AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

# Dreyfus on AI: A Lonerganian Retrieval and Critique

Michael Sharkey
University of Wisconsin-Platteville, USA

**Abstract**
Hubert Dreyfus develops a critique of AI which should interest readers of Bernard Lonergan. He contests its early rationalism in a way that resembles Lonergan's critique of conceptualism. He contests its early representationalism in a way that resembles Lonergan's critique of ocularism. And he makes both criticisms from a cognitional-theoretical perspective which privileges "insight," like Lonergan's. However, Dreyfus ultimately gives short shrift to consciousness, intentionality, and acts, which leads him to throw out the mentalist baby with the conceptualist and ocularist bath. The result is an excessive receptivity to recent (especially neural network) AI, which reduces intelligence to electrical events.

**Keywords:** Hubert Dreyfus, Bernard Lonergan, artificial intelligence, insight, problems of consciousness.

## Introduction

In publications running from *What Computers Can't Do* (1972, 1978) through *Mind Over Machine* (1986) to *What Computers* Still *Can't Do* (1992) and "Why Heideggerian AI Failed" (2007), Hubert Dreyfus develops a critique of artificial intelligence that should interest readers of Lonergan.[1] He shows first variants of the project to possess rationalist philosophical presuppositions and criticizes them in ways that resemble Lonergan's critique of conceptualism. He shows second variants to be in the grips of a representational theory of knowledge and criticizes them in ways that resemble Lonergan's critique of ocularism. And he offers both sets of critique from out of his own cognitional-theoretical perspective, centered as it is on what he entitles "insight."[2]

However, Dreyfus's stance is not fully positional, and this compromises his critique of AI.[3] His method sits uneasily between phenomenology and metaphysics, in the manner of the early Heidegger and Merleau-Ponty. This leads him to give short shrift to consciousness, intentionality, and acts, which in turn leads him to throw out the mentalist baby with the conceptualist and ocularist bath. The result is an undue receptivity to recent (neural network) AI, which reduces intelligence to electrical events.[4]

Both a retrieval and a critique, then, would seem to be in order. In a first part below, I will relate Dreyfus's interpretation and critique of AI, in both its early and more recent variations. In a second, I will explain why I think much of his treatment is consistent with a positional stance. And in a third, I will explain why I think some of his (counter) positions stand in need of reversal.

---

[1] Hubert L. Dreyfus, *What Computers Can't Do: A Critique of Artificial Reason* (New York: Harper Collins, 1972, 1978), *Mind over Machine* (New York: Free Press, 1986), *What Computers* Still *Can't Do* (Cambridge, MA: MIT Press, 1992), and "Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian," in Mark A. Wrathall, ed, *Skillful Coping: Essays on the Phenomenology of Everyday Perception & Action* (Oxford: Oxford University Press, 2016).

[2] Lonergan's masterwork is *Insight: A Study of Human Understanding,* volume 3 of *Collected Works of Bernard Lonergan*, ed. Frederick E. Crowe and Robert M. Doran (Toronto: University of Toronto Press, 1992).

[3] A stance is "positional," for Lonergan, if it cannot be denied without performative contradiction. See Bernard Lonergan, *Insight*, 313-15. And for a rebuttal of the charge that the doctrine is question-begging, see Mark D. Morelli, "Reversing the Counter-Position: The *Argumentum ad Hominem* in Philosophic Dialogue," in Frederick Lawrence, ed., *Lonergan Workshop*, volume 6 (Macon, Georgia: Scholars Press, 1986), 195-230.

[4] I owe the important distinction between an act and an event in this context to Elizabeth Murray.

## I. Dreyfus on AI

### *(A) Early*

In *What Computers* Still *Can't Do* and *Mind over Machine*, Dreyfus shows early variants of the project of AI to possess rationalist philosophical presuppositions. The presuppositions derive from the epistemological programs of Socrates, Descartes, Hobbes, Leibniz, Kant, and Husserl, and tell us that intelligence is a matter of representations and rules.

For Dreyfus, Socrates is a semantic rationalist. He demands that Euthyphro tell him " . . . what is characteristic of piety which makes all actions pious . . . that I may have it to turn to, and to use as a standard whereby to judge your actions and those of other men."[5] Uninterested in this or that example, as rooted in Athenian culture, he requires a general concept or universal definition articulating the necessary and sufficient conditions of the virtue. With one in hand, he might avoid the contingency and imprecision which characterize practical reason. Or so he thinks. He is thus the distant inspiration for AI's "effective procedure" or "set of rules which tells us, from moment to moment, precisely how to behave."[6]

Things are little different with Descartes, Kant, and Husserl. Descartes claims that one can "analyze any problem into its basic, isolatable elements, and explain the complex in terms of rule-like combinations of such primitives."[7] Thus he intuits with certainty that he thinks, deduces that he exists and is a thinking thing, and proceeds therefrom to build up an edifice of new knowledge. Kant holds that "all concepts are really rules,"[8] shows some necessarily to apply to objects of knowledge, and establishes a tribunal of pure reason. Husserl takes concepts to be "hierarchies of rules, rules which contain other concepts under them," and so shows himself to be "father of the information-processing model of the mind."[9]

Things are different, and yet the same, with Hobbes and Leibniz. They are not semantic but syntactic rationalists who would reduce "all . . . appeal to meanings . . . to the techniques of . . . formal . . . manipulation."[10] But they continue to think of intelligence in terms of representations and rules. "When a man *reasons*," Hobbes says, "he does nothing else but conceive a sum total from addition of parcels, for REASON . . . is nothing but reckoning."[11] And Leibniz develops a "universal and exact system

