that connectionists say such things as:

There is ... a large class of computations for which the brain's architecture is the superior technology. These are the computations that typically confront living creatures: recognizing a predator's outline in a noisy environment; recalling instantly how to avoid its gaze, flee its approach or fend off its attack; distinguishing food from nonfood and mates from non-mates; navigating through a complex and ever-changing physical/social environment; and so on. (Churchland & Churchland 1990, p. 36)

What about the logicists? What is their stomping ground? Their forté, traditionally and independent of the clash in question, is the kind of thinking that appears to have very little to do with well-defined neural structures, and everything to do with the kind of thing that one can do while in a sensory deprivation tank, or do limbless, paralysed, and all alone. Paradigmatic cases would seem to be technical philosophers engaged in belief fixation on the basis of competing, formal

arguments—attempts to resolve logical paradoxes, to devise rigorous deductive arguments for God's existence or non-existence, or for the permissibility or impermissibility of abortion, and so on. Many logicians devote large parts of their intellectual lives to activity that seems to be (SYM)-confirming.

If nothing else, logicians and connectionists seem to practise what they preach.

References

- Aleksander, I. (1989) *Neural Computing Architectures* (Cambridge, MA: MIT Press).
- Andersen, A. C. General intensional logic. In D. Gabbay and E. Guenther (eds) In *Handbook of Philosophical Logic*, vol. II (D. Reidel).
- Ashby, W. R. (1952) *Design for a Brain* (Chapman and Hall).
- Bechtel, W. (1988) Connectionism and rules and representation systems. *Philosophical Psychology*, 5–15.
- Bringsjord, S. and Zenzen, M. (1991) In defense of hyper-logician AI. *IJCAI 1991*, 1066–1072, Sydney, Australia (Morgan Kaufman).
- Bringsjord, S., (forthcoming) *What Robots Can and Can't Be*. Cognitive Science Series (Kluwer Academic Publishers).
- Charniak, E. and McDermott, D. (1985) *Introduction to Artificial Intelligence* (Addison-Wesley).
- Churchland, P. M. and Churchland, P. S. (1990) Could a machine think? *Scientific American*, January, 32–37.
- Courant, R. and Robbins, H. (1941) *What is Mathematics?* (London, Oxford University Press).
- Dennett, D. (1991) Real patterns. *Journal of Philosophy*, 91: 27–51.
- Dennett, D. (1984) Cognitive wheels: the frame problem of AI. In C. Hookway (ed.) *Minds, Machines & Evolution* (Cambridge University Press), 129–152.
- Dennett, D. (1978) *Brainstorms* (Bradford Books).
- Dewdney, A. K. (1984). Analog devices. *Scientific American*, 19–26.
- Dietrich, E. and Fields, C. (1988) Some assumptions underlying Smolensky's treatment of connectionism. *Behavioral & Brain Sciences*, 11: 29–31.
- Doyle, J. (1988) Big problems for artificial intelligence. *AI Magazine*, (Spring) 19–22.
- Ebbinghaus, H. D., Flum, J. and Thomas, W. (1984) *Mathematical Logic* (New York: Springer Verlag).
- Elman, J. (1989) Representation and Structure in Connectionist Models, Technical Report CRL 8903, Center for Research in Language, VCSD.
- Fodor, J. (1980) Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral & Brain Sciences*, 3: 63–109.
- Fodor, J. and Pylyshyn, Z. (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28: 3–71.
- Gardner, M. (1970, 1971) His mathematics column, "Mathematical Games", *Scientific American*, October, February, 120–123, 112–117.
- Garson, J. (1990) Cognition without classical architecture. (unpublished manuscript).
- Gold, E. M. (1965) Limiting recursion. *Journal of Symbolic Logic*, 30: 28–38
- Graubard, S. (1988) *The Artificial Intelligence Debate* (MIT Press).
- Harnad, S. (1990) The symbol grounding problem. *Physica D*, 42: 335–346.
- Haugeland, J. (1986) *Artificial Intelligence: The Very Idea* (Cambridge, Massachusetts, Bradford Books, MIT Press).
- Hillis, D. W. (1989) Intelligence as an emergent behavior; or, The songs of Eden. In S. Graubard (ed.), *The Artificial Intelligence Debate* (MIT Press), 175–190.
- Horgan, T. and Tienson, J. (1989) Representations without rules. *Philosophical Topics*, 15.
- Hurlbert, A. and Poggio, T. (1989) Making machines (and artificial intelligence) see. In S. Graubard (ed.), *The Artificial Intelligence Debate* (MIT Press), 213–240.
- Hunter, L. E. (1989) Some memory, but no mind. *Behavioral and Brain Sciences*, 11–37.
- Kaplan, S., Weaver, M. and French, R. (1990) Active Symbols and Internal Models: Towards a Cognitive Connectionism. (unpublished).
- Kugel, P. (1986) Thinking may be more than computing. *Cognition*, 22: 137–198.
- Kugel, P. (1990) Is it Time to Replace Turing's Test? 1990 workshop 'Artificial Intelligence: Emerging Science or Dying Art Form,' sponsored by SUNY Binghamton's Department of Philosophy's Program in Philosophy and Computer & Systems Sciences, and AAAI.
- Laird, J. E., Newell, A. and Rosenbloom, P. (1987) SOAR: an architecture for general intelligence. *Artificial Intelligence*.
- Mendelson, E. (1990) Second thoughts about Church's Thesis and mathematical proofs. *Journal of Philosophy*, May, 225–233.
- Nadel, L., Cooper, L. Culicover, P. and Harnish, R. (1989) *Neural Connections, Mental Computation* (MIT Press).
- Nilsson, N. and Genesereth, M. (1988) *Logical Foundations of Artificial Intelligence* (Morgan Kaufman).

