Hubert Dreyfus: Humans Versus Computers

Philip Brey

This is a preprint version of the following article:

Brey, P. (2001). Hubert Dreyfus: Human versus Machine. In H. Achterhuis (Ed.), *American Philosophy of Technology: The Empirical Turn* (pp. 37-63). Indiana University Press.

Hubert Dreyfus, Critic of Artificial Intelligence

In 1956, a mere ten years after the invention of the first programmable digital computer, the birth of a new field of research called "artificial intelligence" was announced at a conference at Dartmouth College in New Hampshire. Artificial intelligence, or AI (as the field soon came to be known), was described as a basic science which would systematically study the phenomenon of 'intelligence.' Its researchers would pursue this goal by using computers to simulate intelligent processes, and its point of departure was the assumption that the logical operations executed by computers could be structured to imitate human thought processes. AI researchers supposed that it was possible, in principle, for computers to be supplied (through proper programming) with genuine intelligence the same way that artificially produced diamonds are nevertheless genuine diamonds. Because the workings of a computer are understood while those of the human mind are not, AI researchers hoped in this way to reach a scientific understanding of the phenomenon of 'intelligence.'

From the very beginning, AI was a field with high goals and loftypromises. The highest goal was no more and no less than to construct a computer system with the intelligence and reasoning ability of an adult human being. Many AI researchers claimed that this goal would be reached within only a few decades, thanks to the invention of the digital computer and to key breakthroughs in the fields of information theory and formal logic. In 1965, the noted AI researcher Herbert Simon predicted that, by 1985, computers would be able to execute any task that human beings could. The equally famous AI researcher Marvin Minsky predicted in 1967 that all of AI's important goals could be realized within a generation.

It is easy to understand why such predictions were taken seriously at the time, given the apparently limitless possibilities that the computer seemed to offer. In addition, a string of early successes by AI researchers helped to legitimize the ambitious claims. AI scored its first victory

already in 1956, its first formal year of existence -- a computer program able to play chess at a novice level -- and chess-playing programs improved steadily in ability almost every year thereafter. Other breakthroughs followed shortly. In 1964, a program called STUDENT was able to interpret, understand, and solve short textual passages containing algebra problems; two years later, ELIZA was able to carry out a modest therapeutic dialogue with people about their personal problems. Funding agencies took note, including the U.S. Department of Defense. Bright young researchers flocked to the new science. This inaugurated a huge growth spurt for AI, during which it established itself as an exciting, well-funded field in which hundreds of millions of dollars were spent annually world-wide, spread out among thousands of AI researchers.

As the 1960s opened, when AI was still a new field, a young philosopher named Hubert Dreyfus was introduced to it in an indirect way. Dreyfus, at the time an assistant professor at the Massachusetts Institute of Technology (MIT), was teaching courses on philosophical theories of knowledge and perception -- but his students were telling him that the theories he was teaching had been rendered obsolete by the invention of the computer. Under the leadership of his colleague Marvin Minsky in the MIT Artificial Intelligence Laboratory, they informed Dreyfus, MIT researchers were on the way to creating a machine that would be able to know and perceive on its own.

Goaded by this news, Dreyfus began discussing computers and their operations with his brother Stuart, who was then working as a computer specialist for the RAND corporation, a prominent nonprofit research organization. Via this contact, RAND recruited him as a philosophical consultant to evaluate their new AI program. This program was headed by Allen Newell and Herbert Simon, who would later become famous for their AI research. But in his evaluation of their research, Dreyfus came to the conclusion that, while it had successfully demonstrated the ability of computers to solve certain specific types of problems, it had not provided any evidence of insight into the phenomenon of intelligence itself, and indeed was on completely the wrong track in seeking to simulate human intelligence. His pessimistic report, written in 1964 and entitled "Alchemy and Artificial Intelligence," was vigorously criticized by Newell and Simon, but was released by the RAND corporation the next year over their objections.

