ELSEVIER

Artificial Intelligence 80 (1996) 171–191

Artificial Intelligence

Response to my critics

Hubert L. Dreyfus*

Philosophy Department, University of California at Berkeley, Berkeley, CA 94720, USA

Received June 1995

After reading my critics and talking to many present and former AI researchers, I think that my characterization of what John Haugeland calls Good Old-Fashioned AI (GOFAI) as a degenerating research program pursued only by a few "die-hards" was inaccurate. There are really at least three different rather diffuse research programs. The first, and so far most important, is a contribution to cognitive science which found its expression in Newell and Simon's deservedly famous paper on Physical Symbol Systems. It is dedicated to testing the hypothesis that "A physical symbol system has the necessary and sufficient means for general intelligent action" [15, pp. 41 and 49]. But this is not a unified research program. There are several programs dedicated to understanding the human mind as a physical symbol system: Newell's group at Carnegie-Mellon with SOAR, Minsky at MIT with frames, and John McCarthy at Stanford with his logic-based models, to name a few. Each group thinks, or at least until recently thought, that their program was on the most promising track. But none has made enough progress to convince anybody outside their school or group to join them.

Then there are the AI engineers dedicated to making machines behave intelligently at specific tasks regardless of how human being do it. As Jerry Feldman put it in a recent talk here at Berkeley: "AI no longer does Cognitive Modeling. It is a bunch of techniques in search of practical problems." Finally, for many, the work in GOFAI has shifted away from the Newell/Simon program to work on integrated architectures that combine high-level, symbolic problem solving with connectionist models of perception and action. This approach preserves being a physical symbol system as a necessary condition for being intelligent but abandons it as a sufficient condition. But this approach is problematic since it is not clear how the part of the system solving problems using symbolic representations is supposed to talk to the other parts of the system. As Haugeland points out, the very idea that the intellect sends symbolic instructions

* E-mail: dreyfus@cogsci.berkeley.edu.

SSDI 0004-3702(95)00088-7

to the body which then executes these orders may well be a mistaken way of conceiving the relation of intelligence to the world.

Now there is a rival to symbolic representation that has attracted many researchers and graduate students, neural networking modeling. There is some confusion about how radical this departure from symbolic AI is, because neural nets can be simulated with algorithms on digital computers and in a loose sense algorithms use symbols, but the symbols in the GOFAI project were not just strings of bits that represented some state of the computer but were supposed to be semantically interpretable as representing features in the world. Of course, the neural networks are still representations in some sense, but they represent by way of the weights on the connections between simulated neurons and these do not have the properties of precise, context-independent symbols. Even Newell acknowledges that:

> You can, in one sense, say that connectionist systems are systems that are nonsymbolic and see how far you can go when you push them to do the same tasks that have been done with symbol manipulation without their becoming symbolic systems. There is certainly an opposition between the physical symbol system hypothesis and connectionism. [16, p. 153]

Daniel Dennett spells out clearly what this means:

> If you look at the nodes in a connectionist network . . . some of them seem to have careers, suggesting that they are particular symbols. This is the symbol for "cat," this is the symbol for "dog." It seems likely to say that whenever cats are the topic, that symbol is active; otherwise it is not. Nevertheless, if you make the mistake of a simple identification of these nodes—as cat symbol and dog symbol—this does not work. Because it turns out that you can disable this node, and the system can go right on thinking about cats. Moreover, if you keep the cat and dog nodes going and disable some of the other nodes that seem to be just noisy, the system will not work. The competence of the whole system depends on the cooperation of all its elements, some of which are very much like symbols. . . . At the same time, one can recognize that some of the things that happen to those symbols cannot be correctly and adequately described or predicted at the symbol level. [5, pp. 63–64]

Paul Smolensky gives a detailed argument that networks work on a subsymbolic level and are related to symbolic systems as quantum physics is related to Newtonian physics [20].

Whether the shift to network research continues will depend on whether those defending GOFAI are making slow but steady progress as McCarthy claims—and presumably workers on SOAR claim too—or whether what now looks like slow progress comes to look like the diminishing returns that show one is trying to solve the wrong problem. Right now, whether one calls the current situation in GOFAI a degenerating research program left to the die-hards or a winnowing out of the faint-hearted so that only the courageous visionaries remain is largely a

question of rhetoric and, as some of the responses to *What Computers Still Can't Do* show, generates more heat than light.

For the answer finally to become clear, what certainly should be avoided on both sides is the unscientific ploy of never admitting failure even when one fails to achieve one's research goals. Douglas Lenat, whose Cyc project for giving a computer commonsense, Simon, McCarthy and I take to be an important contribution to GOFAI, unfortunately still follows the old approach of rewriting his goals and his time-table so as to be able to claim that his research project is right on schedule and just about to succeed. Eleven years ago Lenat predicted that in 10 years Cyc would cope with novelty by recognizing analogies and would then be able to teach itself by reading the newspapers. Time is up and he seems to have made no progress on this front—at least I have seen no published account of Cyc "noticing patterns and regularities in the data, and drawing from those patterns *useful* new analogies, dependencies, and generalizations" [12, p. 357 (my italics)] as promised. But rather than call attention to this problem Lenat tells us that "After almost a decade, the Cyc project is still on target. The CNL (Cyc-based NL understanding and generation) subsystem is developing synergistically with the Cyc KB, and we expect a sort of crossover to occur in the next two years, by which we mean that most of the knowledge entry will take place by semiautomated NL understanding, with humans able to take the role of tutors rather than brain surgeons." [8, pp. 127–142]. There is no further talk of analogies and the two additional years will be up in one year. For reasons given in the new preface to my book, I do not think the analogy program will work since it presupposes having solved the problem of relevance. But the important thing is to be told clearly just what has been accomplished and what has turned out to be harder than anticipated and why. That might, for instance, be bad news for Simon, who is basing his current optimism on Lenat's success. In a recently published interview, he responds to a question concerning the commonsense knowledge problem by saying:

> Douglas Lenat has a ten-year program of building a huge semantic memory (CYC). Then we will see.... When people start to build programs at that magnitude and they still cannot do what they are supposed to, *then* we will start worrying. [19, p. 239]

Seeing Simon worried would, in itself, be a kind of progress. In the meantime I'll turn to the substantive issues raised by my critics.