- Pinker, S. and Mehler, J. (1988) *Connections and Symbols* (MIT Press).
- Poellock, J. (1989) *How to Build a Person: A Prolegomenon* (Cambridge, MA: Bradford Books).
- Poundstone, W. (1985) *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge* (Morrow).
- Putnam, H. (1981) *Brains in a Vat. Reason, Truth and History* (Cambridge University Press).
- Putnam, H. (1965) Trial and error predicates and the solution of a problem of Mostowski. *Journal of Symbolic Logic*, 20: 49–57.
- Rey, G. (1986) What's really going on in Searle's 'Chinese Room'. *Philosophical Studies*, 50: 169–185.
- Rumelhart, D. and McClelland, J. (1986) *Parallel Distributed Processing* (MIT Press).
- Schwartz, J. (1988) The new connectionism: developing relationships between neuroscience and artificial intelligence. In S. Graubard (ed.), *The Artificial Intelligence Debate* (MIT Press), 123–142.
- Searle, J. (1980a) Author's response. *Behavioral & Brain Sciences*, 3: 450–457.
- Searle, J. (1980b) Minds, Brains, and Programs. *Behavioral & Brain Sciences*, 3: 417–424.
- Searle, J. (1982) Review of Hofstadter and Dennett. *New York Review of Books*, 3–10.
- Servan-Schreiber, D., Cleeremans, A. and McClelland, J. (1988) Encoding Semantical Structure in Simple Recurrent Nets, Technical Report CMV-CS-88-183, Department of Computer Science, Carnegie Mellon University.
- Smolensky, P. (1988a) On the proper treatment of connectionism. *Behavioral & Brain Sciences*, 11: 1–22.
- Smolensky, P. (1988b) Putting together connectionism—again. *Behavioral & Brain Sciences*, 11: 59–70.
- Turkle, S. (1984) *The Second Self* (Simon & Shuster).
- Waldrop, M. M. (1984) Artificial intelligence in parallel. *Science*, 225: 608–10.
- Waltz, D. (1988) The prospects for building truly intelligent machines. In S. Graubard (ed.), *The Artificial Intelligence Debate* (MIT Press), 191–212.
- Zalta, E. N. (1988) *Intensional Logic and the Metaphysics of Intentionality* (MIT Press).

