Two Approaches to Artificial Intelligence: An Analysis of Heideggerian and Dreyfusian Critiques

Nicholas K. Gracilla
Denison University

There exist two broad research paradigms in Artificial Intelligence (AI) which differ radically in their attempts to reproduce human understanding through the use of computers. The dominant paradigm, which I call Traditional AI, has focused on formalizing the process of thinking into rules, symbols and representations of the world. As such, its roots can be found in the philosophical traditions of reductionism and rationalism. The second paradigm, which I call Parallel Distributed Processing (PDP), has focused on using computers to emulate the neurological structure of the brain. Concerned less with formalization than underlying computational structure, its approach has developed from the neurosciences, Gestalt Theory and work in perception. The short history of both paradigms is riddled with fantastic claims and unsupported predictions of success. Hurbert L. Dreyfus was one of the first to critically examine these claims—and concluded that Traditional AI was fundamentally and irreparably flawed in its approach. His criticisms, grounded in the works of Martin Heidegger, make an in–principle argument against the possibility of formalizing human intelligent behavior. Traditional AI has approached the formalization by postulating mental representations, which both Dreyfus and Heidegger reject. Dreyfus' in–principle argument, however, holds no weight against the non–formalized, non–representational paradigm which PDP uses, and he is notably less critical towards it because of this.

I believe his criticisms of Traditional AI are accurate. When addressing the AI project in its entirety, however, his argument appears to slip toward a different claim. He notes that, currently, "human beings are much more holistic [than PDP networks]," and in emphasizing this holism, he suggests the minimal unit of analysis of intelligent behavior may an entire human–like organism in the entire human culture (MVB, p. 39). Dreyfus' holism, which re–emphasizes

As a philosophy and computer science double major, Gracilla wrote this paper for his senior comprehensive project at Denison University. A native of Warren, Ohio, he plans to pursue a doctorate in philosophy at Northwestern University.

"our needs, desires and emotions" as well as the importance of our "... [having] a human-like body with appropriate physical movements, abilities and vulnerability to injury" appears to permit only human organisms to exhibit intelligent behavior (MVB, p. 39).

In this essay, I hope to show why Dreyfus' in-principle argument correctly criticizes Traditional AI, yet may not address some inherent characteristics of PDP architectures. I will also suggest that his broader, holistic argument against AI in its entirety may not be justified. My strategy will be to characterize both paradigms, emphasizing their modes of representing information. I will review Dreyfus' criticisms and their roots, as expanded in an analysis of human understanding offered by Heidegger. I will show why the critique is applicable to Traditional AI, yet inapplicable to PDP systems. His transition to a holistic claim concerning intelligence will then be evaluated, and hopefully shown to be untenable based on his own arguments leveled against Traditional AI.

The computational paradigm used by Traditional AI approaches has been described by Newell and Simon as a physical symbol system. It can be characterized by its use of abstract symbols to represent salient features in a "microworld"—an artificially constructed problem domain which simulates a subset of the real world. Syntactic rules manipulate these symbols to reflect the processes and relations which occur in the microworld. The technique is powerful: symbols and rules are capable of representing every fact and process which can occur within the constraints of the explicitly defined problem domain. It is questionable, however, whether a system which uses this approach can replicate human intelligence.

Terry Winograd's SHRDLU—a Traditional AI program which processed natural language sentences concerning a microworld of blocks, spheres and pyramids—typifies two primary problems of symbolic representation in microworlds. Within the restricted problem domain, SHRDLU could correctly respond to questions or commands such as "Can a pyramid be supported by a block?" and "Find a block which is taller than the one you are holding and put it into the box" (*WCD*, p. 7). Remarkable as this may seem, Dreyfus points out—and Winograd readily admits—that nothing even approaching an *understanding* of natural language is modeled in SHRDLU. For example, in reference to "owning," Herbert A. Simon remarks:

> ... SHRDLU's test of whether something is owned is simply whether it is tagged "owned." There is no intensional test of ownership, hence SHRDLU knows what it owns, but doesn't understand what it is to own something (cf. *WCD*, p. 13).