Dreyfus's report was the first detailed critique of AI to be published, and almost immediately occupied the center stage of a heated debate by computer scientists worldwide. It was his first influential publication on the subject, and was the first of a series of philosophical critiques of AI in the form of books and articles. Dreyfus's most important publication in this area is What Computers Can't Do, which brought him international fame as a critic of AI. First published in 1972, it was substantially revised and republished in 1992 as What Computers Still Can't Do. Another important contribution to the AI discussion is Mind Over Machine (1986), co-authored with his brother Stuart.

A remarkable aspect of Dreyfus's critiques is that they are motivated by a philosophical tradition -- phenomenology -- which at the time was not often associated with science and technology and seemingly far removed in its concerns. Phenomenology, as it appears in the work of Martin Heidegger and Maurice Merleau-Ponty, applies itself to describing the interrelationships between human beings and the world, and uses the first-person experiences of human beings as a point of departure. And while Heidegger, Merleau-Ponty, and other phenomenologists have quite specific things to say about the nature of human perception, thinking, and behavior, their pronouncements about science and technology tend to be rather general and abstract. Dreyfus, however, was able to apply their ideas skillfully in his critique of AI to reach quite specific and concrete conclusions.

Ever since his earliest work on the subject, Dreyfus has progressively honed and extended his philosophical critique of AI by broadening his use of the work of phenomenologists such as Heidegger, Merleau-Ponty, and Husserl, and by making use of the insights of other philosophers including Michel Foucault and S√∂ren Kierkegaard. One of Dreyfus's principal concerns, which appears with regularity throughout his writings, is to articulate the various ways in which human beings experience the world and develop manners of getting around in it. One important book in which he takes up this theme is Being-in-the-World (1991), considered one of the best and most lucid expositions of Heidegger's early work.

Another regularly recurring concern is his critique of Cartesian rationalism. In Dreyfus's interpretation, the key rationalist assumptions are that reality has a rational structure built up out of independent elements in a rule-governed way, that human thinking works in the same rational manner, and that everything that is not rationalizable -- that cannot be expressed and defended in rational principles -- is of little if any value. Dreyfus is convinced that Western culture is still shaped to a large extent by these rationalistic assumptions, but he is also convinced, based on his readings of Heidegger, Merleau-Ponty, and Wittgenstein, that these assumptions are fundamentally flawed. Rational, formal structures are, according to him, human constructions that are only subsequently imposed on reality. Knowable reality itself lacks a rational structure; its features are co-determined by human needs and actions. The most fundamental way of knowing is intuitive rather than rational. Rationalism, as it crops up in AI and elsewhere, knows nothing of these original structures of reality and fails to do justice to the role of intuitive knowledge and skills. Dreyfus is an unrelenting advocate of intuitive knowledge and skills, and a sharp critic of rationalism in all of its contemporary guises.

Dreyfus's critique of AI has been influential to a degree that is exceptional for a philosopher. He has acquired a reputation among AI researchers -- though initially, at least, a negative one -- as well as among philosophers interested in AI. His works are studied by many nonphilosophers in computer science and other technical fields. But most remarkably of all, many influential AI-researchers have taken Dreyfus's critiques to heart, and developed and applied many of his phenomenological ideas to their own research. Via this route, the frequently

abstract philosophical notions to which Dreyfus appeals have had a direct impact on the development of AI.

This presentation of Dreyfus's work will begin with an outline of classical AI research. This will be followed by an exposition of Dreyfus's critique of classical AI, of his alternative theory of intelligence, of his critique of neural networks, an important recent approach to AI, and of his critique of the social applications of intelligent computer systems. Finally, I shall evaluate the soundness and influence of Dreyfus's critique.