I am grateful to Harry Collins for his detailed reading and his open minded and original critique. His new arguments both for and against the possibility of AI bring a breath of fresh air into a stale debate. Also he touches upon most of the issues raised by the other reviewers so I will deal with his comments in detail.

As I see it, Collins raises three problems for my overall approach. Here they are with a preview of my response to each. First, Collins claims that, since I base my work on Heidegger and Wittgenstein, I should follow them in acknowledging that there is no way reality is in itself independent of our social constructions of it. I note for the exegetical record, however, that it is a mistake to read Heidegger

and Wittgenstein as social constructivists. Second, even if all domains are social constructs, does it follow, as Collins claims, that the structure of these domains is up to society and that therefore activity in them can be automated if society so chooses? I will seek to show that chess and natural language, although clearly socially constituted, have, once constituted, intrinsic structures independent of how society happens to interpret them. Third, Collins objects to my preferring neural networks to symbolic representations. In response, I will defend my taxonomy of domains of human activity, distinguishing those that are amenable to the techniques of Symbolic AI and those that are not, as well as my claim that this taxonomy does not apply to simulated neural networks.

First, just for the record, as I read Wittgenstein he holds that meaning and truth presuppose the domain of nature whose structure is independent of the social world it makes possible. He says in the *Philosophical Investigations* that "our interest certainly includes the correspondence between concepts and very general facts of nature." [21, p. 230]. Likewise, Heidegger holds that "What is represented by physics is indeed nature itself." [10, p. 173].

The question that is directly relevant to the possibility of Symbolic AI, however, is not whether all domains of objects are socially constituted but whether, even if they were, it would show that the structure of these various worlds was somehow up to society. The domain of games such as chess provide an illuminating arena in which to refine our distinctions so as to test Collins' claim. Chess is clearly socially constituted; moreover it is constituted as fully digitalizable. No one doubts that with enough power one could calculate moves far enough ahead to play a winning game. But the question for AI that Collins and I are supposed to be addressing is not whether eventually a computer will be able to play master level chess. The question for GOFAI is whether the domain of chess is structured so that a physical symbol system using rules like those used by chess masters could achieve master level play. That is, whether there are principles or rules operating on context-free features that specify a good move in each situation or, in my language, whether there is a theory of the chess domain.

The answer is not at all obvious. Although the micro-world of chess is completely digitalized, to what extent the game has a structure that can be captured in a theory has been a subject of debate between classicists and romantics for over 300 years. The classicist (now represented by Symbolic AI enthusiasts such as Simon) claims that chess skill must consist in the interiorization of a theory of the chess domain that allows the chess master to generate good moves. The romantic (now represented by the Dreyfus brothers and some connectionists) holds, on the contrary, that chess is a domain without a theoretical structure and that therefore the only way for a human being to acquire mastery is to give up looking for chunks and principles and instead learn appropriate responses to tens of thousands of typical whole board patterns [6]. The question is not who is right, the classicist or the romantic, or whether the truth is between the two. The point is that society can invent a digitalized domain such as chess, but whether the domain has a structure that can be captured in theory is up to the domain not to society.

One simple way to focus our disagreement is to see how each of us would answer Collins' interesting question:

> Given that intelligent machines, like calculators, slide rules, logarithm tables, and books in general, are social isolates, how do they work? [2, p. 215]

Collins would explain the success of AI, where it has succeeded, by pointing out that in so far as we can learn to function in a context-free way like machines, machines can replace us in context-free interactions. I would explain how symbolic information processing systems can occasionally behave intelligently by pointing out that machines can behave intelligently in those domains that have a known theoretical structure, since we can program them with the theory of the domain.

An obvious domain in which to test these two accounts is calculation. Collins tells us that people can take what calculators do to be calculating because calculation is an isolated, regimented activity; I claim that there is such a regimented activity as calculating only because mathematics is an isolable, formalizable domain. Collins is making the deep Wittgensteinian point that we could not do mathematics if it were not for social agreement in our judgments and for apprenticeship into certain regimented practices; I want to insist that we could not have developed such rigid practices if the domain of numbers did not have a regular, digitalizable structure.

The fact that mechanical calculators make subtle mistakes—mistakes that Collins has a genius for teasing out—seems in two ways to support, rather than undermine, my thesis that there is an intrinsic structure to the domain, a structure to which our practices are obliged to conform. In *Artificial Experts* Collins points out that given the problem $7/11 \times 11$, we could agree to count 6.99999 as the right answer. We could. But, then, our arithmetic would be a mess and would not give us a grasp of the way things in our world, let alone the entities in the universe, generally behave. Moreover, that we don't perform the operations of division and multiplication in the order indicated reveals that we understand what such operations mean and how they relate to the domain. We do not just carry them out mechanically. This understanding is a skill that requires experience in the domain. That we can outdo the computer in seeing such meaningful relationships supports the Wittgensteinian point Collins and I agree on that applying formal rules requires informal skills. But, again, I want to insist that the appropriate order of operations is also a fact about the domain. It is not up to us, as individuals, or as a society. Thus Collins' example seems to be an argument *for* my two-substance view of knowledge, viz. that some domains are formalizable and some are not.

Another way to put my point is to accept Collins' observation that "one finds mimeomorphic actions in areas as varied as work on Taylorist production lines, the golf swing, high-board competition diving, ideal bureaucracies and *some* arithmetical operations", and ask why, then, does it turn out that of these domains only computation has been digitalized? The high-board diving example suggests that being regimented, while necessary, is not sufficient for formaliza-

tion. Collins would presumably agree. But where bodily skills are concerned, regimentation does not seem to be necessary or sufficient. If walking robots have been very difficult to make, it is not because walking is unregimented. If it were, the robot builders could give up on walking while going on to win prizes in the tightly regimented micro-worlds of high-diving and ballroom dancing. Conversely, a bike riding robot might succeed by following a formula proposed by Polanyi ("for a given angle of unbalance the curvature of each winding is inversely proportional to the square of the speed at which the cyclist is proceeding" [18, p. 50]), even though bike riding cannot be broken down into a series of mimeomorphic actions.