SHRDLU cannot understand "what it is to own something" because it is isolated from the context in which "owning" is meaningful. Moreover, SHRDLU is incapable of *understanding anything* because it is not "in" a context at all—its microworld contains only uninterpreted facts concerning geometric objects and the relationships between them. A context which provides meaning, however, does not consist of a body of uninterpreted facts and relations. "Owning" is meaningful, for example, in a context of social interactions and property rights in which one participates.

This confusion between a meaningful context and the uninterpreted facts which compose a microworld led Dreyfus to reject the idea that microworlds are "worlds" at all.

> A set of interrelated facts may constitute a *universe*, a domain, a group, etc., but it does not constitute a *world*, for a world is an organized body of objects, purposes, skills and practices in terms of which human activities have meaning or make sense (*WCD*, p. 13).

Thus, since the semantics of "owning" are context-sensitive to a human world which SHRDLU is not in, the concept of "owning" is meaningless to SHRDLU.

The second fundamental problem of Traditional AI typified by SHRDLU concerns its method of knowledge representation. SHRDLU's microworld consists of explicitly statable facts and rules—a kind of knowledge Gilbert Ryle calls "knowing that" (Ryle, p. 28). Thus, SHRDLU can express "that a sphere is in the box" since this is explicitly represented in its microworld. Yet, there is a qualitatively different kind of knowledge which SHRDLU is incapable of representing. Ryle calls this "knowing how"—a kind of knowledge which indicates a skill or capability. For example, one might know how to play Bach's preludes, how to swim or how to shoe a horse. Know-

how knowledge concerns an active *doing*, a performance, as opposed to an explicitly stated fact or rule. Such knowledge is obtained through learning from multiple experiences: one learns how to ride a bike by continually getting on the saddle and pedaling.

This mode of acquisition reveals the qualitative difference between the two knowledge types. One cannot explicitly articulate to a child "how to balance" in any meaningful or helpful way. Learning "how to balance" is not a process of studying and memorizing facts concerning one's center of gravity, the effects of motion on objects and so forth. Even after coming to know "how to balance," one cannot easily articulate such facts. Dreyfus remarks "the fact that you can't put what you have learned into words means that know–how is not accessible to you in the form of facts and rules" (*MOM*, p. 16). Since SHRDLU is designed as a physical symbol system, its microworld necessarily consists of explicitly statable facts and rules. Thus, any knowing–how one wishes SHRDLU to have access to must be converted into knowing–that—a deeply problematic undertaking.

Considering the efforts made in AI during the 1970s and beyond, it is clear that the criticisms concerning microworlds and the representation of knowing–how have become influential in Traditional AI theorizing. New proposals, such as Minsky's frame system or Schank's scripts, used complex representational structures in an attempt to address these issues. Consider Schank and Abelson's script system:

> A script is a structure that describes appropriate sequences of events in a particular context. A script is made up of slots and requirements about what can fill those slots. The structure is an interconnected whole, and what is in one slot affects what can be in another. Scripts handle stylized everyday situations. They are not subject to much change, nor do they provide the apparatus for handling totally novel situations. Thus, a script is a predetermined, stereotyped sequence of actions that defines a well–known situation (Schank, p. 41).

Scripts attempt to enrich microworlds by representing human–world interactions, and attempt to capture the kind of common–sense know–how humans use in everyday situations. Schank cites

the following short story as evidence of this: "John went to a restaurant. He asked the waitress for *coq au vin*. He paid the check and left" (Schank, p. 38). Human understanding embodies much more information than presented in the story: we understand that John ate the *coq au vin*, for example, that he sat at a table, ate the meal with utensils and so on. Human knowledge of *how one eats at a restaurant* allows us to understand this story in ways Traditional AI systems, which did not model human practices or common–sense, could not. Do these richer representational schemes offer any significant improvements?