The Classical Paradigm of Artificial Intelligence

From the beginning there have been many different types of AI research with different goals, methods, and formalisms. Yet also from the beginning of AI research in the 1950s up to the beginning of the 1980s the different types of AI research had so much in common as to constitute a paradigm, in the sense articulated by philosopher of science Thomas Kuhn: a collection of methods, goals, assumptions, and exemplary examples of successful research which are shared by scientists and which together define a research program. This paradigm, which continues to characterize much AI research, is known by different names, but I shall refer to it as "symbolic AI" (or sometimes as "classical AI") because its central tenet is that intelligence is symbol-manipulation.[1]

In the first decade of its existence, symbolic AI had as its goal the construction of intelligent computer systems. The grail was a system that possessed universal intelligence; that is, had a universal ability to reason, solve problems, understand language, and carry out other intelligent tasks the way an intelligent human adult could. This research was initially not directed at developing technical applications, and was promoted above all as a science: the new science of intelligence. Some classical AI researchers, including Newell and Simon, set the explicit goal of their research the modeling of the cognitive (thought) processes of human beings; AI with this goal is sometimes called "Cognitive Simulation." Within this approach, AI programs are considered to both simulate and explain intelligent human behavior.

Other researchers who do symbolic AI, including Minsky, do not pretend that their computer programs simulate human thought processes, but rather that their work provides a

[1] The assumptions of classical AI on the nature of intelligence cannot only be found within AI, but have given rise in the seventies and eighties to a new, interdisciplinary science, called cognitive science. Cognitive science is the science of both biology-based and artificial intelligent processes, and has emerged as a result of collaborations between AI researchers, psychologists, philosophers and linguists. Nowadays only part of the research in cognitive science is based on the assumptions of classical AI, and other research traditions has developed as well, like the paradigm of neural networks that will be discussed later on.

theoretical contribution to the understanding of the phenomenon of "intelligence" by laying bare the general properties of intelligent processes. They claim that, while their research may not allow direct insight into the psychological performance of intelligent tasks, it does allow insights into the performance of competent intelligent human behavior; that is, it provides general insights into the cognitive abilities which human beings must possess in order to display intelligent behavior. But the differences between the approach of Cognitive Simulation and this more usual approach are of lesser importance than their points of agreement; both approaches take as their goal the understanding of the phenomenon of intelligence, and they share important theoretical point of departure, methods, and formalisms.

Symbolic AI assumes, as its principal point of departure, that intelligence is a matter of manipulating symbols following fixed and formal rules. A series of assumptions is made to arrive at this idea. A first, necessary assumption is that all intelligent processes, including perceiving, reasoning, calculating, and language use, are forms of information processing, that is, of uptaking information from the environment, processing or manipulating of this information, and providing a response. Thus when one adds numbers, one first determines the addend, performs a certain operation on this information, and then exhibits the solution. Chess playing, somewhat more sophisticated, nevertheless has the same structure; one inventories the pieces and their positions, analyzes the situation, and determines which move is to be made. The implication of such examples seemingly is that intelligent organisms and systems have in common that they are information processing systems.

Once this assumption is made, one is naturally led to pose the following two questions: what is the character of this information, and how is it "processed"? At this juncture symbolic AI makes two key assumptions. In response to the first question, it assumes that, to be processed by an information processing system, information must first be represented. In order to handle information, a system must first operate with a medium in which the information can be entered. Such a medium, which provides information about an external reality, is called a representation. Familiar examples of representations include photographs, pictures, images, and spoken and written sentences -- but these are not the kinds of representations that can be used by information processing systems. Information processing systems are assumed to use internal representations, all given in a form adapted to what they can handle. Thus human thinking is supposed to work via a system of internal mental representations in which our thoughts, perceptions, and memories are all inscribed.

The most fundamental assumption of classical AI is that the internal representations of intelligent systems are symbolic in nature. An alternative possibility is that internal representations are more like photographs and images in that they are iconic, carrying information by a physical resemblence with that to which they refer the way a portrait conveys information about its subject by virtue of its likeness. But classical AI takes as its starting point the assumption that the internal representations are more like the words of a natural language.

Language is symbolic; its tokens are arbitrary in the sense that they have neither resemblances nor inherent references to that which they represent. The word 'dog' neither looks like a dog nor has any intrinsic connection to one.

