ELSEVIER Artificial Intelligence 80 (1996) 119–128

Artificial Intelligence

Book Review

Body and world: a review of *What Computers Still Can't Do: A Critique of Artificial Reason* (Hubert L. Dreyfus)[☆]

John Haugeland*

Department of Philosophy, University of Pittsburgh, Pittsburgh, PA 15260, USA

Received July 1993

1. Returning to Part III

The original edition of *What Computers Can't Do* comprised three roughly equal parts: (i) a harsh critical survey of the history and state of the art in AI, *circa* 1970; (ii) a brilliant philosophical *exposé* of four hidden assumptions shoring up AI's misplaced optimism; and (iii) a much more tentative exploration of ways to think about intelligence without those assumptions. Part I, because it was the most combative (and also the easiest to understand), got most of the attention. Also, since that discussion was the most timely—hence the most quickly obsolete—it is what the excellent substantive introductions to the later editions have mainly brought up to date. An unfortunate consequence of these concentrations, however, is that the more interesting and enduring parts of the book, Parts II and III, have been somewhat eclipsed and even neglected.

The third part, because it is the deepest and the most pioneering, is simultaneously the most difficult and the least developed—and, accordingly, I think, also the most rewarding to reconsider. Its principal theses are: that human intelligence is *essentially* embodied; that intelligent bodies are *essentially* situated (embedded in the world); and that the relevant situation (world) is *essentially* human. And these, I would like to argue, all come to the same thing: namely, to understand the possibility of intelligence is not to understand a property of some possibly isolable system, such as an "intellect", or a "mind" (≈ intellect + affect), or even an "agent" (≈ intellect + affect + body). Rather, it is to understand a larger whole comprising a number of cultured, embodied individuals living

[☆] (MIT Press, Cambridge, MA, 1992); liii + 354 pages, $13.95.

* E-mail: haugelan+@pitt.edu.

Elsevier Science B.V.
SSDI 0004-3702(95)00084-4

together in an already meaningful world. In what follows, I would like to give enough of a feel for this radical, threefold thesis that its plausibility and bearing on current research can be reassessed. (Readers familiar with recent discussions of "embedded" or "situated" cognition may get a sense of *déjà vu*—anachronistic, to be sure.)

2. Empirical metaphysics

First, a preliminary clarification: the claim about embodied situatedness is not epistemological but ontological. That is, it's not about prerequisites on theorizing, or on empirical discovery, or even on practical applications of the theories (e.g., boundary conditions). "Understanding the possibility of intelligence" means *making sense* of it as a genuine phenomenon—something accessible to investigation and explication at all. The idea is that intelligence apart from embodied living in a world doesn't make any sense—not only *is* there no such thing, but there *couldn't* be (it's *nonsense*). This thesis, of course, is substantive and tendentious: it's not at all a matter of armchair conceptual analysis or *a priori* reasoning from first principles. It may not be amenable to direct confirmation or disconfirmation, but it is nevertheless continuous with scientific research in just the way that foundational issues in quantum mechanics, evolutionary theory, economics, and so on, are continuous with more empirical scientific questions in their respective domains.

Ironically, I think, it's the scientists who are more often guilty of "armchair philosophizing" here, than are philosophers like Dreyfus. That is, it's typically the *opposition* to the necessary situatedness of intelligence that has been based on *a priori* presuppositions. (In a way, of course, this was the charge made by Part II of *What Computers Can't Do*.) For instance, opponents are tempted to retort: How *could* situatedness be essential to intelligence? After all, we *know* (that is: only a kook would deny) that intelligence is realized in the nervous system—mainly the neocortex, presumably—and that it interacts with the outside world only through specific sensors and effectors (transducers). To be sure (the objection continues), the brain needs metabolic support from the rest of the organism, and hence the environment; but those are "implementation details", in the sense that they are at the wrong level of description for the analysis of intelligence *per se*. When abstracted from such details, it is clear that *intelligence*—the essential functional structure that the nervous system implements—could be just what it is quite apart from either the body or the world.