The same issue comes up again when we turn to expert systems. I argue that expert systems work in domains that we independently see how to formalize, while domains in which expert systems have failed to be as good as experts are domains which we have no independent reason to think can be formalized. I would agree with Collins that in most cases formalizable domains coincide with domains in which we have learned to discipline our minds or our bodies to perform a series of context-free tasks. So, for example, expert systems for loading transport planes, analyzing mass spectograph output, and configuring VAXes—all regimented domains—have achieved expertise, while expert systems based on heuristic rules for medical diagnoses and chess playing, have not, and no one would know how to begin to build an expert driver, or even an expert diver. But I would want to add that it is not just a contingent social fact that we have formalized the first three domains, and not the last three. Plane loading can be captured by a combination of geometry and trial and error, spectography has a theoretical structure, and computer component choice can be carried out in a context-free way by using look-up tables and heuristic rules for how to combine components having various capacities. Experts in these fields were, indeed, replaceable because they were behaving in a digitalizable way, engaging in a calculating kind of expertise, and they were behaving digitally because the domain had a digitalizable structure. Collins' taxonomy of domains into those that society has disciplined and those it has not presupposes my taxonomy into regimentable and non-regimentable domains. To see the dependence going the other way seems to me literally preposterous.

I think that many of these limitations on activities that can be regimented has to do with embodiment. Collins seeks to show that embodiment is not important so as to replace it with his idea of social embedding. He elaborates convincingly Wittgenstein's point that if lions could talk we could not understand them:

> Circus lions talking among themselves would, presumably, group what we call a household chair along with the other weapons they encounter in the hands of "lion tamers", not with object to do with relaxation. They would not distinguish between sticks and chairs and this is why their language would be incomprehensible to us. But this does not mean that every entity that can recognize a chair has to be able to sit on one. That confuses the capabilities of an individual with the form of life of the social group in which that

> individual is embedded. Entities that can recognize chairs have only *to share the form of life* of those who can sit down.

Collins is right, but I don't think it matters whether our intelligent behavior depends *directly* upon our having the sort of bodies we have or whether our form of life depends *in the end* on having our kind of bodily dispositions—what Wittgenstein calls "the facts of natural history". I grant Collins that we can understand stories about chairs without being able to sit on them, and, of course, being able to sit in chairs is necessary for thinkers but irrelevant for much of their thinking. The question is, can we understand stories about chairs or any everyday objects or events without sharing a lot of the characteristics of the embodiment of our consocials? I want to argue that one would need to have experience with our kind of body to make sense of our kind of world. Collins denies this claim. He grants that:

> The shape of the bodies of the members of a social collectivity and the situations in which they find themselves give rise to their form of life. Collectivities whose members have different bodies and encounter different situations develop different forms of life.

Yet he concludes:

> But given the capacity for linguistic socialization, an individual can come to share a form of life without having a body or the experience of physical situations which correspond to that form of life.

I contend that Collins again has it backwards. Having the sort of bodies we have is a necessary condition for social embedding in a society of similarly structured human beings.

To see this, we need only notice what Collins grants, viz. that our form of life is organized through and through by and for beings embodied like us; people with bodies that have insides and outsides; that have to balance in a gravitational field; that move forward more easily than backwards; that have to approach objects by traversing the intervening space, overcoming obstacles as they proceed, etc. Our embodied concerns so pervade our world that we don't even notice the way our body is at home in it. We would only notice it by experiencing our disorientation if we were transported to an alien world set up by creatures with radically different–say spherical or gaseous—bodies, or by observing the helpless confusion of such an alien creature brought into our world. One thing is sure, nothing could be more alien to our life-form than a big, metal box with no body, no special way of moving, etc. The computer has no built-in *preunderstanding* of how our world is organized and so of how to get around in it. The odds against its being able to acquire all the knowledge it needs of the embodiment familiar and obvious to us because we are it, are overwhelming.

Lenat's Cyc provides a perfect opportunity to notice the shared world we normally take for granted, and how the cards are stacked to enable creatures who share our embodied form of life to learn to cope intelligently, while making all

other creatures look hopelessly stupid. As I noted in my Introduction to *What Computers Still Can't Do*, Lenat collects some excellent examples of the difficulties involved. In order to answer Collins (and suggest a problem for McCarthy later) I will start with an excerpt from my Introduction and develop it further. Lenat says: "Take the following sentence: 'Mary saw a dog in the window. She wanted it'." [13, p. 200]. He then asks: "Does 'it' refer to the dog or the window? What if we'd said 'She *smashed* it', or 'She pressed her nose up against it'?" [13, p. 200].

Note that the last case—she pressed her nose up against it—seems to appeal to our ability to *imagine* how we would act in the situation, rather than requiring us to consult *facts* about dogs and windows and how a typical human being would react. We draw on what we *are*, not what we know. We imagine getting around in the world, such as getting closer to something on the other side of a barrier to see it more clearly. Merleau-Ponty describes this general, body-based tendency in his phenomenological account of what he calls *maximal grip*.

According to Merleau-Ponty, higher animals and human beings are always trying to get a *maximum grip* on their situation. Merleau-Ponty's inspiration for his notion of maximal grip comes from perception and manipulation. When we are looking at something, we tend, without thinking about it, to find the best distance for taking in both the thing as a whole and its different parts. When grasping something, we tend to grab it in such a way as to get the best grip on it.

> For each object, as for each picture in an art gallery, there is an optimum distance from which it requires to be seen, a direction viewed from which it vouchsafes most of itself: at a shorter or greater distance we have merely a perception blurred through excess or deficiency. We therefore tend towards the maximum of visibility, and seek a better focus as with a microscope. [14, p. 302]...
>
> My body is geared into the world when my perception presents me with a spectacle as varied and as clearly articulated as possible, and when my motor intentions, as they unfold, receive the responses they expect from the world. [14, p. 250]

That is not meant to deny that a robot could also have a body with certain capacities for motion and perception and that the robot would have to organize its knowledge and generate its actions taking its own structure into account. But it is meant to suggest that the robot would presumably do this not by representing its capacities and then reasoning from them. Rather, the capacities would be represented, as it were, in its procedures, for coping with things. This is not a new idea in AI, but it seems to me an important one that needs to be followed out. A new problem would then arise, however, when the robot had to solve problems involving perception and action without acting, as for instance in understanding stories. This is when human beings resort to imagining their actions. Could the robot run a simulation of itself? If it could not represent its capacities and make inferences from them, could it somehow run itself hypothetically? If such ideas make sense, working them out would be an important step towards making an intelligent robot.