---

[5] Plato, *Euthyphro*, VII, trans. F. J. Church (New York: Library of Liberal Arts), 1948, 7, as quoted in Dreyfus *What Computers* Still *Can't Do*, 67.
[6] Marvin Minsky, *Computation: Finite and Infinite Machines* (Englewood Cliffs, N.J.: Prentice-Hall, 1967), 106, as quoted in Dreyfus, *What Computers* Still *Can't Do*, 67.
[7] Hubert Dreyfus, *Mind over Machine*, 3. Italics removed.
[8] Hubert Dreyfus, *Mind over Machine*, 4.
[9] Hubert Dreyfus, *Mind over Machine*, 4.
[10] Hubert Dreyfus, *What Computers* Still *Can't Do*, 69. Parentheses removed.
[11] Thomas Hobbes, *Leviathan* (New York: Library of Liberal Arts, 1958), 45, as quoted in Dreyfus, *What Computers* Still *Can't Do*, 69.

of notation, an algebra, a symbolic language" to which concepts can be reduced. On their basis "and the rules for their combination all problems [can] be solved and all controversies ended."[12] Leibniz writes that if someone were to contest his results, he would say to him, "'Let us calculate, Sir,' and thus by taking pen and ink, we should settle the question."[13]

Semantic and syntactic rationalism drive early AI. The successor to the latter, Cognitive Simulation, means "to reproduce the steps by which human beings actually proceed," whereas the successor to the former, Semantic Information Processing, means just to achieve the same results.[14] But between them they take concepts to be rules, of a kind, or to be formal stand-ins for meanings which, when manipulated by rules, produce intelligence. They thus incarnate the commitment to representations and rules.

Among examples of Cognitive Simulation, Dreyfus considers programs for playing games, translating languages, solving problems, and recognizing patterns. Among examples of Semantic Information Processing, he considers programs for understanding language and finding analogies.

Newell and Simon's program for playing chess is a fine example of Cognitive Simulation. Chess is a game in which pieces of varying capacity are moved across a board in a rule-like way to achieve certain ends. Intelligent play involves finding the best means of achieving those ends. So a computer program for playing chess must include at least representations (or definitions) of the pieces and a list of the rules for manipulating them. But it must include more, for of course there is a difference between intelligent and unintelligent manipulation. Enter what Newell and Simon call "heuristics," or "rules of practice," or "rules of thumb," gleaned from the greats. These are not rules followed invariably but just occasionally in order to reduce calculation. They are "aids to discovery" meant to replicate the judgment *in situ* that is characteristic of human play.[15]

Another example of Cognitive Simulation is Oettinger's Russian-English dictionary. On one understanding of how language works, such as is to be found in Augustine's *Confessions* and Wittgenstein's *Tractatus*, there is a one-to-one correspondence between words and things, or sets of words and states of affairs. A dictionary translating from one language to another, then, must exhaustively correlate the more or less complex correspondences on each side. "It was soon clear that a mechanical dictionary could easily be constructed in which linguistic items, whether they were parts of words, whole words, or groups of words, could be processed independently and converted one after another into corresponding items in another

---

[12] Hubert Dreyfus, *What Computers* Still *Can't Do*, 69.
[13] Leibniz, *Selections*, ed. Philip Wiener (New York: Scribner, 1951), 18, as quoted in Dreyfus, *What Computers* Still *Can't Do*, 69.
[14] Hubert Dreyfus, *What Computers* Still *Can't Do*, 85.
[15] Hubert Dreyfus, *What Computers* Still *Can't Do*, 74-77, 94, 102-107.

language."[16] In this way it was thought the difficulties in understanding a foreign tongue could be reduced to low-level matching.

A striking example is Newell, Simon, and Shaw's General Problem Solver, which sought "rules for converting any sort of intelligent activity into a set of instructions." But again, because studies showed subjects "tended to use rules or shortcuts which were not universally correct, but which often helped," heuristics were employed. If, in solving logic problems, "[s]uch a rule of thumb might be, . . . try to substitute a shorter expression for a longer one,"[17] or if, in playing chess, it might be "maintain center position" or "sacrifice queen,"[18] in this context it was held that by generalizing such strategies the human capacity for solving problems in any area could be mimed.

> In short, we now have the elements of a theory of heuristic (as contrasted with algorithmic) problem-solving; and we can use this theory both to understand human heuristic processes and to simulate such processes with digital computers. Intuition, insight, and learning are no longer exclusive possessions of humans; any large high-speed computer can be programmed to exhibit them also.[19]

Last examples of Cognitive Simulation come from pattern recognition. They are programs for transliterating hand-sent Morse code, as well as for "recognizing a limited set of handwritten words and printed characters in various type fonts."

> These all operate by searching for predetermined topological features of the characters to be recognized, and checking these features against preset or learned "definitions" of each letter in terms of these traits. The trick is to find relevant features, that is, those that remain generally invariant throughout variations of size and orientation, and other distortions.[20]

Here, the human capacity to discern according to necessary and sufficient conditions is modelled.

Turning to Semantic Information Processing, Bobrow's STUDENT program is exemplary. It makes no pretense to the humanoid, but still solves algebra word problems and "understands English."[21]

---

[16] Hubert Dreyfus, *What Computers* Still *Can't Do*, 91.
[17] Hubert Dreyfus, *What Computers* Still *Can't Do*, 75.
[18] Hubert Dreyfus, *What Computers* Still *Can't Do*, 101-102.
[19] Herbert A. Simon and Allen Newell, "Heuristic Problem Solving: The Next Advance in Operations Research," *Operations Research*, Vol. 6 (January—February, 1958), 6, as quoted in Hubert Dreyfus, *What Computers* Still *Can't Do*, 77.
[20] Hubert Dreyfus, *What Computers* Still *Can't Do*, 97.
[21] According to Marvin Minsky, in his "Artificial Intelligence," *Scientific American,* Vol. 215, No. 3 (September 1966), 257, as quoted in Hubert Dreyfus, *What Computers* Still *Can't Do*, 132.

> The program simply breaks up the sentences of the story problem into units on the basis of cues such as the words "times," "of," "equals," etc.; equates these sentence chunks with x's and y's; and tries to set up simultaneous equations. . . . [T]he . . . scheme works . . . because there is the constraint, not present in understanding ordinary discourse, that the pieces of the sentence, when represented by variables, will set up soluble equations.[22]

In other words, the program reduces typical human expression to algebraic formalism and rules.