Scripts offer a significant advantage in their ability to use *expectations*, in the form of unfilled or default data slots. One such slot, for example, could contain information about John's sitting position at the table. Thus, if the story had an additional line, such as "When the gun fired, John hit his knee on the table," the script could account for John's unfortunate reflex by already having information concerning his leg placement beneath the table.

We imagine that such a script, in order to handle the incredible amount of information encountered in a restaurant, would be quite complex. Minsky, commenting on his frame system (similar in many regards to a script) notes,

> ... the list [of facts] is not endless. It is only large, and one needs a large set of concepts to organize it. After a while one will find it getting harder to add new concepts, and the new ones will begin to seem less indispensable (*WCD*, p. 11).

Minsky's approach of decomposing the common–sense knowledge of, say, "how to use a spoon" is characteristic of AI's information processing model: the use of a spoon is a conglomeration of a huge number of actions and rules—the degree of tension the fingers must use to hold the spoon, the proper angle to hold the spoon so that food will not slide off and so on.

Moreover, Jerry Fodor, another Traditional AI theorist, questions the importance Ryle and Dreyfus place on the distinction between knowing–how and knowing–that. He remarks,

> there is a real and important distinction between knowing how to do a thing and knowing how to

> explain how to do that thing. ... But what has this to do with the relation between knowing how and knowing *that*" (Fodor, p. 71)?

In refuting one's inability to articulate knowing–how as evidence of a qualitatively different kind of knowledge, Fodor offers a distinction between mental competences or skilled abilities and mental traits—like intelligence or sensibility. Knowing–how to do something is evidence of a competency, but not necessarily a trait like intelligence. Moreover, traits like intelligence are not dependent on competencies. By drawing this distinction, he suggests that,

> if John is intelligent, there is no *specific* activity he need be good at ... being intelligent is not a matter of *doing* something ... [since] "Being intelligent" and "being stupid" do not name actions or types of actions (Fodor, p. 72).

He suggests that knowing–how appears to have the character of a qualitatively different kind of knowledge only because humans have no conscious access to it: we must, subconsciously, rapidly process large amounts of knowing–that knowledge in every action and ability. Thus, having larger amounts of information in richer representational schemes presumably addresses both problems of impoverished microworlds and the representation of commonsense knowing–how.

This technique, however, has met with serious difficulties. In the attempt to "bolster" the information a script can contain, Traditional AI theorists hope to work upwards from isolated, constrained problem domains towards the world of human knowledge and experience. Yet, at every turn, more and more information must be explicitly represented within the script. The magnitude of the project does not go unnoticed; Minsky later (1975) comments:

> Just constructing a knowledge base is a major intellectual research problem. ... We still know far too little about the contents and structure of common–sense knowledge. A "minimal" common–sense system must "know" something about cause–effect, time, purpose,

> locality, process and types of knowledge ... (Minsky, p. 124).

An obvious solution would be to construct a machine which could move around in the world and learn to create its own representations. Yet this approach has encountered a serious paradox. Richer knowledge representations require advances in robotic movement, vision and interaction to learn from the environment: yet such advances in robotics first require advances in knowledge representations in such fundamental areas such as representing the robot's own body, the solidity of objects, the effects of movement on perspective and more (*WCD*, p. 46). Dreyfus does not consider enriched representational schemes any kind of advance towards machine understanding at all. The problem lies in an unjustified belief concerning human ability in the world: why would one consider, as Minsky and Fodor do, the explicit representation of human practices to be formalizable? This makes sense only in the context of the highly constrained microworld in which a program operates, and reveals serious discrepancies between microworlds and the real world of human experiences. Indeed, in an analysis of the attempt to represent the knowledge of even a small part of the world we live in, Dreyfus concludes that microworlds are completely *unlike* the human experience of the world. He suggests that we may work in subworlds, such as the university or the theater, but they are not related to each other in an isolated mode of "composing" a larger, shared world as microworlds are. Human subworlds instead *presuppose* a larger unified whole, and work as local elaborations of it (*WCD*, p. 14).