Notes

1. A very remote ancestor of this paper was presented at a 1990 workshop entitled 'Artificial Intelligence: Emerging Science or Dying Art Form.' This workshop was sponsored by SUNY Binghamton's Department of Philosophy's Program in Philosophy and Computer & Systems Sciences, and by AAAI. I am indebted to those at the workshop who made trenchant comments on this ancestor, especially John Sowa, Jim Hendler, Eric Dietrich (who kindly offered further insightful comments on the predecessor to the predecessor to (!) this final draft), and William Rapaport. I am especially indebted to a participant from this conference who anonymously refereed a previous version of this paper: this referee's criticisms and suggestions, the majority of which led to an improvement in the paper, are too numerous to cite on a case by case basis. Even those criticisms with which I heartily disagreed (and in some cases still do), were trenchant, and some of them would be worth devoting self-contained papers to. I am, finally, also indebted to those connectionists at Rensselaer Polytechnic Institute who have done so much to enlighten this die-hard logician about their views, most notable among which is Michael Skolnick.
2. The phrase here is Roger Schank's, made in reference to 'parallel processing' (Waldrop 1984). It's a phrase Smolensky (1988a, p. 7 and note #5, p. 23) mentions in connection with the connectionist–logicist clash; and he goes to great pains to try to show that this clash *isn't* a red herring. Like Smolensky (1988a), I don't know if Schank was in particular referring to the connectionist–logicist clash, and nowhere in this paper are the grounds for his phrase addressed.
3. It may be suggested that the connectionist–logicist clash be viewed as the question of which position better satisfies a set of 'soft constraints', i.e. as a kind of optimization problem rather than an 'I'm right, you're wrong' problem. There are at least two barriers to viewing the problem this way: (i) Given that the clash is inevitably in large part about the truth or falsity of propositions at issue, optimization would in turn inevitably reduce to questions about how many and which (etc.) propositions are truth or false on both sides of the clash. But these questions would give rise to an 'I'm right, you're wrong' problem once again, and progress on the clash would thus turn out to be illusory. (ii) Proponents on both sides of the clash generally treat the clash as an 'I'm right, you're wrong' problem.
4. I assume, to ease exposition, that specifications of \mathcal{C} and \mathcal{L} come in k opposing pairs, where the opposition is (at least on the surface) short of outright contradiction.
5. I believe there are some versions of connectionism which may in fact entail type physicalism, which by the (perhaps misguided) lights of nearly all philosophers of mind is an untenable theory. This is an issue to be tackled on another day.
6. It would be nice, but is by no means necessary, if readers have worked through Smolensky (1988a) and the peer review and reconstruction that followed it in *Behavioral & Brain Sciences*.
7. My focus on, and explication of, symbols systems should not be allowed to obscure the uncontroversial distinction between how a computer program is *analysed* and how it *computes*.

I'm *not* saying that all logicist programs are in fact invariably analysed by AIniks as symbol systems; I *am* saying that, at bottom, mathematically speaking, all logicist programs are in fact implemented symbol systems (of varying type). And on the connectionist side: I'm not saying that all connectionist systems are invariably analysed as systems of differential equations with calculus machinery; I *am* saying that connectionist systems, formally speaking, are implemented systems of differential equations (of varying type).

8. The first operation is to cut, say, $n = 10$ pieces of spaghetti proportional to the numbers to be sorted. This operation seems to require only a time proportional to n . Next, holding the strands in one hand, slam the ends down on table, aligning one end of each strand. This is (so the story goes) a single operation. Finally, holding the aligned strands in one hand, simply remove them individually in turn, going from the tallest to the shortest, measuring each and recording the result as they are picked. We are finished with our sorting in what is *apparently* linear time.
9. As good a place as any to concede that the continuum probably isn't one-dimensional.
10. Along these lines, see (Dennett 1984).
11. PBP is a mnemonic for 'Person Building Project.'
12. Readers reluctant to allow even an ontology which includes a generic, non-question-begging, exposition-easing concept of an agent are free to (e.g.) paraphrase (PER_{TUR}) as some such thing as 'Some computational theory of the Turing machine variety is the best way to explain human cognition'. The tactic of paraphrasing would have to be performed *uniformly* throughout this paper—which would lead, I think, to some rather cumbersome locutions which are, I think, rather harmlessly compressed in the ways I have chosen.
13. Physicalist versions of the (PER₂) theses will of course have the drawback that they by definition rule out, or perhaps beg the question against, agent dualism. Agent dualism is at least taken seriously by a number of philosophers, some of whom work on the foundations of AI and Cog Sci (e.g., Pollock 1989).
14. Consider, for illumination of this point, my attack on (PER_{AUT}) from free will, carried out in Chapter VIII of *What Robots Can and Can't Be* (forthcoming in Kluwer Academic's Cognitive Science Series.) In that chapter I capitalize on the fact that automata cannot compute the so-called Busy Beaver function (a well-known uncomputable function). If my argument there is any good, an appeal to sensors will accomplish nothing, because sensors don't (can't!) increase the baseline computing power of automata.
15. There's an important thesis that I leave off our list herein, namely

(ROB) AIniks will eventually build a robot whose observable behaviour is generally indistinguishable from the observable behaviour of human persons.