A final example of Semantic Information Processing is Evan's Analogy Finder. It does not purport to reproduce human intelligence any more than does Bobrow's STUDENT, yet it too is set out in mentalistic terms. "Given a set of figures, [the program] constructs a set of hypotheses or theories as follows." First, a description of figure A may be transformed into one for B. Second, the parts of A may be set into correspondence with the ones for C, suggesting a relation like the first, but now relating C and other figures. Third, the differences between C and another figure may be reduced to the same degree as between A and B, so that, Fourth, it may be determined that A stands to B as C does to, say, D3, this having been determined by measurement.[23] Evans's editor even adds that he feels sure "rules or procedures of the same general character are involved in any kind of analogical reasoning."[24]

Now, Dreyfus does not take any of these programs to rise to the level of intelligence. He takes the examples from Cognitive Simulation to fail to do so because they do not employ "fringe consciousness," "contextually disambiguate," "distinguish the essential from the inessential," and "perspicuously group," as do all human beings when behaving intelligently. And he takes the examples from Semantic Information Processing to fail to do so because they do not have "bodies," are not "in situations," and do not have "needs."[25] He offers a hermeneutic-phenomenological argument for the view that human intelligence involves more than rule-following and representing.

As against Newell and Simon's program for playing chess, Dreyfus points out that human beings do more than count out possible moves and responses and occasionally employ rules of thumb. For "[a]lternative paths multiply so rapidly that we cannot . . . run through all the branching possibilities" and it is necessary not just to "look . . . every once in a while for a Queen sacrifice but . . . look in those situations in which such a sacrifice is relevant."[26] For this, "fringe consciousness" is required. It

---

[22] Hubert Dreyfus, *What Computers* Still *Can't Do*, 133.
[23] Marvin Minsky, ed., *Semantic Information Processing* (Cambridge, Mass.: M.I.T. Press, 1969), 16, as quoted in Hubert Dreyfus, *What Computers* Still *Can't Do,* 139.
[24] Marvin Minsky, "Artificial Intelligence," *Scientific American,* Vol. 215, No. 3 (September 1966), 250, as quoted in Hubert Dreyfus, *What Computers* Still *Can't Do*, 139. Dreyfus's italics removed.
[25] In fact, what Dreyfus says here applies to programs from Cognitive Simulation too. But since being embodied, being in situations, and having needs are central to his overcoming of representationalism, his primary target is programs for Semantic Information Processing, with their emphasis on representations more than rules.
[26] Hubert Dreyfus, *What Computers* Still *Can't Do*, 101.

is "marginal awareness" that "concentrate[s] information concerning our peripheral experience."[27] In virtue of it, promising areas of the board may be identified.

Consider the following player's report. "Again I notice that one of his pieces is not defended, the Rook, and there must be ways of taking advantage of this. Suppose now, if I push the pawn up at Bishop four, if the Bishop retreats I have a Queen check and I can pick up the Rook."[28] At the end, Dreyfus notes, "we have an example of … "counting out"—thinking through the various possibilities by brute force enumeration." But at the start, we have something very different, a kind of sussing out, perhaps. "[T]he subject "zeroed in" on the promising situation."[29]

As against Oettinger's program for machine translation, Dreyfus calls attention to context. It invariably produces ambiguity in expression, which makes one-to-one translation difficult. It therefore turns out that "in order to translate a natural language, more is needed than a mechanical dictionary—no matter how complete—and the laws of grammar—no matter how sophisticated." For "[t]he order of the words in a sentence does not provide enough information to enable a machine to determine which of several possible parsings is the appropriate one, nor do the surrounding words—the written context—always indicate which of several possible meanings . . . the author had in mind."[30] What is required is "contextual disambiguation."

"A phrase like 'stay near me,'" Dreyfus writes, "can mean anything from 'press up against me' to 'stand one mile away,' depending upon whether it is addressed to a child in a crowd or a fellow astronaut exploring the moon."[31] And human beings can determine which is which. Again, a child can learn the names of things without being unduly thwarted by situational change. "It is this ability to grasp the point in a particular context which is true learning; since children can and must make this leap, they can and do surprise us and come up with something genuinely new."[32]

As against Newell, Simon, and Shaw's program for general problem solving, Dreyfus presses this point about getting the point. "[I]nsight," he declares, "has proved intractable to stepwise programs such as Simon's General Problem Solver."

> If a problem is set up in a simple, completely determinate way, with an end and a beginning and simple, specifically defined operations for getting from one to the other, . . . then Simon's General Problem Solver can, by trying many possibilities, bring the end and the beginning closer and closer together until the problem is solved.[33]

---

[27] Hubert Dreyfus, *What Computers* Still *Can't Do*, 103.
[28] Allen Newell and H. A. Simon, *Computer Simulation of Human Thinking*, The RAND Corporation, P-2276 (April 20, 1961), 15, as quoted in Hubert Dreyfus, *What Computers* Still *Can't Do*, 102.
[29] Hubert Dreyfus, *What Computers* Still *Can't Do*, 102.
[30] Hubert Dreyfus, *What Computers* Still *Can't Do*, 107.
[31] Hubert Dreyfus, *What Computers* Still *Can't Do*, 108.
[32] Hubert Dreyfus, *What Computers* Still *Can't Do*, 111.
[33] Hubert Dreyfus, *What Computers* Still *Can't Do*, 112.

Or it can do so in concert with heuristics. But when the problem is complex more than slavish rule-following is required and the heuristics themselves can be seen to be nothing more than that. This is borne out by analysis of the reports given by human beings while they are solving problems.