The attempt to gain machine understanding through enriched representational schemes of the world has, so far, met with failure. Yet Dreyfus' arguments have indicated an even deeper problem with the approach: the question does not concern the degree of complexity a representational scheme must have, but rather whether human understanding involves representations *at all*. To further develop this, I turn to Heidegger's analysis of human understanding.

Dreyfus' criticisms of AI's use of microworlds, and his concern for the human context in general, can clearly be traced to Heidegger's analysis of human existence (*Dasein*) Being–in the world. "Being–in" conveys a sense of "in" entirely different from the way objects may

be "in" other objects. A sphere, for example, may be "in a box" in the sense that it is surrounded on three or four sides; but this sense of "in" is an *unengaged* one: the sphere, Heidegger says, is really "along with" the box (Heidegger, p. 79). Humans, on the other hand, are very different: we *are* engaged in the world; we dwell in a familiar and involved way in it. Heidegger notes "there is no such thing as the 'side-by-sideness' of an entity called 'Dasein' with another entity called 'world'" (Heidegger, p. 81) to emphasize that Being-in is not like an "object inside an object." Indeed, the world is not a thing at all, nor is it a composition of things. Instead, the world is a context, a background for which entities have always already been in.

Entities in the world can be encountered by *Dasein* in two ways. In use, an object is **ready-to-hand** (Heidegger, p. 98). Heidegger's examples of ready-to-hand entities typically involve skilled activities, such as hammering. The hammer, when actively used, is unnoticed: "an entity of this kind is not *grasped* thematically as an occurring Thing" (Heidegger, p. 98). Thematic grasping of an entity *qua* entity requires detached contemplation, a way of revealing objects as **present-at-hand**. Thus, a hammer could be revealed as present-at-hand—if it is sitting on a table and *Dasein* is analytically examining it, or if its head suddenly breaks when prying a nail and *Dasein* attempts to repair it. But typically entities are known in their use, as ready-to-hand.

The distinction between use and detached contemplation clearly corresponds to the distinction between knowing-how and knowing-that. Heidegger's analysis, which shows that *Dasein* is always in-the-world, sets the ready-to-hand encountering of entities as the fundamental, typical way *Dasein* understands. This understanding is knowing-how—encountering an entity as ready-to-hand in its use. Revealing an entity as present-at-hand in detached contemplation yields an entity *qua* entity. This is knowledge of the knowing-that sort, concerning facts and information about objects distinct from *Dasein*. Thus Dreyfus notes that Being-in-the-world cannot be understood solely on the model of a relationship between subject and object, because such a model does not account for understanding a thing as ready-to-hand (*BW*, p. 45). Heidegger, also refuting the subject/object model of understanding, insists that "... the perceiving of what is known is not a process of returning with one's booty to the 'cabinet' of consciousness after one has gone out and grasped

it ..." (Heidegger, p. 89). He refutes the traditional representational theory of mind, which hold that we form meaningful mental representations of the world and manipulate them when thinking. Heidegger does not deny the possibility of mental phenomena: he does, however, reject the idea that such phenomena create "internal meanings" of the world.

Traditional AI has had, as its primary focus, an analysis of the way humans–as–subjects "grasp" objects in the world and interpret them in an internal, mental sphere. This is the attempt to analyze understanding as a collection of knowing–that knowledge. But this "knowing" is only knowing the present–at–hand: it involves symbolic representations of the world and the rules needed to meaningfully manipulate them. It completely neglects understanding as primarily understanding entities as ready–to–hand. Fodor's earlier argument that intelligence is not a skill like hammering makes this fundamental mistake. This is the attempt to formalize understanding as something distinct from the way *Dasein* is in–the–world—an impossible project, since the world revealed ready–to–hand cannot be represented by a set of context–free elements. The use of a hammer, for example, is nested in the context of a human social world with purposes and roles, which need not be represented as a set of facts (MVB, p. 29).