Lenat, unlike Collins, does see that the body is indispensable for commonsense intelligence. He intends "in principle and in Cyc—to describe perception, emotion, motion, etc., down to some level of detail that enables the system to understand humans doing those things, and/or to be able to *reason* simply *about* them" [13, p. 218 (my italics)]. But again it seems to me this whole approach is misguided. We don't normally reason *about* our bodily capacities, we reason *in terms* of them. That is, when we reason in a commonsense way we already use our sense of our body to guide our reasoning. If this is so, the conclusions human beings find reasonable will differ from those of a disembodied inference-making machine.

Mark Johnson, in his book, *The Body in the Mind*, tries to work out in detail how reasoning depends on our sense of our embodiment. He writes:

> The epistemic sense of modals, such as *must*, *may*, and *can*, find their home in the domain of reasoning, argument, and theorizing. . . . I am claiming that . . . the basis for this connection is that we understand the mental in terms of the physical, the mind in terms of bodily experience. In particular, we understand mental processes of reasoning as involving forces and barriers analogous to physical and social forces and obstacles. [11, p. 53]

To illustrate the way the body works in structuring the basic metaphors in terms of which we organize even our concepts Johnson uses our understanding of force:

> We learn to move our bodies and to manipulate objects such that we are centers of force. Above all we develop patterns for interacting forcefully with our environment—we grab toys, raise the cup to our lips, pull our bodies through space. We encounter obstacles that exert force on us, and we find that we can exert force in going around, over, or through those objects that resist us. Sometimes we are frustrated, defeated and impotent in our forceful action. Other times we are powerful and successful. . . . In each of these motor activities there are repeatable patterns that come to identify that particular forceful action. These patterns are embodied and give coherent, meaningful structure to our physical experience at a *preconceptual* level. [11, p. 13]

Johnson concludes:

> Understanding is never merely a matter of holding beliefs, either consciously or unconsciously. More basically, one's understanding is one's way of being in, or having, a world. This is very much a matter of one's embodiment, that is, of perceptual mechanisms, patterns of discrimination, motor programs, and various bodily skills. [11, p. 137]

On this view embodiment is presupposed for acquiring the skills and knowledge which amount to social embedding. We move and meet resistance, etc., as embodied individuals even before we are socialized. Indeed, each human being begins as a cultureless animal that must acquire the culture. That is why the form of life, if it is to be acquired, must be structured in keeping with the individual's embodiment. "Intelligence abides bodily in the world", as John Haugeland argues

in his contribution, because the social world is organized by and for beings with our kind of bodies. That is not to deny that, as we grow up our body becomes a social body, but the structure of any social world is constrained by what our body can learn and make sense of.

Collins, however, as we have just noted, claims that social embedding is more basic than individual embodiment. He thinks this claim is supported by Oliver Sacks' account of the case of Madeleine, who, Sacks says, has acquired her understanding of our culture linguistically—by being read to from books. Collins concludes:

> We can say with confidence that if we can't train a computer without a body to act like a socialized human, giving it the ability to move around in the world encountering the same physical situations is not going to solve the problem. On the other hand, if we can find out what is involved in the sort of socializing process undergone by a Madeleine—let us call it "socialisability"—we may be able to apply it to an immobile box.

If Merleau-Ponty, Mark Johnson and I are right, however, Collins' understanding of Sacks' account must be science fiction. And, indeed, if one goes back to Sacks' account, one sees that Madeleine was far from being an immobile box. What was special about Madeleine was merely that she was blind and could not use her hands to read Braille. True, she was overprotected as a child, but there is no hint that she was carried everywhere. If she crawled, walked, balanced, overcame obstacles, had to find the optimal distance for kicking and hearing, etc. she would have acquired the body schemata which would have allowed her imaginatively to understand and project into new domains the events and cultural norms she heard about from books. Collins misses these facts when he says:

> Madeleine has imagination; she can empathize with those who have more complete bodies. But under this argument a body is not so much a physical thing as a *conceptual structure*. *If* you can have a body as unlike the norm and as unable to use tools, chairs, blind persons' canes and so forth as Madeleine's, yet you can still gain commonsense knowledge, *then* something like today's computers—fixed metal boxes—might also acquire commonsense given the right programming.

My point is that Madeleine *has* our body structure and many of our body-skills. That is why she can be socialized into a human world set up in terms of human bodies. So we should not agree with Collins' conclusion:

> In sum, the shape of the bodies of the members of a social collectivity and the situations in which they find themselves give rise to their form of life. But given the capacity for linguistic socialization, an individual can come to share a form of life without having a body or the experience of physical situations which correspond to that form of life. . . . [H]uman bodies are not *necessary* for human-like socialization.

On the contrary, what little we know about the capacity for linguistic

socialization suggests that we have to have a body with a structure like our consocials and skills for grasping and moving like theirs, if we are to acquire their language at all. For, as Wittgenstein points out, and Collins would surely agree, to learn a language is not just to learn a fixed set of words and grammatical constructions, but to use this linguistic equipment in ever new situations. As Wittgenstein argues, it is this ability to project a language into new situations that shows we have understood it, and as Johnson shows, we can't do this projecting without appeal to bodily analogies that we sense directly because we have to move, overcome opposing forces, get a grip on things (and ourselves), etc. So it looks to me like Collins has things back to front again. To learn a natural language a computer has to have a body; it must be embodied if it is to be embedded.

On the importance of repairs for AI successes I fully agree with Collins. In footnote 45 of *What Computer Still Can't Do* I mention in passing that NetTalk depends upon our making sense of the rather garbled English it produces. This is not an objection. We make the same sort of repairs in perception and in communication. But on Collins' account it remains a mystery why repairs work well in some domains but not in others. We repair NetTalk, printed English, or our perceptual errors without even noticing the noise, but we have to stop and reprogram our computer or ignore its result if it makes an illegal move in chess or gets the wrong answer in arithmetic. In an informal domain like pronunciation the context does a lot of the work; as long as the program does roughly what it is supposed to, we don't notice the deviation. But in a digitalized domain, like chess or calculation, there is no ambiguity tolerance—no everyday context to help us fill in the gaps—so we can't ignore even a small error. For that reason in digitalized domains we are not given much opportunity for repairs.