Consider the example of one such 'protocol' given by a person solving a problem in logic. In it he reports that having received a list of rules for transforming symbolic expressions, he applied "the rule (A · B → A) and the rule (A · B → B), to the conjunction ( — R v — P) · (R v Q)." Newell and Simon note that in so doing he "handled both forms of rule 8 together," whereas their machine "took a separate cycle of consideration for each form." But they assume that the subject "covertly" took each form in turn, while Dreyfus notes that, on the face of it, he "grasped the conjunction as symmetric with respect to the transformation operated by the rule, and so in fact applied both forms of the rule at once." That is, Dreyfus shows that the phenomenological evidence suggests the subject had an insight. He was able to "discriminate between occasions when it is was appropriate to apply both forms of the rule at once and those occasions when it was not."[34]

Again, "[a]t a certain point, the protocol reads: " . . . I should have used rule 6 on the left-hand side of the equation. So use 6, but only on the left-hand side." Simon sees that "[h]ere we have a strong departure from the GPS trace," for "[n]othing exists in the program that corresponds to this." And "[t]he most direct explanation," he avers, "is that the application of rule 6 in the inverse direction is perceived by the subject as undoing the previous application of rule 6." He seems to recognize the act of insight. But he does not see that this counts against his approach.[35]

Part of the explanation for this must be that Newell and Simon think they have covered the phenomenon of insight with heuristics. Such aids in discovery are supposed to take the program beyond the automatic to the selective, but in fact they just take it beyond the invariant to the occasional. And the programmers determine what counts as occasional. It is this "insightful predigesting of their material" that enables them to pass off as intelligent what is just mechanical.[36]

Lastly, as against the programs for pattern recognition, Dreyfus contests the primacy of the concept. "A computer must recognize all patterns in terms of a list of specific traits," he notes. And "in simple cases artificial intelligence workers have been able to make some headway with mechanical techniques." But "patterns as complex as artistic styles and the human face reveal a loose sort of resemblance which seems to require a special combination of insight, fringe consciousness, and ambiguity tolerance beyond the reach of digital machines."[37] This Dreyfus calls "perspicuous grouping."

Consider even the apparently simple task of identifying a shape. How do we do it? We are not, most of us, like aphasics, who "can only recognize a figure such as

---

[34] Hubert Dreyfus, *What Computers* Still *Can't Do*, 113.
[35] Hubert Dreyfus, *What Computers* Still *Can't Do*, 113-114.
[36] Hubert Dreyfus, *What Computers* Still *Can't Do*, 119.
[37] Hubert Dreyfus, *What Computers* Still *Can't Do*, 120.

a triangle by listing its traits, that is, by counting its sides and then thinking: 'A triangle has three sides. Therefore, this is a triangle.'" We do not need to "conceptualize . . . the traits common to several instances of the same pattern in order to recognize that pattern."[38] We do not need to employ a classification rule. Instead, we zero in on relevance and grasp the point in a context, irrespective of some ambiguity.

We can see that Dreyfus's main reservation about Cognitive Simulation is its emphasis on rule-following. By contrast, his main reservation about Semantic Information Processing is its emphasis on semantics. But for Dreyfus "semantics" always has to do with "representation," and we will be able better to see his critique of it if we turn to material beyond Bobrow's STUDENT and Evans's Analogy Finder.

It is true that we must use our bodies in order to see, hear, taste, touch and smell. In the language of early AI theorist Marvin Minsky, such "meat machine" operation is essential. But it is not sufficient, according to Dreyfus, for we must also use our "lived bodies" to get at meanings. We do not just receive sense-impressions, re-present those presentations to ourselves, and string the representations together to form ideas and thoughts. We are aware of ourselves as sensing, and indeed as seeking understanding, which supplies us with a "global anticipation" in whose light we make sense of parts.[39]

For example, "in recognizing a melody, the notes get their values by being perceived as part of the melody, rather than the melody's being recognized in terms of independently identified notes." Similarly, the "hazy layer which I would see as dust if I thought I was confronting a wax apple might appear as moisture if I thought I was seeing one that was fresh."[40] My gulp of milk will leave me disoriented if what I was expecting was water.[41] And I will be unable to identify silk as silk, if I lack the appropriate anticipations developed in me by long familiarity with fabric.[42] It is only because I am anticipatorily involved with my world, that I am able to understand any bit of it. But machines lack embodiment, and so lack the condition of the possibility of understanding.

Again, it is because I am in situations that I am able to affix meanings correctly. On a walk I know that my friend's gesture towards "the Old Man of the Woods" refers to a plant and not a person.[43] In front of a pet store I know my daughter's desire for "it" refers to a doggie and not the window.[44] In hearing from a gift-giver that I "can take it back if I already have one," I know he means the item he has given and not the one I may already have.[45] And in a Berkeley restaurant, I know the suggestion to "order

---

[38] Hubert Dreyfus, *What Computers* Still *Can't Do*, 123.
[39] Hubert Dreyfus, *What Computers* Still *Can't Do*, 237.
[40] Hubert Dreyfus, *What Computers* Still *Can't Do*, 238.
[41] Hubert Dreyfus, *What Computers* Still *Can't Do*, 242.
[42] Hubert Dreyfus, *What Computers* Still *Can't Do*, 249.
[43] I feel sure this example, borrowed from Wittgenstein, is in one of Dreyfus's texts. But I am unable to find it.
[44] Hubert Dreyfus, *What Computers* Still *Can't Do*, xix.
[45] Hubert Dreyfus, *What Computers* Still *Can't Do*, 57.

anything" does not include the chef.[46] My ability to understand depends on familiarity with situations and their criteria. But machines are not in situations, and so they cannot "compute."

Finally, both my embodiment and being-in-situations are tied up with needs. It is because I require nourishment, both physical and aesthetic, that I listen to melodies, look at apples, drink water, and touch silk. And it is because I need love and friendship that I walk with friends, spend time with my daughter, have birthday parties, and go to restaurants. My ability to understand, therefore, is rooted not just in embodiment and being in situations, but in the needs which drive me to both. And yet computers do not have needs any more than are they embodied or in situations. This is another reason why they are blocked from cognition.[47]

In summary, Dreyfus criticizes early AI because it models intelligence on representations and rules. Its first variant, Cognitive Simulation, emphasizes rule-following, and so ignores the fringe consciousness, contextual disambiguation, insight, and perspicuous grouping which are essential to the real article. And its second variant, Semantic Information Processing, emphasizes representations, and so ignores the embodied anticipation, situational sensitivity, and neediness which are the conditions of representation. Both variants are indebted to the rationalist tradition in Western philosophy, against which Dreyfus would set Heidegger, Wittgenstein, and Merleau-Ponty.[48] However, as we will see, this surprisingly does not stop him from endorsing AI of a kind.