The assumption that whatever bears the information in intelligent systems is symbolic is more convincing than that it is iconic. It is difficult, for instance, to imagine an iconic representation of abstract things with no perceivable structure. Symbolic representations are also much easier than iconic ones to combine, and a finite number of symbols can be used in different combinations to represent an infinite amount of content. Natural language is a case in point; a finite number of words can be combined to create, in principle, an infinite number of sentences. Also, symbols clearly play a prominent role in the sorts of cognitive tasks, like mathematical calculations and logical reasoning, that are often taken to represent the highest form of intelligence.

From the earlier assumption that intelligence consists of the capacity to process information, together with the assumption that information processing consists of symbol manipulation, it follows that intelligent systems are symbol processing systems. Up to now nothing has been said about how these symbols are processed. It is a further assumption of symbolic AI that symbols are only processed on the basis of their formal properties, that is, the form of the symbol as opposed to its content or meaning. The meaning of a symbol, therefore, does not play a direct role in its processing. Thus when a computer processes the symbol 'dog' in a particular way, it does so because of what it recognizes in the form of the symbol and not because it has any insight into the meaning of the symbol.

But how is it determined which processes are carried out by the system on the basis of these formal properties? Here we encounter another key assumption: an information processing system includes rules according to which these symbols are interpreted and processed. These are of necessity formal rules, because they are only linked to the formal properties of these symbols. They work automatically, as it were; when presented with a symbol or series of symbols, the system executes a certain process which results in a new symbol or set of symbols, which then once again is linked automatically to another rule, and so forth. In the absence of such rules, intelligence would be a mystery, at least, in the perspective of symbolic AI, because there would be no easy way of accounting for the symbol-processing ability of intelligent systems.

More support for the assumption that intelligent information processing consists of the application of rules appears to come from the role which rules appear to play in many tasks requiring intelligence. Thus language comprehension appears to involve knowledge of rules of grammar, logical reasoning appears to involve the application of analytical rules, and problem-solving in mathematics and natural science appears to involve the application of mathematical principles or natural laws. Notice the assumption that the knowledge required for intelligent

behavior is theoretical; to know and understand something means to possess an abstract, symbolic theory expressed in rules with which one apprehends the phenomenon.

The theory of intelligence just sketched out can be formulated and elaborated without reference to the nature and possibility of the digital computer. Clearly, however, the development of the computer makes this theory of intelligence considerably more attractive. A digital computer is conceived to be an information processing system which makes use of symbolic representations (strings of zeros and ones) and which processes these symbols according to formal rules (a likewise symbolically represented program). The computer's existence therefore seems to offer the opportunity to test and explore assumptions about the symbolic and rule-governed character of intelligence and build models of intelligent processes in a scientific manner. The above-mentioned assumptions about intelligence thus offer the prospect of a potentially fruitful scientific research program that might well lead to interesting technical applications. The initial successes of symbolic AI in the design of intelligent computer programs seems to supply an additional legitimacy.

The ideas about the nature of intelligence promoted by symbolic AI are sometimes described as innovative, but Dreyfus emphasizes that it is merely the latest reincarnation of an ancient view, generally called rationalism, which periodically emerges in the history of thought. Plato laid the groundwork for this theory. Plato proposed that wisdom consisted of the ability to formulate knowledge in explicit definitions, and scorned human beings whose behavior was based on learned abilities or intuition. He believed in the possibility of discovering a system of theoretical, objective principles which, much like the fundamental axioms of geometry, could be used to justify behavior and explain reality on a rational basis.

The most important representative of this idea in the modern era is Rene Descartes, who in the seventeenth century claimed that each problem can be broken down into simple and independent elements, and that every complex situation or thought can be clarified by discovering the system of rules which govern how this situation or thought has been built up out of this simple elements. He claimed that even the human mind operates according to such rules and simple elements. After Descartes, this conception of the nature of intelligence crops up again in other rationalistically minded thinkers including Leibniz, Kant, and Husserl, but to a lesser extent also in empiricists such as Locke and Hume, and more recently in scientists such as linguist Noam Chomsky, the psychologist Jerry Fodor, and the different representatives of symbolic AI.