The conjunction of \neg (PER_{AUT}) and (ROB) seems to me to be an interesting position. It's been my experience that many in AI haven't entertained it. Those who are especially sanguine about what robots will in the future *do*, are, it seems, invariably sanguine about what they will (in some deep sense) *be*, and so are affirmers of something like (PER_{AUT}). One might reasonably suspect that this is due to a hasty conflation of person-like *behaviour* with *personhood*.

16. I use the generic term 'associated' so as not to beg any metaphysical questions (e.g.) against the agent dualist or physicalist. I apologize for the fact that the wording here is cumbersome, but I'm choosing awkwardness over begging questions.
17. It might be said that 'some' is too cautious here (and also too cautious in (SUB-SYM)), since (so the claim goes) Smolensky's conscious rule interpreter and Pollock's Q&I modules already give this much away. This claim is mistaken, for a number of reasons. First, in the case of Pollock, the claim is clearly incorrect, as is born out by close reading of the above-presented quote from him on (among other things) Q&I modules: that quote implies that Q&I modules do *not* enter into the kind of occurrent deliberation (see (SYM)) required for personhood. In the case of Smolensky: his position is that the conscious rule interpreter is completely dispensable in favour of the intuitive processor (see my detailed discussion of Smolensky's views later in the paper). Given this, and given that (SYM) is making the claim that the objects of S^* 's occurrent deliberations *just are* represented ... (as opposed to being such that they can be *viewed* by AIniks as representing ...), it's clear that Smolensky has hardly conceded (SYM). Third, the most charitable construal of the claim that leads to this note may be calling for a quantifier between \exists and \forall (at the moment, \exists is employed). Developing such a quantifier, and using it in the context of the \mathcal{C} – \mathcal{L} debate, would presumably enlarge the aforementioned continuum, and certainly would embroil us in issues that simply can't be covered herein for lack of space. (For a discussion of issues involving the use of \forall in (SYM)ish theses, at the juncture under scrutiny, see Bringsjord & Zenzen 1991.)
18. (SYM) is ubiquitous in the strong logicist literature. See, for example, Nilsson & Genesereth (1988), Pollock (1989), and Charniak & McDermott (1985).
19. Readers wanting to enquire into the plausibility of Leibniz's dream that all of natural language be formalized can start by introducing themselves to intensional logic; Montague, after all, had the same dream, and took, via the core of the system now known as IL, appreciable steps towards

its realization. A good place for a technical philosopher to start is Andersen (1984). I also highly recommend Zalta (1988).

20. I have found that mathematically mature AIniks tend to be strong in the standard foundations of theoretical computer science, but astonishingly weak in the rudiments of philosophical logic. I know some rather illustrious and seemingly technically sound AIniks who are unaware of standard motivators of intensional logic, such as the following puzzle. The English sentence 'Ponce de Leon searched for the fountain of youth,' if symbolized in first-order logic, might be 'S1f,' where 'S' is a two-place predicate letter standing for '___ searched for ___,' and '1' is a constant denoting Ponce de Leon, and 'f' is a constant denoting the fountain of youth. But by the rule of Existential Generalization of first-order logic, 'S1f' gives rise to '∃xS1x.' But '∃xS1x.' is false, since (if we assume that Ponce de Leon didn't go on any other wild goose chases), it's false that there exists something such that Ponce de Leon searched for it.
21. Or change with the equivalent 'none of its processing involves subsymbolic encodings representable in \mathcal{L}^T .'
22. In the proof-sketch that follows I quantify only over propositions to be captured by \mathcal{L}^T , 'S*' is an arbitrary constant denoting the robot agent of the future upon which our entire discussion is predicated, and $\phi(x)$ denotes a first-order formula in which x occurs.
23. That there is no certified logic of analogical reasoning makes this so alone.
24. The coming proposition, (*), isn't threatened in the least by the fact that, to use the connectionist slogan (compare, for example Smolensky 1988a, 1989), *neural nets are dynamical systems, not von Neumann machines*. It is exceedingly hard to fathom such statements as