What is important about repairs, I think, is that they again reveal that some domains can be formalized or digitalized and other domains cannot. Indeed, some domains *must* be regimented to do their work while others must remain context-dependent to function at all. It is essential to natural language, for example, that we be able to project our words into new situations, using them in new ways analogous to the old ways, and that our fellow language users can, nonetheless, understand us. In such cases we are not repairing mistakes but seeing—presumably on the basis of our embodied and embedded intuitions—that what is being said, although it breaks the rules, still makes sense.

Collins makes a similar point, but in a misleading way, when he suggests we could choose to regiment even natural language.

> [I]n the case of natural language interactions with computers we would have to learn to make our speech mimeomorphic, and that would mean restructuring our lives. We are *unwilling* to do this kind of thing and that means computers don't work in the corresponding roles.

In another paper, Collins expands on this possibility:

> More translation machines might mean more and more routinization of the

> way we write until our linguistic form-of-life becomes, as in George Orwell's *1984*, fully instantiable in mimeomorphic action and fully computerizable. [3, p. 729]

I contend that it is not up to us whether, in the case of language, we choose to behave like a computer. Another way to put this point is that natural languages not only have a syntax and a semantics but also a pragmatics. Even if we could agree to regiment, and so render formalizable, syntax and semantics—even if we could regiment parts of pragmatics like turn-taking—we would still not be able to use language in such a rigid way as to render pragmatics formalizable. If it is essential to natural language that it can be extended into ever-new situations, as I take it all parties to this debate agree, to turn our language into a context-free code is not just something we are *unwilling* to do, it is something the domain prohibits. No one doubts that language is a social construction, but once it is constructed, it is not up to any one of us, nor to society as a whole, to decide, by making our linguistic behavior mimeomorphic, to turn the world that everyday, situationally relevant language opens up into a formalizable domain. If we did this there would be no language, no society, no world, and no us.

The touchstone of Collins' and my disagreements is our attitudes towards simulated neural networks. In his article, "Why Artificial Intelligence is not Impossible", Collins tells us:

> The theory of mimeomorphic action makes no distinction between one kind of machine and another. Machines can only reproduce tasks that can be broken down into a series of mimeomorphic actions. [4]

But from my point of view, as Collins clearly sees, this covers up an essential difference between using computers to instantiate symbolic representations and using them to model neural networks. On Collins view, any device should be able to replace experts in regimented domains and no device should be able to replace them in non-regimented domains. My view leads me to make a sharp distinction not only between types of domains, but also between what we can expect of different types of computers. Stephen Palmer, for instance, notes that:

> We have never been able to get very good models of human perception, pattern recognition, or human learning. We have never been able to get contextual effects to arise naturally out of physical symbol system models. These are exactly the domains where connectionism is strong. [17, p. 162]

So, although we cannot expect physical symbol systems to exhibit expertise in domains for which we have no theory, this limitation does not apply to neural networks. Networks learn from examples and respond in similar ways to similar cases without needing to be given, or needing to extract, any rules or principles which serve as a theoretical representation of the structure of the domain in which they work. It we grant that nets can learn to respond to patterns without analyzing the patterns into context-free features, then one would expect simulated networks, unlike symbolic programs, to be able to learn to respond appropriately

in domains that have not been, and perhaps cannot be, digitalized. In so far as nets are succeeding in some areas where symbolic AI has had a hard time, this is just what has happened. Thus it looks like my two-knowledge-stuffs view turns out to be marking a deep distinction, and Collins' taxonomy turns out to be superficial.

One last query for both Collins and Haugeland. Why does intelligence have to be social? If one disqualifies the tasks animals perform as unintelligent because animals do not have our sort of language and institutions, we will need to be told in a non-question begging way what intelligence is. After all, Kohler wrote a book on the intelligence of apes, including the example of an ape figuring out how to move a box and then get on the box and use a stick to knock down a banana that was out of reach. This is clearly using flexible means to achieve a desired end. That is not enough to count as intelligence, however, since a chess program using brute force can do that. What is essential is that apes learn from experience how to achieve their ends. That is, when something works, they are able to adapt it to other contexts. If we define intelligent beings, then, as beings, that can learn from their experience to find flexible means to achieve their ends, apes are surely intelligent.

I do not doubt that some intelligent behavior, like say, investing, is inherently cultural, but I do not see why one would hold that only such institutional and linguistic behavior can count as intelligent behavior. Why should we suppose that animal behavior, no matter how seemingly flexible, adaptive and skilled it may appear, since it is not cultural, is *ipso facto* mimeomorphic, i.e., that animals always respond rigidly to the same situation with the same action? But how else could Collins reach the conclusion that: "whatever can be done by neural nets—or dogs, performing seals (and babies, for that matter)—could be described in rules even if we have not actually so analyzed them."? [3, p. 732]. To me it looks like animals reveal a whole domain of context-sensitive, skillful coping. Even if context sensitive behavior cannot be simulated by GOFAI, I don't see why it could not, at least in principle, be successfully simulated by networks regardless of whether such networks could be socialized.

John Haugeland has always understood me better than I understood myself, and written what I should have written. He positions himself, as I would position myself, between those who seek an abstract, representational account of mind such as John McCarthy and those, like Harry Collins, who hold that mind is a product of social embedding. He wants to describe embodied-being-in-the-world, as I do. He does not have much to say about the actual structure of the body and the role it plays in having a world. He does, however, make a strong argument that, thanks to the body, we are much more richly coupled to the world than the traditional representational view allows. In fact, he argues that there cannot be intelligence apart from embodied living in the world.

Haugeland reads me so carefully and extends my view so sympathetically that, in the end, it seems to me he makes explicit and embraces a contradiction in my own account I had not noticed until now. He works out my account of being-in-the-world in a way that makes embodiment basic, then goes on to make social

embedding basic so that I end up sounding like Collins. Haugeland himself seems of two minds on this important issue and I now realize that I am too. I must admit that I generally agree with Heidegger, and that Haugeland and Collins, in emphasizing the fundamental role of the social, are being good Heideggerians, but I am also pulled toward Merleau-Ponty, who would claim that intelligence arises from the way individual animals are tightly coupled with the *perceptual* world.