*(B) Recent*

Dreyfus is more sanguine about the prospects for recent, neural network AI, and this precisely because it does not employ representations and rules. Instead of trying to make a mind, as at least Cognitive Simulation did, it seeks to model the brain; and Dreyfus believes it is partly on its way.

In *What Computers* Still *Can't Do*, Dreyfus argues that "we should set about creating artificial intelligence by modelling the brain's learning power rather than the mind's symbolic representation of the world" because of what we have learned from neuroscience. Already in the '50's that discipline had suggested that "a mass of neurons could learn if the simultaneous excitation of neuron A and neuron B increased the strength of the connection between them." In the present, then, AI might "attempt to automate the procedures by which a network of neurons learns to discriminate patterns and respond appropriately."[49] But how? "[A] designer could tune a simulated

---

[46] Hubert Dreyfus, *What Computers* Still *Can't Do*, 311, note 102.
[47] Hubert Dreyfus, *What Computers* Still *Can't Do*, 276-280.
[48] Hubert Dreyfus, *What Computers* Still *Can't Do*, 212, 233. And see Dreyfus, *Mind Over Machine*, 4-5, 7 and 11.
[49] Hubert Dreyfus, *What Computers* Still *Can't Do*, xiv.

multilayer perceptron (MLP) neural network by training it to respond to specific situations and then having it respond to other situations in ways that are (the designer hopes) appropriate extrapolations of the responses it has learned." In this case the modeler "provides not rules relating features of the domain but a history of training input-output pairs, and the network organizes itself by adjusting its many parameters so as to map inputs into outputs, situations into responses."[50]

Consider a famous example. In order better to wage the Gulf War, a neural net was trained to distinguish rocks from mines at the bottom of a sea. First, visual and sonar data on these items was assembled. Second, our (or our brain's) ability to identify patterns in this data was modelled by "input and output nodes," "middle layer nodes," and the variable strengths of their relations expressed as "weights." Third, an expert at identifying and distinguishing rocks and mines "tuned" the network of nodes-in-their-relations (adjusted their relative strengths) to correspond to that obtaining in the world. And fourth, the network was afterwards able to discriminate on its own.[51]

Dreyfus even argues this approach is consistent with phenomenology. In "Merleau-Ponty and Recent Cognitive Science," he draws a parallel between understanding as Merleau-Ponty conceives of it and understanding as modelled by neural nets. Just as, for Merleau-Ponty, "the life of consciousness—cognitive life, the life of desire or perceptual life—is subtended by an 'intentional arc,' which projects round about us our past, our future, our human setting,"[52] and so establishes a "dialectic of milieu and action," so for neural net AI "past experience with a large number of cases . . . modifies the weights between the simulated neurons, which in turn determine the response." In neither case is there need to "represent . . . past experience as cases or rules for determining further action," and in both it is thus possible "to avoid the problem . . . concerning how to find the *relevant* rule."[53]

Again, in "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," Dreyfus likens understanding as Heidegger conceives of it to Freeman's Neural Dynamics. For Heidegger, understanding is an affair of practical know-how, of knowing one's way around in the world. It is "more basic than *thinking* and solving problems" and is "not representational at all." In fact, in understanding at our best, "we are drawn in by solicitations and respond directly to them, so that the distinction between us and our equipment vanishes."[54] "I *live* in the

---

[50] Hubert Dreyfus, *What Computers* Still *Can't Do*, xv.

[51] R. Paul Gorman and Terence J. Sejnowski, "Learned Classification of Sonar Targets Using a Massively Parallel Network," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* 36/7 (July 1988), 1135-40, referenced in Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," Mark Wrathall, ed., *Skillful Coping,* 238, note 12.

[52]Maurice Merleau-Ponty, *Phenomenology of Perception*, tr. Colin Smith (London: Routledge Classics, 2002), 136, as quoted in Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," 234.

[53] Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," 236.

[54] Hubert Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," 258.

understanding of writing, illuminating, going-in-and-out, and the like," Heidegger says. And "[my] being in the world *is* nothing other than this . . . understanding."[55]

It is much the same in Freeman's dynamics. He "proposes a model of rabbit learning based on the coupling of . . . brain and . . . environment," and to the degree that these remain distinct they stand in circular relation. The rabbit is thrown on to a horizon of longing. It "sniffs around until it falls upon food, a hiding place, or whatever else it . . . needs." Its "neural connections are then strengthened to the extent that" it is satisfied. And its new configuration of synapses-in-relation contextualizes further desire.[56] No representations or rules are required. Only a kind of natural analogue of the hermeneutic circle.

Or so it might seem. However, Dreyfus is alert to some limitations of neural modelling. As against the (putatively) Merleau-Pontyan version, he argues that the problem of relevance resurfaces. "When a net is trained by being given inputs paired with appropriate responses," he writes, "the net can only be said to have learned to respond appropriately when it responds appropriately to *new* inputs similar to, but different from, those used in training it." Otherwise, it may seem just to have engaged in the low-level matching characteristic of GOFAI. Yet, in any given instance, there will be many different candidates for "similar to," and even different candidates for the relevant sort(s) of similarity. So the net designer will have to set parameters.[57]

Likewise, there is a problem with Freeman's dynamics. For "to program Heideggerian AI, we would not only need a model of the brain functioning underlying coupled coping, . . . but . . . a model of our particular way of being embedded and embodied such that what we experience is significant for us in the particular way that it is."[58] We would need a model of ourselves in all our materiality, and not just our brains. And failing this, "Heideggerian AI can't get off the ground."[59]

In summary, then, Dreyfus is hopeful and hesitant about neural modelling. He is hopeful about both versions we have considered because they seem to proceed without representations and rules. But he is hesitant about the first because it requires help from the net designer, and he is hesitant about the second because it seems focused on brains and nut full persons. It is noteworthy, however, that for him there does not seem to be any in-principle block to the latter approach: it might well just be a matter of time and labor before we model the human brain and body. By contrast, the typical neural net procedure seems subject to the "insoluble problem of a

---

[55] Martin Heidegger, *Logic: The Question of Truth*, tr. Thomas Sheehan (Studies in Continental Thought: Bloomington, IN: Indiana University Press, 2010), 121, as quoted by Hubert Dreyfus in "Why Heideggerian AI Failed," 258-59.