According to Dreyfus, three characteristic notions of rationalism can be found in symbolic AI. The first he calls the psychological assumption, the already-mentioned supposition that human intelligence is a question of symbol-manipulation according to formal rules. This assumption provides the theoretical basis for the claim that a computer can be programmed to think like a human being. Not all AI research programs share the psychological assumption, though most do; it is the explicit point of departure, for instance, of the Cognitive Simulation

approach. All variations of symbolic AI, however, share a second, epistemological assumption, that all knowledge is formalizable; that everything which can be understood by human beings can be expressed in context-independent, formal rules or definitions. If true, this supposition would guarantee the success of the project of symbolic AI even in the event that the psychological assumption is false, because a formalized version of informal human knowledge and behavior would have the same cognitive value as the unformalized original. While a computer might not be in a position to simulate human thought-processes, it would be able to reproduce human intelligent behavior.

Both the epistemological and the psychological assumptions are frequently based on the ontological assumption that reality, insofar as it can be known by human beings, has a formalizable structure which is built up out of a series of objective, determinate elements, each of which is independent of the others. If reality lacked such a structure, it would be highly unlikely that it would be knowable with the aid of a set of context-independent, formal rules whose existence is taken for granted in both the epistemological and the psychological assumption.

The fact that classical AI is based on a rationalistic conception of intelligence which belongs to the history of philosophy would not be of interest except for the fact that that history also includes important philosophical critiques of that conception. Dreyfus has been heavily influenced by these antirationalist critiques, especially those of Heidegger, Merleau-Ponty, and Wittgenstein, and he makes extensive use of their arguments in his critique of symbolic AI and its three central assumptions, as well as in his development of an alternative theory of intelligence.

Dreyfus's Critique of Symbolic AI

The two scientific fields which could furnish proof in support of the psychological assumption of classical AI are psychology and neuroscience. Neuroscience is important because thinking with the aid of rules and symbols would be possible only if these rules and symbols are implemented in the human brain in the same way that a computer program is implemented by the hardware of a computer. Dreyfus calls the assumption that brain functioning involves the implementation of a symbol processing system the biological assumption, a fourth assumption which frequently appears in classical AI. But Dreyfus emphasizes that both psychological and brain research have not so far provided good empirical evidence for the psychological and biological assumptions, though neither is their any conclusive proof for the incorrectness of these assumptions.

Dreyfus's most important criticism, however, is directed against the epistemological assumption, underlying all forms of classical AI, that intelligent behavior can be reproduced by formalizing human knowledge (i.e., codifying it in rules) in a way that can be followed by a machine. Dreyfus argues against this assumption that, while such formal rules may be one way

of describing human knowledge, they cannot provide the basis for reproducing such knowledge. The possession of knowledge, Dreyfus points out, entails the ability to apply this knowledge in relevant situations by applying it in reasoning, communication and behavior. The idea that fire is hot, for instance, entails being able to apply this knowledge at appropriate moments in thinking about or dealing with fire; if this didn't take place one could not truly say that this knowledge were present.

The application of formalized, rule-given knowledge, however, appears to run up against an important problem. If a computer which has been given a set of formal rules is to be able to apply them to a new datum -- for example, to a typed-in sentence, an arithmetic sum, or a photographic image -- it must first be told under what precise circumstances they are indeed applicable. This may seem to be simply an issue of symbol matching: if the datum (encoded in symbolical form) has the appropriate formal properties referred to in a rule in the computer program, then apply the rule. Almost invariably, however, it turns out that contextual factors also play a role in rule application. Consider, for instance, the attempt to program a computer to understand language. The simplest approach would be to give the computer a set of rules of interpretation which define the meanings of different words and a set of rules of grammar to analyze the sense of the sentences. The computer would then recover the meaning of strings of text by applying these two sets of rules.