Well, look at how much *metaphysics*—tendentious and substantive metaphysics—lurks unexamined in those comfortable phrases! Let us begin with "the essential functional structure that the nervous system implements". Which structure is *that*? Of course, the brain, like any reasonably complex physical system, exhibits any number of distinct abstract structures that could (in principle) be identified and defined apart from anything else. But how was it determined that intelligence is one of *those* structures? In fact, no such thing was ever determined at all—because the question was never asked. What we *know* is that, in the case of ordinary human beings, *something* about the structure of their brains is crucial to their intelligence. We know this because, if their brains get messed up, so does their intelligence. But that fact does not begin to show that anything about their brains *alone*—i.e., any structure that is *local* to their brains—is *sufficient* for intelligence as an intelligible phenomenon.

Needless to say, what suffices as a separately intelligible system (or subsystem) depends on the respects in which—or the "level" at which—it is to be understood. For instance, the central processing unit of a desktop computer is, from many points of view, a perfectly intelligible self-contained subsystem. But if one wants to understand the system as a word processor, then one cannot focus on the c.p.u. alone: word processing as such is intelligible only as a characteristic of a more encompassing system. Note: the point is *not* just that it must be "described at a higher level": the c.p.u. is simply the wrong place to look—it's too *small*—to understand word processing, at any "level" (which is not to deny that it's a crucial component, in the sense that the relevant system couldn't work without it). Likewise, the brain may be, from some points of view, a perfectly intelligible self-contained subsystem. But, if Dreyfus is right, human intelligence is intelligible only as a characteristic of a more encompassing system: the brain alone is the wrong place to look (at *any* level of description)—indeed, an entire individual organism may not be, by itself, encompassing enough.

3. Transduction and the body

How does one go about developing and defending such a point? What sorts of considerations are relevant to determining the scope and boundaries of intelligent systems? Dreyfus writes:

> Generally, in acquiring a skill—in learning to drive, dance, or pronounce a foreign language, for example—at first we must slowly, awkwardly, and consciously follow the rules. But then there comes a moment when we finally can perform automatically. At this point we do not seem to be simply dropping these same rigid rules into unconsciousness; rather we seem to have picked up the muscular gestalt which gives our behavior a new flexibility and smoothness. The same holds for acquiring the skill of perception. [248f]

A "*muscular* gestalt"? What have the *muscles* got to do with it? We react with questions like these, perhaps even with a trace of exasperation, because of a very seductive traditional story. When we are acting *intelligently*, our rational intellect is (consciously and/or unconsciously) taking account of various facts at its disposal, figuring out what to do, and then issuing appropriate instructions. These instructions are converted by output transducers into physical configurations (mechanical forces, electric currents, chemical concentrations, . . .) that result in the requisite bodily behavior. The transducers function as a kind of *interface* between the rational and the physical; and, as such, they provide a natural point of *subdivision*—in the sense that any alternative output subsystem that responded to the same instructions with the same behavior could be substituted without making any essential difference to the intellectual part. So, on that picture, the muscles would fall on the physical side, and not belong to the intelligent (sub)system at all.

Well, *are there transducers* between our minds and our bodies? From a certain all-too-easy perspective, the question can seem obtuse: *of course* there are. Almost by definition, there *has to be* a conversion between the symbolic or conceptual contents of our minds and the physical processes in our bodies; and that conversion just is transduction. But Dreyfus is, in effect, denying this—not by denying that there are minds or that there are bodies, but by denying that there needs to be an interface or conversion between them. Transduction, it's worth remembering, is the function that Descartes assigned to the pineal gland: it is required if and only if the mind and the body are fundamentally disparate and distinct—that is, intelligible in their own terms, quite apart from one another.

The fateful die is already cast in the image of the intellect figuring things out and then issuing instructions. What is an *instruction*? By fairly conventional wisdom, it is a syntactic expression which, by virtue of belonging to a suitably interpretable formal system, carries a certain sort of semantic content. Specifically, its content does *not* depend on how or whether it might be acted upon by any *particular* physical system. For instance, if I decide to type the letter "A", the content of the forthcoming instruction wouldn't depend on it being an instruction to *my* fingers, as opposed to any others, or even some robotic prosthesis. Any output system that could take that instruction and type an "A"—and, *mutatis mutandis*, other instructions and other behaviors—would do as well. The idea that there are such instructions is morally equivalent to the idea that there are transducers.