Formalizing understanding to gain commonsense knowledge is at an impasse because Heidegger's commonsense understanding—everyday know–how—does not consist of procedural rules, but rather an unformalizable knowing–what–to–do in everyday situations (MVB, p. 33). Dreyfus suggests that a child comes to know–what–to–do by constant exposure to the world, and that "the same might well be the case for the social world. If background understanding is indeed a skill and if skills are based on whole patterns and not on rules, we would expect symbolic representations to fail to capture our commonsense understanding" (MVB, p. 33).

The Heideggerian perspective provides a useful background to Dreyfus' criticisms of Traditional AI. The danger, however, lies in the ease at which one can overemphasize the holistic nature of human understanding. That "one cannot build up the phenomenon of world out of meaningless elements" (*BW*, p. 119) does not necessarily imply that human understanding is dependent on the entire human culture, as Dreyfus does. To show this, I will first show how

the PDP approach to AI satisfies Heidegger's and Dreyfus' criticisms of Traditional AI, and then critically examine Dreyfus' "wholer than holism" criticisms.

In their most general case, PDP systems are simulations of neural networks found in the brain. They consist of large numbers of individual processing units connected together in varying degrees of complexity. Individual units typically perform simple computations; they process information by sending excitatory or inhibitory signals to other units in varying degrees of intensity, dependent entirely upon the signals of the units simultaneously connected to them. Such a network will have two primary edges of multiple connection lines: the first can be considered as an input edge, where received information can pass through the network of connections and computation units to an output edge.

A PDP network is not programmed with explicit rules nor does it create representations of the world to manipulate. Instead, a network is repeatedly exposed to "input" information concerning the world and "output" expected responses. By adjusting its internal connections, the network *learns* to associate the expected response to the situation. For example, one might "train" a network to predict weather patterns by presenting facts about the current weather conditions and what followed from them. The trained network could then associate similar future conditions to what had happened. More importantly, since there are no explicit rules concerning barometric pressures, wind patterns, etc., the network can *generalize* to new conditions based on past experience. Salient features of new experiences can be associated with past experiences, allowing for responses to conditions which the network had not seen before.

The method of knowledge retention, too, is non–explicit and non–representational. Note well that Heidegger does not deny internal psychical entities or mental states—he merely denies that they are "internal meanings" or representations. The same holds true for PDP connectionist networks. In the context of weather recognition, the value of an individual node at some position is meaningless. No individual or group of nodes "represent" a rule which might state "if the barometric pressure is high, it is likely to be a nice day," nor is the value of some individual processing unit a meaningful representation of a feature in the world.

Indeed, a trained network is only meaningful when considered

as a whole in the context of the world. Its internal adjustments, are entirely dependent on the information presented. No previously structured system akin to a script is used to deconstruct and process particular salient features of the problem. Instead, problem situations are presented to the network, which independently determines which are, and which are not, useful features.

PDP networks are consistent with Dreyfus' and Heidegger's accounts in several ways. First and foremost, they do not create internal representations in the spirit of a representational theory of mind. Activity certainly occurs between nodes, but this activity cannot be meaningfully related to external phenomena. An individual node's value is meaningful only in the context of the entire network. Secondly, the network is not independent of the context of the situation. Its only rule—which might be stated as "adjust to suit the expected response"—is entirely dependent on the information presented as well as the expected response. Any rule–like behavior which a network appears to follow can not be the result of the formation of rules, since rules cannot be represented in the network. Such behavior must be said to be emergent: a complex activity gained through the interaction of processing units which are not explicitly concerned with the more complex overall goal (Wallich, p. 128). The problem of machine representations of knowing–how appears to be readily addressed by the emergent quality of PDP systems. Just as, for humans, the acquisition of such knowledge requires repeated practice and development, parallel acquisition in connectionist systems require repeated training sessions.