The problem for me is that Haugeland wants to link intelligence with the meaningful and the meaningful with the objective or normative, and that with the social. As he puts it, "the meaningful is that which is significant in terms of something beyond itself, and subject to normative evaluation . . .". This could still be compatible with Merleau-Ponty if meaning meant something like a gestalt, and norm meant that a perceptual gestalt demanded a certain sort of completion and resisted others. Since there is more to the figure than is directly present, e.g. what I take to be a house seen from the front looks thick and looks like its concealing an inside and back, it allows us to "deal reliably with more than the present and the manifest"—Haugeland's definition of intelligence. But for Haugeland the norm has to be a *social* norm.

As I read Haugeland, the first two thirds of his comments makes a terrific case for the Merleau-Pontian approach to embodied intelligence in general. Then Haugeland adds his own convincing version of the Heideggerian view that social norms—"equipment, public places, community practices"—are essential to *human* intelligence. What I don't agree with, and what lands Haugeland in the Collins camp, is his claim that human intelligence and meaning are the *only* kind of intelligence and meaning there is, so that part two of his comments seems to take back part one. This becomes clear when he is led to the conclusion that "In my own view (and I suspect also Dreyfus'), there is no such thing as animal or divine intelligence." I don't know about God, but as I argued in my answer to Collins above, apes seem to show intelligence, precisely by using something like equipment, and no community practices need be involved.

Rather than letting a social theory of intelligence take back a general theory of embodied intelligence, I would like to build Part II of Haugeland's paper on Part I. That would amount to building Heidegger on top of Merleau-Ponty by showing how social intelligence grows out of and presupposes non-social intelligence. Merleau-Ponty clearly has such a project in mind when he says:

> The body is our general medium for having a world. Sometimes it is restricted to the actions necessary for the conservation of life, and accordingly it posits around us a biological world; at other times, elaborating upon these primary actions and moving from their literal to a figurative meaning, it manifests through them a core of new significance: this is true of motor habits such as dancing. Sometimes, finally, the meaning aimed at cannot be achieved by the body's natural means; it must then build itself an instrument, and it projects thereby around itself a cultural world. [14, p. 146]

T.D. Koschmann's remarks are based not only on *What Computers Still Can't*

Do; he seems to have read all my books with sympathy and understanding, and gives a succinct summary and critique of my views. His comments are primarily directed to my brother's and my account of expertise in *Mind over Machine*. We are not, as he realizes, repeating the current view that experts "*reason*" from a large data base by "working with patterns" or that "expert problem solving consists of reasoning by analogy". It is precisely our point that experts don't normally solve problems and do not normally reason at all. The expert is not aware of using inference rules, and the rules he gives knowledge engineers don't work to reproduce his expertise. The expert is not aware of using cases either. Indeed, knowing which cases are relevant presupposes his expertise. One can claim that the relevant cases are identified by certain context-free features, but the problem is, which ones? The expert is not aware of any features, and no more able to suggest reliable features than reliable rules.

Because it was confused with the case-based approach, in the second edition of our book we retracted our fanciful account of expertise in terms of tens of thousands of stored holograms, and adopted the model of a simulated neural network being trained by experience to make more and more refined pattern discriminations. But with this model in mind it is hard to make sense of what Koschmann's holistic, non-analytical *problem solving* would be. If one is mapping input vectors to output vectors there is no problem solving because there is no problem. This is not to deny that problem solving is precisely what beginners do and also what experts do when faced with novel and complicated situations. We do, however, as Koschmann sees, "privilege" intuitive expertise over such problem solving. We feel justified in assuming that it is better not to have a problem than to solve one.

Koschmann has a valid point, however, when he rejects the strong view that "representation is irrelevant to expert performance". He correctly points out that "representation is important to instruction at all levels, if for no other reason, because it allows us to see the world in new and different ways". Unfortunately, he confuses the point when he continues, "Traditional AI was all about representation-defining schemes for mechanically storing, retrieving, and applying knowledge". Linguistic representation in natural language, which is what Koschmann is defending, is public and informal. It should not be confused with the formal, internal representations used in GOFAI. We stick to our argument that the sort of formal representations used in AI are never useful for acquiring or sharpening expertise. Linguistic representations in the form of books, films, etc., all of which are public, informal representations, however, are essential for communication, education and the sharing and preservation of cultural knowledge. Still, whatever sharing of information, justification and public accountability etc. one can gain by putting intuitive know-how into such representational form is purchased at the expense of expertise. Nonetheless, Koschmann's idea of a collaborative classroom makes good sense for dealing with problematic situations. Discussion (of course in a natural language) does sharpen intuitions by enabling experts to see things from different perspectives.

It is hard to reply to the Strom and Darden contribution since its statement of

my views are so far afield that I have the feeling it is responding to a different book. But then, its claim that the GOFAI research program has been advanced by brute-force search is equally off the wall. Is the paper, perhaps, written by a sophisticated GOFAI program that reveals its limitations by its lack of a sense of relevance and of context sensitivity? I fear that in entering into a dialogue with Strom and Darden I may discover that I have been duped into admitting that a computer can, after all, pass the Turing test, but I have to take that risk, so here goes. I will, out of caution, refer to my interlocutor as SD.

In the next few pages I will take up the following questions: Why favoring holograms and simulated networks does not constitute a research program? What the GOFAI program was trying to do? And whether, as SD claims, "Symbolic AI [has succeeded] in areas where Dreyfus predicted failure"?

What could have been an interesting discussion about whether I have a research program never gets off the ground. First, because SD seems to think that when, in the first edition of *Mind over Machine*, my brother and I resorted to holograms to answer the what-else-could-it-be argument, we were endorsing that far-out view as plausible. Second, even if SD had read the second edition and noted that we had switched to invoking neural networks, the idea that we were backing that research program would have quickly been dispelled by the end of the new Preface to *What Computers Still Can't Do*. A human reader would have seen as relevant to this claim the fact that there I argue that neural networks too will fail to solve the commonsense knowledge problem. One could then infer that I did not put my bets on the neural network research program. In fact, I don't think of myself as having a research program at all but rather of doing the job philosophers have always done, seeking conceptual clarification and looking critically at various sorts of knowledge claims.