4. Output patterns that aren't instructions

A contrary—i.e., incompatible—view would be the following. There are tens of millions (or whatever) of neural pathways leading out of my brain (or neocortex, or whatever) into various muscle fibers in my fingers, hands, wrists, arms, shoulders, etc., and from various tactile and proprioceptive cells back again. Each time I type a letter, a substantial fraction of these fire at various frequencies, and in various temporal relations to one another. But that some particular pattern, on some occasion, should result in my typing an "A" depends on many contingencies, over and above whatever pattern it happens to be. In the first place, it depends on the lengths of my fingers, the strengths and quicknesses of my muscles, the shapes of my joints, and the like. In other words, there need be *no* way—even in principle, and with God's own microsurgery—to reconnect my neurons to anyone else's fingers, such that I could type with them. A set of connections that might work for one letter in one posture would be all wrong when it came to typing the next letter, or in a slightly different posture. But, in that case, what any given pattern "means" depends on it being a pattern specifically for *my* fingers—that is, for fingers with my "muscular gestalts".

Perhaps an analogy would help—even if it's a bit far fetched. Imagine an encryption system based on very large encryption keys, and such that all brief encrypted messages turn out to be comparable in size to the keys themselves (tens of millions of bits, just for instance). Now, consider one such message, and ask whether it could possibly mean anything *apart from its* particular *key*. It's hard to see how it could. Then the analogy works like this: each individual's *particular body*—his or her own muscular gestalts, so to speak—functions like a large encryption key, apart from which the "messages" are mere noise.

But even this may be overly sanguine. Whether a given efferent neural pattern will result in a typed "A" depends also on how my fingers happen already to be deployed and moving, how tired I am, which keyboard I'm using, and so forth. On different occasions, the same pattern will give different letters, and different patterns the same letter. In other words, there need be no similarity structure in the patterns themselves, at any level of description, that reliably correlates with the actions they produce. The reason that I can type, despite all this, is that there are comparably rich *afferent* patterns forming a kind of feedback loop that constantly "recalibrates" the system. (In terms of the above analogy, think of the encryption key itself constantly changing in real time.) But that would mean that the "content" of any given neural output pattern would depend not only on the particular body that it's connected to, but also on the *concrete details* of the current situation. Surely the idea of well-defined instructions and interchangeable transducers is hopeless here. But with no coherent notion of mental/physical transduction, the boundary, and hence the very distinction, between mind and body begins to blur and fade away.

5. The world itself as meaningful

Dreyfus, however, wants to go even further: the distinction between us and our world is likewise under suspicion. Of course, just as the brain is identifiable apart from the rest of the human organism for certain purposes, so also that organism is identifiable apart from its surroundings. The question is not whether the surface of the skin is easily discernible, but whether that surface amounts to an important division or interface when what we want to understand is human intelligence. Much of *What Computers Can't Do* is an attack on what might one be called "classical" or "symbol-processing" cognitive science. According to that view, internal symbolic representations constitute the realm of meaning in which intelligence abides.

Now Dreyfus, as everybody knows, emphatically rejects the primacy of internal *symbolic* representations. But he shares with his opponents the conviction that intelligence abides in the meaningful. So the question becomes: Is that realm of meaning that is the locus of intelligence representational at all, and is it bounded by the skin? Dreyfus' answer is clear:

> When we are at home in the world, the meaningful objects embedded in their context of references among which we live are not a model of the world stored in our mind or brain: *they are the world itself.* [265f; italics in original]

There are really two (closely related) points being made here: a negative point against cognitive science, and a positive point about the meaningful as such.

The negative thesis is simply a repudiation of the view, almost ubiquitous in cognitive science, that the meaningful objects amidst which intelligence abides are, in the first instance, *inner*. "Classical" cognitive scientists restrict these inner objects to *symbolic* expressions and models, whereas others are more liberal about mental images, cognitive maps, and maybe even "distributed representations". But Dreyfus wants to extend the meaningful well beyond the inner: meaningful objects are "the world itself". So, it's not just symbolic models that are being rejected, but the representational theory of the mind more generally—as we shall see more fully when we get to the positive thesis.

But first, we should guard against a misunderstanding. Everyone would allow, of course, that worldly objects can be meaningful in a *derivative* way, as when we assign them meanings that we already have "in our heads". You and I, for instance, could agree to use a certain signal to mean, say, that the British are coming; and then it would indeed mean that—but only derivatively from our decision. (Many philosophers and scientists would hold further that this is the *only* way that external objects can come to be meaningful.) By contrast, when Dreyfus says that meaningful objects are the world itself, he means *original* meaning, not just derivative. That is, intelligence itself abides "out" in the world, not just "inside"—contra cognitive science, classical or otherwise.