Dreyfus recognizes the compatibility of PDP systems with his and Heidegger's in–principle position concerning knowing–how representations and Being–in–the–World. Yet Dreyfus is still critical of network systems. He comments "Intelligent behavior requires as a background the totality of practices which make up the human way of Being in the world ... [yet] the capability for providing such a background is, at present, beyond the horizon" (FMK, p. 132) and "... human beings are much more holistic than neural nets. Intelligence has to be motivated by purposes in the organism and goals picked up by the organism from an ongoing culture" (MVB, p. 39).

It is this increasing dependence on an argument based on more and more holism which I find incompatible with his earlier, consistent views. We say of a child, who clearly has *not* gained the "totality

of practices which make up the human way of Being in the world," that she is still intelligent despite this deficiency. On Dreyfus' account, at what point could we determine that a person was intelligent? How many human social practices would one have to know? It would be absurd to think that traveling to a country whose social practices were not known in their entirety would render a person unintelligent.

Increased holism appears to be a *quantitative* argument. Yet Dreyfus refuted Traditional AI's attempt to use a *similar* argument. At that point, the Traditional AI approach was to include more and more information into highly structured representational systems. Dreyfus had shown that the quantity of information a Traditional AI system had was irrelevant: its representations had no knowledge of the ready–to–hand. Yet now that PDP systems meet such criteria, Dreyfus reverts to the quantitative argument he refuted: a PDP system must now have access to an entire human culture, with innumerable goals and purposes.

I do not believe that PDP systems are the final answer to the many questions involved in modeling human intelligence. Yet I have shown the significant advances they do offer: networks are capable of exhibiting the non–formalizable behavior both Dreyfus and I believe are vital to human understanding. This capability renders Dreyfus' in–principle argument against them inapplicable.

WORKS CITED

[FMK] Dreyfus, H.L. "A Framework for Misrepresenting Knowledge." cf. M. Ringle, ed. *Philosophical Perspectives in Artificial Intelligence.* New York: Humanities Press, 1979.

[*BW*] Dreyfus, H.L. *Being–in–the–World: A Commentary on Heidegger's Being and Time, Division I.* Cambridge: The MIT Press, 1991.

[*MOM*] Dreyfus, H.L. and S.E. Dreyfus. *Mind Over Machine.* New York: The Free Press, 1986.

[MVB] Dreyfus, H.L. and S.E. Dreyfus. "Making a Mind Versus

Modeling the Brain: Artificial Intelligence at a Branchpoint." cf. *Daedalus* , (winter 1988): 15–43.

[*WCD*] Dreyfus, H.L. *What Computers Can't Do.* New York: Harper and Row, 1979.

Fodor, J. *Representations: Philosophical Essays on the Foundations of Cognitive Science.* Cambridge: The MIT Press, 1981.

Heidegger, M. *Being and Time.* trans. John Macquarrie & Edward Robinson. San Francisco: Harper & Row, 1962.

Minsky, M. "A Framework for Representing Knowledge." cf. J. Haugeland, ed. *Mind Design.* Cambridge: The MIT Press, 1975.

Newell, A. and H. Simon. "Computer Science as Empirical Inquiry." cf. J. Haugeland, ed. *Mind Design.* Cambridge: The MIT Press, 1976.

Ryle, G. *The Concept of Mind.* Great Britain: The Mayflower Press, 1949.

Schank, R.C. and R.P. Abelson. *Scripts, Plans, Goals, and Understanding.* Hilladale: Lawrence Erlbaum Assoc., 1977.

Wallich, P. "Silicon Babies." *Scientific American* , (Dec. 1991): 124–134.