Since SD and I at least agree that "Research programmes are identified by a 'Hard core' of centrally important hypotheses", we should look for GOFAI's. It's not hard to find. The basic claim of Symbolic AI, as defined by Newell and Simon's paper on physical symbol systems, was that human beings and heuristically programmed computers, at a suitable level of description, were operating on the same principles—principles that could be implemented in brains or digital computers—and that therefore work in AI was a contribution to cognitive psychology. Thus Newell and Simon assumed that the principles in question included ways of avoiding massive brute-force search since this is not available to human beings. They say:

> The potential for the exponential explosion of the search tree that is present in every scheme for generating problem solutions warns us against depending on the brute force of computers—even the biggest and fastest computers—as a compensation for the ignorance and unselectivity of their generators. . . . The task of intelligence, then, is to avert the ever-present threat of the exponential explosion of search. [15, p. 57]

There is, of course, a branch of AI that would be happy to make machines behave intelligently in any way at all, but that cannot be the research program

GOFAI researchers are engaged in since it would cast no light on human intelligence. SD seems to agree with Newell, Simon and me on this point. Indeed, in saying that "when an AI program performs as it is intended, it offers a potential for theoretical confirmation . . . especially in fields such as cognitive psychology and philosophy of mind", SD seems to explicitly endorse the human-oriented research goal as essential to AI.

If psychological relevance *is* the goal, then GOFAI work in checkers and chess must seek heuristics to prune the search tree so that the program does not have to search through millions of moves. This should be obvious but just to nail down the point I'll quote again from the Simon interview I already cited:

> [Interviewer:] Now, one could imagine a machine that wins every game of chess by pure brute force. But we know that human beings don't think that way.
>
> [Simon]: So that machine would not pass the Turing Test, because it would not fit the behavioral evidence, which shows that no human being looks at more than a hundred branches of the tree before making a move. If we want to have a theory of human chess playing, we would need a machine that also matches that piece of evidence.
>
> [Interviewer:] So the program with brute force would not be Artificial Intelligence?
>
> [Simon]: It would be Artificial Intelligence, but not cognitive science. Take the best chess programs—Deep Thought, for example . . . It does not tell anything about how a chess grandmaster thinks, or very little. [19, p. 243]

Since for SD, as for Simon, the kind of AI in question is the kind that can make a contribution to cognitive science, it is hard to understand how SD can reach the conclusion that "Deep Thought is clearly a vindication for the traditional AI programme of heuristic search".

Since the distinction between GOFAI as a contribution to cognitive psychology and AI as any sort of technique using symbolic representations was obvious from the start, I never predicted failure for brute-force calculation, but only for chess programs based on heuristics supposedly used by chess masters to cut down the search to a few hundred moves. I was all for trying brute-force but I was not ready to take a stand on what it could achieve. In the second edition of *What Computers Can't Do*, I pointed out that chess was a micro-world and added that "while the game's circumscribed character makes a world champion chess program in principle possible, there is a great deal of evidence that human beings play chess quite differently from computers" (p. 29). I have not changed my mind on this nor is there any reason to. Obviously, I would never say, as SD reads me as saying (without page references), that what I called zeroing-in was "necessary for expert performance", nor that brute-force counting out was "hopeless". I did say in *What Computers Can't Do*, and I still hold, that:

> With present programs what is really at stake is how far computers . . . can make up by sheer brute force for the use of long-range strategy, the

> recognition of similarity to other preanalyzed games, and the zeroing in on crucial aspects characteristics of advanced human play. (p. 32)

I glossed this in *What Computers Still Can't Do* by saying that "good chess players don't seem to figure out from scratch what to do each time they make a move". But I never would have thought that a reader could be so insensitive to context as to retort, "However, Deep Thought is at least a 'good player', yet it essentially does figure out from scratch what to do each time".

Of all the responses the one I most appreciate is the one with which I most disagree. That is, of course, the response by John McCarthy. I appreciate his comments on a personal level because he has kept his cool and stayed a friendly opponent all these years in spite of my often abrasive remarks. Intellectually, I admire McCarthy's respect for philosophy and philosophical argument.

Obviously, McCarthy and I come from two totally different paradigms. His research program of "formalizing human reasoning" assumes that intelligence is based on making inferences and asks how to make the right ones. I think the preliminary question needs to be posed whether everyday intelligent behavior involves reasoning at all. As I have been arguing in this response, we seem to use the know-how that comes with having a body without inferring from facts about the body, and we seem to avoid the problem of selecting relevant data to reason from by responding to whole situations. I agree with McCarthy that on this point neither of us can give an argument for our intuitions that should convince the other. So we each do what we can. He works on solving the problems internal to his view; I try to bring up problems that such a view will have to face further down the line.

But there is an asymmetry. McCarthy reads my papers and I regret that, due to my lack of expertise in logic, I have not read his technical papers on circumscription. This is especially troubling to McCarthy since, as he says, this work is meant in part to answer my early claim that human beings exhibit ambiguity tolerance which he translates as the ability to reason from a commonsense knowledge data base even when there are inconsistencies in it. But what I had in mind in speaking of ambiguity tolerance was the way the context enables one to repair incomplete or contradictory inputs. The question of ambiguity tolerance for me is not how to reason when there are exceptions but how the context allows us to ignore most exceptions altogether. More generally, my problem is how to store commonsense knowledge so as to be able to retrieve the relevant information, not how to reason from the facts once millions of facts are collected and organized and the relevant facts already retrieved. For this reason I do not feel that McCarthy has solved the problem I meant to pose, although I appreciate his having turned my rather general remarks into a clear problem that could then be solved.

I presume that McCarthy would agree with Lenat that the way to solve the relevance problem is to use relevance axioms. Lenat proposes two kinds of relevance axioms: specific and general. The idea behind specific relevance axioms is that different sections of the knowledge base "can be ranked according to their relevance to the problem solving task at hand" [1, p. 15]. So, for example, if the

task given to Cyc is to solve a problem in chip design, the program will be guided in its search for relevant information by an axiom to the effect that the computer section is more relevant than the botany section (although the botany section cannot be ruled out completely, for it might be the source of a useful analogy or two) [1, p. 15]. But as my old example of goldenrod on the race track being relevant to a jockey with hay-fever suggests, what is relevant can, indeed, come from a completely unrelated domain such as botany.

For this sort of problem Lenat proposes general relevance axioms. These are formalizations of such statements as "It is necessary to consider only events that are temporally close to the time of the event or proposition at issue" [9, p. 7]. This would bring in the goldenrod all right, but of course it would bring in an indefinitely large number of other facts about the race track, so the relevance problem would not be solved. Moreover, in explaining and defending this axiom, Guha and Levy say "it is rare that an event occurs and . . . [then after] a considerable period of time . . . suddenly manifests its effects" [9, p. 7]. But promises and all sorts of health problems, to take just two examples, have exactly the form that what is relevant can be far in the future, and all sorts of historical and psychological facts relevant to my present can be found in my more or less distant past.