[56] Hubert Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," 263.

[57] Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," 236.

[58] Hubert Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," 272.

[59] Hubert Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," 273.

disembodied mind responding to what is relevant."[60] In time, we will see that Lonergan can offer resources for transcending such difficulties. But for now, let us notice how much in Dreyfus he can affirm.

## II. A Lonerganian Retrieval

### *(A) Early*

To very much in Dreyfus's critique of early AI, Lonergan can utter a resounding "yea." For the most part, this is because of their similar understandings of understanding. Dreyfus's fringe consciousness, contextual disambiguation, insight, and perspicuous grouping remind one of Lonergan's patterns of experience, transcendental intention, insight, and anti-conceptualism. And Dreyfus's embodiment, situations, and needs remind one of Lonergan's anti-ocularism, history, and carnality. Let us briefly consider their affinities.

Fringe consciousness, as we saw, is a tacit awareness that human beings possess but computers do not, and by which they zero in on relevance. It is not yet the grasp of relevance, but something which makes it possible, and is in this way like Lonergan's patterns of experience. These organize and direct the flow of conscious awareness in biological, dramatic, aesthetic, or intellectual ways, and render it selective. This prepares the mind to identify specific relevance.[61]

Contextual disambiguation, of course, is the overcoming of ambiguity due to context. It permits us, but not computers, to reach beyond the confines of variable situations and get things right. In this way, it is like Lonergan's transcendental intention, which intends not this or that meaning datum, but intelligibility per se, and so supplies a criterion in terms of which to advance.[62]

Insight, for Dreyfus, is that by which we do advance, or grasp relevance, or distinguish the essential from the inessential, in a situation. It is thus the same as or similar to what Lonergan means by the same term. For him, insight is the grasp of intelligibility in the concrete, as prepared for by the patterning of experience and transcendental intention. It is the understanding that defines us as human beings and places us beyond machines, among else.[63]

Perspicuous grouping, we may recall, is that combination of fringe consciousness, contextual disambiguation, and insight by which we approach intelligibility pre-conceptually. If eventually, we do classify, and express our understanding in terms of lists of necessary traits, we do not begin there, as does a

---

[60] Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," 236.
[61] Bernard Lonergan, *Insight,* 204-212.
[62] Bernard Lonergan, *Insight,* 372-398.
[63] Bernard Lonergan, *Insight, passim.*

computer. And this marks another affinity with Lonergan, for whom understanding drives conception, and not the other way around. This is his anti-conceptualism.[64]

When it comes to embodiment, we are reminded of Lonergan's anti-ocularism. Or, we are reminded of his anti-representationalism, which is implicit in his anti-ocularism. For Dreyfus, we do not take in the presentations of sense, re-present these to ourselves, and try to mirror the world with our minds. Instead, our lived body anticipates wholes in the light of which we identify parts, and this supplies the field on which distinctions between subject and object occur. Likewise, for Lonergan, our understanding is not an affair of seeing what is out there, set out over against us, but of increasingly making good on our in-built orientation to the transcendentals, understood as essence, existence, and good. And this too is the condition of any encounter or confrontation.[65]

Again, Dreyfus's situations remind us of Lonergan's history, or commitment to historicity. If, for Dreyfus, it is in part the situated character of the human knower that permits her to know how to go on in situ, so for Lonergan it is in part her embeddedness in history that enables her to do so. For we do not, like computers, purport to operate *sub specie aeternitatis*, but inhabit time.[66]

Finally, Dreyfus's needs call to mind Lonergan's insistence on our carnality. For Dreyfus, not only must we meet the physiological demands of sight, hearing, taste, touch, and smell, but these drive us to reach out to nature, family, friends and society more generally. For Lonergan, too, the exigencies of neural demands, and the like, propel us beyond the biological to the dramatic, aesthetic, common sensical and intellectual. He is a soft, and not a hard, dualist, we might say.[67]

*(B) Recent*

To a much lesser degree, can Lonergan affirm Dreyfus's criticisms of recent, neural AI. But this is only because he would press them more strongly and add to them. In part III, section (b) below, we will see that and how this is so. Here, let us try simply to identify what Lonergan can admire.

In an early work, Lonergan writes that "With remarkable penetration Aquinas refused to take as reason the formal affair that modern logicians invent machines to

---

[64] Bernard Lonergan, *VERBUM: Word and idea in Aquinas,* volume 2 of *Collected Works of Bernard Lonergan*, ed. Frederick E. Crowe and Robert M. Doran (Toronto: University of Toronto Press, 1997).
[65] Bernard Lonergan, "Cognitional Structure," in *Collection*, volume 4 of *Collected Works of Bernard Lonergan*, ed. Frederick E. Crowe and Robert M. Doran (Toronto: University of Toronto Press, 1993), 205-221.
[66] Bernard Lonergan, "The Transition from a Classicist World-View to Historical-Mindedness," in W. J. F. Ryan and Bernard J. Tyrrell, eds., *A Second Collection* (Toronto: University of Toronto Press, 1996), 1-9.
[67] Bernard Lonergan, *Insight,* 204-212.

perform."[68] And he gives a painful example of who we can become if we do not do the same.

> A sergeant-major with his manual-at-arms by rote knows his terms, his principles, his reasons; he expounds them with ease, with promptitude, and perhaps with pleasure; but he is exactly what is not meant by a man of developed intelligence. For intellectual habit is not possession of the book but freedom from the book. It is the birth and life in us of the light and evidence by which we operate on our own. It enables us to recast definitions, to adjust principles, to throw chains of reasoning into new perspectives according to variations of circumstance and exigencies of the occasion.[69]

The passages make clear Lonergan's pity for the dependence and rigidity of early AI, but suggest a possible openness on his part to the learning and flexibility of more recent variants. If indeed this is what they possess. The difficulty, of course, is that Dreyfus is not at all sure that they do.