But numerous problems arise here, one being the fact that many words are ambiguous. Suppose for instance that the computer is given the word "hot" in a sentence and is asked to describe its meaning. One rule that might apply in this case has the form, "If something is hot, it has a high temperature" -- but a second possible rule has the form, "If something is hot, it has a sharp peppey taste and will burn the mouth." In order to know which of these two rules of interpretation is to be applied, other elements are relevant, for instance whether the text refers to food. Higher-order rules must therefore be formulated for the correct application of the rules of interpretation; "If in the preceding text reference is made to a peppery dish, then apply the second interpretive rule." But there are also exceptions to these rules of application. A text can be about hot Mexican dishes, but nevertheless use the word "hot" to refer to something with a high temperature. Thus there need to be still higher-order rules of application for the appropriate use of the rules of application -- threatening to give rise to an infinite regress of rules, making interpretation impossible.

In short, the problem appears to be that the correct interpretation of many elements depends heavily on surrounding elements. Formal rules, however, need to be tied as little as possible to the context and to be related only to the elements in question, or to a few which are important in determining its application. If one sought to make rules sensitive to context, all possible contexts would have to be formulated, or separate rules of application would have to be formulated. Both solutions appear to be without an end. Human beings, Dreyfus observes, are able to interpret elements effortlessly from the context. Thus if they encounter a misspelled

word in a text, they automatically fill in the right meaning while computers grind to a halt. Human beings, Dreyfus concludes, have "common sense," by virtue of which they know which interpretations are meaningful and which not. Computers lack common sense, which is why they often reach absurd interpretations. Dreyfus calls providing computers with common sense the greatest challenge of classical AI, and calls it the commonsense understanding problem of classical AI. But it is a problem Dreyfus considers insoluble, for the reasons mentioned above.

Computers function the best when the "world" which they encounter and interpret is an artificial and formal world. A formal world consists of elements whose identity can be directly read off from their form, independently of whatever other elements there are in the world, and which are linked to these other elements in clearly ordered ways. In such a formalized reality the common-sense knowledge problem rarely crops up. This is the case, for instance, in simple games like tic-tac-toe and to an extent even complex ones like chess, and also for mathematics and formal logic. The problems encountered in such knowledge domains are nearly always characterized by a clear goal or "final state" to be reached -- three 'x's' in a row, checkmate, or a the numerical value of a mathematical equation -- as well as by a set of clearly defined steps to be taken in order to reach the goal. And in fact classical AI has had its greatest successes in solving just these kinds of problems.

But more mundane problem situations involving intelligent behavior appear to have an entirely different structure than that of these artificial 'worlds." Let's take for instance the problem situation that occurs when you accidentally lock your car keys inside the car after a trip to the supermarket. Clearly this situation involves a problem which requires a solution, though the problem differs from those mentioned in the previous paragraph on two counts.

The first is that the goal of this solution may not be clear in advance. Clearly, you are posed with a problem that you will have to overcome, but what is your goal? Is it to regain access to the car keys in the car? Not if one has access to a set of spare keys which a friend can readily bring over. Is it to be able to drive the car again? Not necessarily if your greatest need is to return home as rapidly as possible. The search for a solution is not directed toward a single, unique goal, but involves a continuous weighing of different needs, including the estimated damage to the automobile, the cost of lost time, returning home or making appointments promptly, and so on.

The second difference is that, even when the goal is clear, the problem is not readily formalizable because it is not apparent in advance which facts are potentially relevant to finding a solution. The situation, that is to say, is not characterized by a fixed set of elements with objective properties and characteristics to which rules can be applied. Some facts become potentially relevant only during the actual solving of the problem, such as the car window that turns out to be ajar or weakly fastened, or a wire hanger lying on the street which might be used to break in, or a previous owner who just might happen to have kept a set of reserve keys. The process of solving this kind of problem, that is, typically progresses through several stages in the

course of which one conceptualizes and reconceptualizes it in search of a representation of the problem that gives one the best feeling that one has a confident grasp of the situation. The ability to creatively reformulate a problem appears to be a more essential skill than the ability to find a solution to a problem whose definition is clear and well-defined from the outset. Formal rules and their application appear to play no role in the search for a good definition of the problem.

In short, there appear to be sound arguments against the epistemological assumption that intelligent behavior can be reproduced by a system of consisting of formal rules and symbols. Moreover, human intelligence itself does not seem to work in this way.