The mathematical category in which [dynamical systems] live is the continuous category, not the discrete category, so we have a different kind of mathematics coming into play.

if 'different kind of mathematics' is supposed to be taken seriously. Classical mathematics incorporates not only the differential equations near and dear to the heart of strong connectionists, but symbol systems near and dear to the heart of logicians. Furthermore, it's well-known that problems solved by way of differential equations can be solved, in principle, using only a first-order language. Students taking an undergraduate course in mathematical logic can be fairly called upon, rather early in the course, to formalize, using a first-order language, the continuity of a function ρ on \mathbb{R} .

25. The proof in first-order logic that (PER_{NN}) and (PER_{TUR}) are interderivable is trivial given (*). Here is a Fitch-style proof of (PER_{NN} ⊢ (PER_{TUR})):

1	$\forall x(Px \rightarrow \exists y(Ny \wedge x = y)) = (\text{PER}_{NN})$	assump
2	$\forall x(Nx \rightarrow \exists y(My \wedge x = y))$	(*)
3	$Pa \rightarrow \exists y(Ny \wedge a = y)$	1
4	$Na \rightarrow \exists y(My \wedge a = y))$	2
5	Pa	assump
6	$\exists y(Ny \wedge a = y)$	3, 5
7	$Nb \wedge a = b$	assump
8	Nb	7
9	$a = b$	7
10	Na	8, 9
11	$\exists y(My \wedge a = y))$	10, 4
12	$\exists y(My \wedge a = y))$	6, 7-11
13	$Pa \rightarrow \exists y(My \wedge a = y))$	5-12
14	$\forall x(Px \rightarrow \exists y(My \wedge x = y)) = (\text{PER}_{TUR})$	13

26. Part of the driving force behind (AI-F) comes from observations (made by Pollock 1989) like the following:

- (O1) An everyday object of kind T (a can opener, a car, etc.) could be made of different matter and yet still qualify as a T .
- (O2) There are significant anatomical differences between the brains of different persons, borne out, for example, by the fact that though linguistic functions are normally represented in the left hemisphere of right-handed persons, insult to the left hemisphere can lead to the establishment of these functions in the right hemisphere. In fact, brains suffering massive insult occasionally display an almost miraculous plasticity which allows their owners to reclaim mental capacity.
- (O3) Eventually it will be possible to replace human eyes, and indeed even parts of human brains, with equivalently operating circuitry.
- (O4) Different computers can run the same programs while exhibiting important hardware differences.

27. There are those whose ultimate aim in AI is this side-effect—thinkers who hope to build a computational device whose structure is appropriate for 'ensoulment', a device alongside which

a person, an immaterial entity, will pop into existence and connect up with the device in some fruitful way (for a glance at such thinkers see Turkle 1984). These thinkers aren't our concern herein. We're concerned with those who think that AI techniques are near the *essence* of personhood, or mindedness, itself.

28. By the phrase 'we have never found' I mean to imply that the machines in question are all in some sense *useable*. So-called trial-and-error machines (introduced independently by Putnam 1965, Gold 1965, and discussed logically and philosophically by Kugel 1990, 1986) are capable of, for example, solving the halting problem. It's highly doubtful, however, that such machines can be used.
29. All page numbers in this sub-discussion to refer to Smolensky (1988a).
30. So as to keep our discussion of Smolensky here in line with the vocabulary introduced and employed above, I take the liberty of changing 'conceptual' to 'symbolic' and 'subconceptual' to 'subsymbolic.'
31. Gardner (1970, 1971) introduced the game to a wide audience. Poundstone (1985) provides a 'deeper' look at the game and its philosophical implications.
32. Lest it be thought that the point I've just made is trivial, I would point out that the trivial is not always consciously affirmed even by the learned. Two connectionists (who shall go nameless) have told me in conversation that though the main thesis of the present paper is substantiated, their brand of connectionism is immune. Their brand? That ultra-high speed parallel computation on an analog device will lead to revolutionary advances in AI. My point here (missed by these connectionists, I'm afraid) is that logicians are already salivating over super-duper hardware. The thought of implementing even restricted first-order circumscription on anything less than a connection machine should be banished by reflex.
33. Expressed by Ashby (1952), and, more recently, by Dietrich & Fields (1988).