6. Not just representations

The positive thesis, alas less clearly worked out, is about the meaningful as such. If I may try to put it into my own words, I would say (very roughly) that the meaningful is that which is significant in terms of something beyond itself, and subject to normative evaluation according to that significance. *Representations* are familiar paradigms of the meaningful in this sense; and when cognitive scientists speak of meaningful inner objects, they *always* mean representations (the only dispute being about the possible forms of these representations—i.e., whether they're symbolic). That in terms of which a representation is significant is that which it purports to represent—its *content*—and it is evaluated according to whether it represents that content correctly or accurately. Descartes, in effect, *invented* the "inner realm" as a repository for cognitive representations—above all, representations of what's outside of it. Cognitive science hasn't really changed this; in particular, nothing *other than* representations has ever been proposed as inner and meaningful.

But when Dreyfus holds that meaningful objects are the world itself, he doesn't just (or even mostly) mean representations. The world can't be representation "all the way down". But that's not to say that it can't all be meaningful, because there are more kinds of significance than representational content. A number of philosophers earlier in the twentieth century—Dewey, Heidegger, Wittgenstein, and Merleau-Ponty, to name a few of the most prominent—have made much of the significance of equipment, public places, and community practices; and Dreyfus has these very much in mind. A hammer, for instance, is significant beyond itself in terms of *what it's for*: driving nails into wood, by being wielded in a certain way, in order to build something, and so on. The nails, the wood, the project, and the carpenter him or herself, are likewise caught up in this "web of significance", in their respective ways. These are the meaningful objects that are the world itself (and none of them is a representation).

There's an obvious worry here that the whole point depends on a pun. *Of course*, hammers and the like are "significant" (and even "meaningful") in the sense that they're *important* to us, and *interdependent* with other things in their proper use. But that's not the same as meaning in the sense of bearing content or having a semantics. Certainly! That's why they're not representations. So it's agreed: they are meaningful in a broader sense, though not in a narrower one. The real question is: Which sense matters in the context of understanding human intelligence?

7. Intelligence and the meaningful

To address this question, we ask what meaningfulness has to do with intelligence in the first place. To say that intelligence *abides in* the meaningful is not to say that it is surrounded by or directed toward the meaningful, as if they were two separate things. Rather, intelligence has its very existence in the meaningful as such—in something like the way a nation's wealth lies in its productive capacity, or a corporation's strength consists in its market position. Intelligence is nothing other than the overall interactive and interdependent structure of meaningful behavior and objects.

Perhaps the basic idea can be brought out this way. Intelligence is the ability to deal reliably with more than the present and the manifest. That's surely not an adequate definition of intelligence, but it does get at something essential, and, in particular, something that has to do with meaning. Representations are clearly an asset in coping with the absent and covert, insofar as they themselves are present, and "stand in for" something else which they "represent". *How* can they do that? A typical sort of story goes like this. Individual representations can function as such *only* by participating, in concert with many others, in a larger and norm-governed *scheme* of representation. Then, assuming the scheme itself is in good shape, and is used correctly, a system can vicariously keep track of and explore absent and covert represented phenomena by keeping track of and exploring their present and manifest representational stand-ins. (Really, what it means for a scheme to be "in good shape" is for this coping at-one-remove to be generally workable.) In effect, the structure of the extant representations, in conjunction with that of the scheme itself, "encodes" something of the structure of what is represented, in such a way that the latter can be accommodated or taken account of, even when out of view.

To abide in the meaningful is to abide in those structures, both inert and dynamic, that make this extended effectiveness possible. It's clear enough how representations fill the bill. But, equally, it should be clear how tools, structured places, and institutionalized practices extend present capacity by "encoding" the unobvious. To take the crudest example: the problem of securing shelter from the wild is anything but simple and manifest. It took our forebears many generations to work out the basic solutions that we now take more or less for granted. And *how*, exactly, are these solutions "granted"? Well, in various ways; but one of the most important is in the shapes and qualities of hammers and nails, boards and saws, along with our standard practices for using them. These do not "represent" the accumulated wisdom of our ancestors, at least not in any semantic sense, but they do somehow incorporate it and convey down to us in a singularly effective manner.