Recall Dreyfus's account of the problem of similarity and its would-be solution in designer parameters. "All neural net modelers," he writes, "agree that for a net to be intelligent it must be able to generalize; that is, given sufficient examples of inputs associated with one particular output, it should associate further inputs of the same type with that same output." But what, he asks, counts as the same type? "The designer of the net has in mind a specific definition of the type required for a reasonable generalization and counts it a success if the net generalizes to other instances of this type." In other words, the task of abstraction falls to the designer, not the net.[70]

A similar point is made by an exponent of Lonergan in the philosophy of law. In the law, of course, we must not only abstract in order to determine initial law, but abstract again in order to apply it. And this re-raises the problem of similarity. For an application must be legitimate, and not just arbitrary. Yet for it to be legitimate, it must regard a case which is similar to the original in relevant respects. Thus, "application of our habitual insight to any particular concrete case always involves a further insight, at least the insight that this situation is the same as the original."[71] And such an insight does not seem to be the province of computers any more than of law tables.

Again, Lonergan can affirm Dreyfus's critique of Freeman's neural dynamics, although it does not go nearly far enough. For if the latter models brain, but not full nervous function, it may well be incomplete as a model of intelligence, even if it is

---

[68] Bernard Lonergan, *Verbum*, 71.
[69] Bernard Lonergan, *Verbum*, 193-194.
[70] Hubert Dreyfus, "Making a Mind versus Modelling the Brain: Artificial Intelligence Back at a Branchpoint," in Mark Wrathall, ed., *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action* (Oxford: Oxford University Press, 2014), 229.
[71] Frederick E. Crowe, "Law and Insight," in Michael Vertin, ed., *Developing the Lonergan Legacy: Historical, Theoretical, and Existential Themes* (Toronto: University of Toronto Press, 2004), 271.

more so in virtue of its inattention to consciousness, intentionality, and acts, and the difference between a model and what it models.[72]

## III. A Lonerganian Critique

### *(A) Early*

As we have seen, Lonergan can affirm much in Dreyfus's critique of early AI, and some in his critique of more recent variants. However, not even the former would meet with his full approval. The reason, again, is to do with cognitional theory. If Dreyfus's doctrines of fringe consciousness, contextual disambiguation, insight, and perspicuous grouping resemble Lonergan's patterns of experience, transcendental intention, insight, and anti-conceptualism, and his strictures regarding embodiment, situations, and needs resemble Lonergan's regarding anti-representationalism, history, and carnality, his account of insight is by Lonergan's standards nevertheless incompletely differentiated. And this fuels in him an undue receptivity to recent, neural AI, as we will soon see.

The "insight" which Dreyfus brings to bear against early AI is a "grasp of . . . essential structure."[73] It is an exercise of "the ability to distinguish the essential from the inessential . . . necessary for learning and problem solving, yet not amenable to the mechanical search techniques which . . . operate once this distinction has been made."[74] It thus explains the fact that "[t]he grandmaster is somehow able to "see" the core of the problem immediately, whereas the expert or lesser player finds it with difficulty, or misses it completely, even though he analyzes as many alternatives and looks as many moves ahead as the grandmaster."[75] And it does not assume that "a human being, like a mechanical pattern recognizer, must classify a pattern in terms of a specific list of traits."[76] That is, it is not a species of the conceptualism against which Lonergan inveighs.

However, if insight in Dreyfus's sense is prepared for by fringe consciousness and made possible by contextual disambiguation, it is sufficient unto itself for the grasp not just of possibility but fact. And this Lonergan would contest. For he takes the act of insight to grasp a possibly relevant intelligibility, and to require verification before it can be judged truly to be so. Or, he takes one sort of insight (direct) to grasp possibly

---

[72] A model of the mind does not get us intelligence any more than a model of the weather gets us wet, Searle quips. See John Searle, Consciousness in Artificial Intelligence | John Searle | Talks at Google - YouTube.

[73] Hubert Dreyfus, *What Computers* Still *Can't Do*, 114.

[74] Hubert Dreyfus, *What Computers* Still *Can't Do*, 119.

[75] Eliot Hearst, "Psychology Across the Chessboard," *Psychology Today* (June, 1967), 32, as quoted in Dreyfus, *What Computers* Still *Can't Do*, 118.

[76] Hubert Dreyfus, *What Computers* Still *Can't Do*, 121.

relevant construal, and another (reflective) to grasp the sufficiency of the conditions for its affirmation.[77] Let us see more closely how this is so.

In response to a What is it? or How often? question, for Lonergan, we grasp unities and relations in the data of sense (or consciousness), and body forth a conception or formulation of that intelligibility in separation from the concrete. We move from so-called apprehensive to formative abstraction, and express what we have understood.[78] But we do not leave things there. For "the desire to understand, once understanding is reached, becomes the desire to understand correctly; in other words, the intention of intelligibility, once an intelligible is reached, becomes the intention of the right intelligible, of the true and, through truth, of reality."[79] And so we inquire further. Of the formulation in hand, we now ask, Is it so?, Is it true? We are not interested in bright idea but confirmed fact; we do not care for possibility but act. We identify a link between our hypothetical and what would confirm it, a tie between our conditioned proposition and its fulfilling conditions. We return to the data, to see if the conditions are fulfilled, and if they are, we affirm, we judge, with greater or lesser assurance.[80]

What is the significance of this? It is that Dreyfus is a direct, while Lonergan is a critical realist, rendering Dreyfus susceptible to over-correction in his criticisms of rationalism. Correctly seeing that intelligence is not a matter of representations and rules, but envisioning no alternative beyond pre-reflective grasp, he needlessly scorns reflection and the distance on oneself it involves. Rightly recognizing the subject not to be set over against a world out there, but envisioning no alternative to (near) self-world identity, he unhelpfully reduces the knower to the known. Or close. It would even appear, at times, that he endorses the physicalist reductionisms of recent, neural AI.