Nonetheless, McCarthy envisages "a system [that] would . . . represent what it knew about the world in general, about the particular situation and about its goals by sentences in logic". I think there is a general reason such a project will not succeed. As the relevance axioms suggest, what counts as relevant depends on the current context. But, as I argued in my comments on Koschmann, how we classify the current context itself depends on what we count as the relevant information. This circularity suggests that people cannot proceed by first determining the context and then looking for information relevant in that context nor by first finding the relevant information and then using it to determine the context. It suggests to me—as it suggested to Merleau-Ponty—that both rationalism, with its inferences from stored data, and empiricism, with its claims that we associate to stored cases, are bound to fail. It further suggests that perhaps some neural network sort of association with no inferences and no stored cases is more promising. Jeffrey Elman's paper, "Finding Structure in Time" [7, pp. 179–212], gives a first clue as to how the behavior of a neural net could become sensitive to relevant context without developing a symbolic model of what in the context is relevant. Such results are surely one reason why many young AI researchers are working on simulated networks and fewer are going into GOFAI, or at least, fewer are going in with the sort of long range vision of fully intelligent logic machines which McCarthy so stalwartly defends. There are plenty of problems with neural networks, however, so if McCarthy can solve the problem of relevance for common sense knowledge he can still turn things around.

But I think McCarthy would agree with me that the relevance problem is too hard for now. He would like me to give him a simpler one. I will try. In my response to Collins I explain why I think an experience of the body is essential even to reasoning and why, if true, this poses a serious problem for the logicist

approach to AI. It is not "empathy" that is important but the experience of moving about in a world, dealing with obstacles, etc. To return to Lenat's example of Mary and the dog in the window, the problem is not just getting the right antecedent for the pronoun naming the object of Mary's desire, but getting the right antecedent for what she presses her nose against. How does one spell out propositionally our know-how concerning the body and its motor and perceptual systems? Of course, one could write out all the information about paths, obstacles, distance required for a maximal perceptual grip, etc., for this single case, and draw inferences from them, but that we all agree would be cheating. But no one, as far as I know, has a clue as to how to store and access the facts about the body that would be needed to reason about such situations in general. There is no reason to think this problem can be solved. We certainly don't solve it. To avoid it, we *imagine* moving around in the situation and see what we would press our nose against. McCarthy, however, thinks the problem of finding the antecedents in Lenat's example is "within the capacity of some current parsers", so I would be happy to submit this example as the next thing a logic approach should try to deal with. I sincerely promise to try to read anything McCarthy and his followers write on the subject.

Meanwhile, I remain skeptical about both GOFAI and neural networks. Although I expect that there will someday be androids like Data in *Star Trek: The New Generation*, I agree with McCarthy's cautious assessment that "reaching human level AI is not a problem that is within engineering range of solution. Very likely, fundamental scientific discoveries are still to come." We only differ in that I doubt that the activity of these future androids' "positronic brains" will be based on the research of any of the current contenders in the AI race.

References

[1] P. Blair, R.V. Guha and W. Pratt, Microtheories: an ontological engineer's guide, MCC Technical Report No. CYC-050-92 (1992).
[2] H.M. Collins, *Artificial Experts* (MIT Press, Cambridge, MA, 1981).
[3] H.M. Collins, Hubert L. Dreyfus, forms of life, and a simple test for machine intelligence, *Social Stud. Sci.* **22** (1992).
[4] H.M. Collins, Will machines ever think?, *New Scientist* **1826** (June 20, 1992).
[5] D.C. Dennett, In defense of AI, in: P. Baumgartner and S. Payr, eds., *Speaking Mind: Interviews with Twenty Eminent Cognitive Scientists* (Princeton University Press, Princeton, NJ, 1995).
[6] H.L. Dreyfus and S.E. Dreyfus, *Mind over Machine* (Free Press, New York, 1988) Chapter 1.
[7] J.L. Elman, Finding structure in time, *Cognitive Sci.* **14** (2) (1990).
[8] R.V. Guha and D.B. Lenat, Enabling agents to work together, *Comm. ACM* **37** (7) (1994).
[9] R.V. Guha and A.Y. Levy, A relevance based meta level, MCC Technical Report No. CYC-040-90 (1990).
[10] M. Heidegger, Science and reflection, in: *The Question Concerning Technology and Other Essays* (Harper & Row, New York, 1977).
[11] M. Johnson, *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason* (University of Chicago Press, Chicago, IL, 1987).
[12] D. Lenat and R.V. Guha, *Building Large Knowledge-Based Systems* (Addison-Wesley, Reading, MA, 1990) 357.

[13] D.B. Lenat and E.A. Feigenbaum, On the thresholds of knowledge, *Artif. Intell.* **47** (1991) 185–250.
[14] M. Merleau-Ponty, *Phenomenology of Perception* (translator C. Smith) (Routledge & Kegan Paul, 1962).
[15] A. Newell and H.A. Simon, Computer science as empirical inquiry: symbols and search, in: J. Haugeland, ed., *Mind Design* (MIT Press, Cambridge, MA, 1988).
[16] A. Newell, The serial imperative, in: P. Baumgartner and S. Payr, eds., *Speaking Mind: Interviews with Twenty Eminent Cognitive Scientists* (Princeton University Press, Princeton, NJ, 1995).
[17] S.E. Palmer, Gestalt psychology redux, in: P. Baumgartner and S. Payr, eds., *Speaking Mind: Interviews with Twenty Eminent Cognitive Scientists* (Princeton University Press, Princeton, NJ, 1995).
[18] M. Polanyi, *Personal Knowledge* (Routledge & Kegan Paul, 1962).
[19] H.A. Simon, Technology is not the problem, in: P. Baumgartner and S. Payr, eds., *Speaking Mind: Interviews with Twenty Eminent Cognitive Scientists* (Princeton University Press, Princeton, NJ, 1995).
[20] P. Smolensky, On the proper treatment of connectionism, *Behav. Brain Sci.* **11** (1988) 1–74.
[21] L. Wittgenstein, *Philosophical Investigations* (Basil Blackwell, Oxford, 1953).