Here's another angle on much the same idea. A c.p.u. cannot be understood as "word processing" all by itself; only in conjunction with (at least) some suitable software and processable text (both appropriately accessible in RAM), plus pertinent input/output facilities (keyboard, display, disk, and printer, e.g.), can there be word processing going on at all. But none of these others suffices in isolation either: the software is only suitable given the way the c.p.u. responds to it and the keyboard, by modifying the text and the display; the RAM contents are only text given the way . . . ; and so on.

Classical cognitive science and AI wanted to import an essentially comparable structure *into* the brain. And, exactly parallel to the word processing, you only got intelligence when you considered all the parts in conjunction—the primitive operations (cognitive architecture), learned expertise (scripts, productions, common sense), current models and plans, etc. These would all be characterized at the relevant level of description—a level which, however, only makes coherent sense when all the parts are characterized together. So what we have is a kind of *holism* (except it's broader than the familiar holism of semantic interpretation, because it includes processors and real-time operations). In particular, the fundamental meaningful activities and objects, in which the system's intelligence abides, are only meaningful in virtue of the way the overall system works as a whole.

Now Dreyfus can be understood as proposing yet a third variation on the same basic picture; only, in his version, each person is effectively just a "processing unit" (in a multi-processor configuration), and the relevant operations take place in the public world. That is, the meaningful "data-structures" on which they operate are public objects—hammers, cities, movies, election campaigns, corporations, technologies, revolutions—and these meaningful objects *are the world itself.* As before, the meaningfulness and intelligence reside only in the integrated whole—and, in particular, not in the "processors" alone. Nevertheless, those processors are crucial, in the sense that without them, none of the rest of it would work, and the whole structure would collapse.

8. This world is our world

From here it is but a short step to my third and final point. Dreyfus says:

> The human world . . . is prestructured in terms of human purposes and concerns in such a way that what counts as an object or is significant about an object already is a function of, or embodies, that concern. [261]

What does he mean by the *human* world? Obviously, he means the world all around us, the one we live in every day. But that might still be misleading, for it might be taken to imply a contrast: our (human) world as opposed to some other or others—animal worlds, God's world, or something like that. But there's only *one* world, *this* one—and it's ours.

Well, in what sense is it "ours"? Surely not just that our species has overrun the planet, and, in the meantime, is arrogant beyond compare. No, it is our world in the sense that we understand it and live it: it is the substance and shape of our lives. Now it goes without saying that the world is multifarious, to the extent that we sometimes speak of "different" worlds: the world of the theater, the wide world of sports, the brave new world of cognitive science, not to mention the nomadic Bedouin world, the agrarian Hopi world, and so on. But these are all still human, all still understood, all still the meaningful abode of human intelligence. In my own view (and I suspect also Dreyfus'), there is no such thing as animal or divine intelligence. But if that's wrong, it only extends the scope of who "we" are. And the same can be said about the possibility of intelligent extraterrestrials. The *world* just is the realm of the meaningful; in other words, it is where intelligence abides.

But what about the physical universe: countless stars and galaxies, vast mindless forces, fifteen billion years of atoms and the void? Isn't that the *real* world, a fleeting speck of which we happen to throw an interpretation over, and regard as meaningful? No, that's backwards. The physical universe is *a part of* our world. It is a peculiar and special part, because a peculiar and special intelligence abides in it—namely, an intelligence that has invented the meaning "meaningless", and made that hang together in a new kind of whole. This, however, is not the place to examine the sense that the physical makes, but only to note that it is *not* primary. Accordingly, cognitive science would be trying to build from the roof down if it began its investigations of intelligence with our understanding of physics. The foundations of intelligence lie not in the abstruse but in the mundane.

9. Conclusion

I started by proposing a return to Part III of *What Computers Can't Do*, attributing to it three principal theses: that human intelligence is *essentially* embodied; that intelligent bodies are *essentially* situated (embedded in the world); and that the relevant situation (world) is *essentially* human. And I suggested that these all come to the same thing. What they all come to, we can now see, is the radical idea that intelligence abides bodily in the world. If this is right—as I believe it is—and if science is ever to understand it, then research agenda must expand considerably. Not only is symbolic reasoning too narrow, but so is any focus on internal representation at all. When cognitive science looks for its closest kin, they will not be formal logic and information processing, but neurobiology and anthropology.