*(B) Recent*

In "Overcoming the Myth of the Mental," Dreyfus writes that "[t]he meaningful objects . . . among which we live are not a *model* of the world stored in our mind or brain; *they are the world itself*."[81] In "Depth Psychology to Breadth Psychology," he follows an approach that "do[es] not refer to the mind at all." For "the whole human being is related to the world. Indeed, even 'relation' is misleading, since it suggests the

[77] Bernard Lonergan, *Insight*, 304-340.
[78] Bernard Lonergan, *Verbum*, passim.
[79] Bernard Lonergan, "The Subject," in William F. J. Ryan and Bernard J. Tyrrell, eds., *A Second Collection* (Toronto: University of Toronto Press, 1974), 81.
[80] Bernard Lonergan, *Insight,* 296-340.
[81] Hubert Dreyfus, "Overcoming the Myth of the Mental," in Mark Wrathall, ed., *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action* (Oxford: Oxford University Press, 2014), 106, quoting himself from *What Computers* Still *Can't Do*, 265-266.

coming together of two separate entities."[82] And in "Why Heideggerian AI Failed," he says that "in our most basic way of being . . . we are not minds at all but *one with the world* . . . [T]he inner-outer distinction becomes problematic. There's no easily askable question about where the absorbed coping [practical insight] is—in me or in the world."[83]

In other texts, Dreyfus gives examples to support such claims. He cites Sartre's insistence that, in running to catch a streetcar, there is neither runner nor car, but just the situation.[84] He notes Larry Bird's report that he is unaware of what he is doing on the court until after he has done it, as well as the Israeli fighter-pilot's comment that he blacks out in situations of high performance.[85] He even claims that, in his own minimal experience of excellence in tennis, he disappears into the game.[86] It is not just that, in such events, one's awareness of oneself is tacit, and not focal. It is that the distinction between the self and world breaks down.[87] Heidegger calls this "primordial understanding." It "dispenses altogether with the need for mental states like desiring, believing, following a rule, and so on, *and thus with their intentional content*."[88] It is even "zombie-like."[89]

It is this view, then, which would seem to lead Dreyfus to endorse recent, neural AI, in spite of its apparent physicalism. For if distinctions between inner and outer, and even mind and world, break down, then so perhaps do ones between conscious intentionality and nonconscious materiality. And this is just the sort of suggestion we saw in our reviews of Dreyfus on neural net AI and Freeman's neural dynamics. In the former, apparently material transactions were likened to Merleau-Ponty's dialectic of action and milieu, and in the latter they were likened to Heidegger's hermeneutic circle.

However, it is a good question how Dreyfus arrives at his views. What is his method? It cannot be straightforward phenomenology, since it requires claims to be based in the data of consciousness, one's first-personal awareness of oneself and one's acts; and here claims to such realities are abrogated. Nor can it be straightforward science, or any third-personal approach, since it would only reveal non-conscious, meaningless transaction; and what is here being discussed is understanding. Probably Dreyfus would claim his approach is similar to that of early Heidegger and Merleau-

---

[82] Hubert Dreyfus, "Depth Psychology to Breadth Psychology," in Mark Wrathall, ed., *Skillful Coping,* 170.

[83] Hubert Dreyfus, "Why Heideggerian AI Failed," in Mark Wrathall, ed., *Skillful Coping,* 259. I owe this and the former note's quotation to Wrathall, who helpfully lists them in his editor's Introduction to this volume, 4-5.

[84] Hubert Dreyfus - Is Consciousness an Illusion? - YouTube

[85] Hubert Dreyfus, "Responses," in Mark Wrathall and Jeff Malpas, eds., *Heidegger, Coping, and Cognitive Science* (Cambridge: The MIT Press, 2000), 323.

[86]Hubert Dreyfus, "Responses," 329.

[87] Hubert Dreyfus, "Responses," 323.

[88] Hubert Dreyfus, "Husserl, Heidegger, and Modern Existentialism," in Brian Magee, ed., *The Great Philosophers* (Oxford: Oxford University Press, 1987), 258.

[89] Hubert Dreyfus, "Husserl, Heidegger, and Modern Existentialism," 266.

Ponty, who examine being-in-the-world or *etre au monde*, taking subject- and object-poles at once. But such a strategy sits uneasily between phenomenology and metaphysics, and does not lead Heidegger or Merleau-Ponty themselves to any degree of receptivity to naturalism.[90]

No, Dreyfus's method would seem simply to be bad phenomenology. As Searle points out, it cannot be true that in high performance I lose awareness of myself and my goals altogether, otherwise when things cease to go well my attention would not be drawn to the problem.[91] And as Lonergan might observe, the fighter-pilot is likely blurring the difference between tacit and focal awareness in his report. For one can hardly be aware of not being aware of oneself at all.[92] There would not seem to be any reason to suppose that in excellent action we lose awareness of ourselves and the criteria of our success. But if this is so, there is no evidence for Heidegger's "primordial understanding," in which intentional acts and their objects disappear into the world. And there is certainly no evidence for the view that we are naught but electricity acting on circuits.

Dreyfus is right to conclude his opus by saying that "Our risk is not the advent of super-intelligent computers, but of subintelligent human beings."[93] He may not, however, see all that the latter risk entails. For this reason, we should be grateful for the Thomist phenomenology of Lonergan, which offers a verifiable account of the differentiated intelligence that distinguishes us from machines.

---

[90] As Dreyfus himself recognizes. See Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," 245-247.

[91] John Searle, "The Limits of Phenomenology," in Wrathall and Malpas, eds., *Heidegger, Coping, and Cognitive Science*, 77-81. In his "Replies" to his critics assembled in this volume, Dreyfus claims to grant Searle's point (26, 384). But that he does is, I think, belied by the bulk of his replies.

[92] One could perhaps infer from one's present situation, that one just performed well under blackout. But in this one would base a large claim about oneself on material not given in consciousness, which is un-phenomenological.

[93] Hubert Dreyfus, *What Computers* Still *Can't Do*, 280.