Intelligence is Embodied and Situated

In his alternative theory of intelligence Dreyfus argues that one must begin by recognizing that human beings generally do not apply rules in their intelligent behavior -- and generally do not even make use of internal representations. For Dreyfus, intelligence is situated; it is co-determined by the situations in which human beings find themselves. The insights on which intelligent behavior are based are constructed locally, from concrete situations, with the aid of information which is a direct product only of this situation and without the aid or necessity of prior rules or internal representations. This vision, which derives mainly from the philosophy of Heidegger -- and to a lesser degree from that of Wittgenstein and Merleau-Ponty -- is probably the most difficult part of Dreyfus's work to understand.

That the psychological assumption that human beings need representations and rules in order to interpret the world seems so reasonable stems, according to Dreyfus, from a particular conception of how the world is and how it is known by human beings. The world, in this view, is interpreted as a material structure, independent of human beings. This is just the view offered to us by the natural sciences. This world is inherently meaningless, and is spatially separate from human beings, so that no direct, unmediated experience of it is possible. Dreyfus does not deny the value of the perspective offered by the natural sciences about the world, but emphasizes that another perspective is possible -- the phenomenological perspective. In its description of the world, this perspective takes human experience as its point of departure. The 'world,' as the word is used by phenomenologists, thus refers to the world as it is manifested in human experience.

This 'human' world is a world that is not entirely objective, as it is filled with experienced structures, like smells, feelings, frustrations, threats,obstacles, and goals. Nor is it completely subjective in the sense that the structures which we learn to perceive in the world are not our own arbitrary mental constructions; smells and obstacles are not things which we invent but which are manifested in our encounters with the world. Human beings are born into, interact with, and learn to perceive, behave, and think in a world that is neither entirely objective nor entirely subjective. This world evolves alongside these activities, for new worldly structures are always manifesting themselves in and through human activity. Thus while the world of a

newborn baby is to a great extent unstructured, that of an adult human beings contains countless structures that have crystallized out in the course of years.

Dreyfus emphasizes that, for human beings, the experience of the world as a whole precedes the experience of independently distinguished elements. Thus a depressed person experiences the world as "gray" and "meaningless" before specific elements stand out in it, and experiences a new environment as 'safe' or 'threatening' before distinguishing discrete objects; it is the situation as a whole that calls for the experience.

Specific elements in a world or situation are distinguished and experienced from out of this more general experience of meaning and sense. As a result, these elements stand always in a meaningful relation of significance with their context. While at work, a carpenter experiences a hammer that lies close at hand as a 'thing-with-which-to-hammer' and as 'thing-that-is-useful-with-nails,' but in a more threatening context might experience it as a 'weapon-to-use-against-an-intruder.' In neither case does the carpenter perceive the hammer as just one item in the environment among all the rest, whose meaning and significance are still to be determined. In the same manner a chess grandmaster 'sees' a meaningful board situation and its associated possibilities, without having to first build it up by inventorying the locations of specific pieces and surveying their possibilities as allowed by the rules.

Intelligent behavior involves human beings discovering meaningful structures in situations in which they find themselves, which call in turn for meaningful behavior. The meaningful structure which human beings find in such situations is a local product of their needs, actions, and perceptions. The different elements in this structure derive their meaning from it. Actions flow automatically from the meaningful context; just as the eye automatically 'understands' the amount of light it receives and reacts by increasing or shrinking the size of the pupil, so human beings 'understand' the situations in which they find themselves and react accordingly with actions appropriate to the context. These actions can be generated fairly automatically out of the experienced situation because these situations are already structured in 'manageable' ways, that is, with an eye to meaningful behavior.

The meaningful structure which is experienced in a situation is thus not one that has been built up according to the application of a number of fixed rules out of separate, context-independent elements. If it were, the assumptions of symbolic AI would be sound. But the reality is exactly the opposite: the global, holistic structure which belongs to each situation makes it possible, by a process of abstraction, to discover and represent elements in it as separate objects and facts, and then to apply rules to them. For intelligent behavior it is usually not necessary to abstract in this way, except when the problem situation is defined in abstract ways from the outset.

Dreyfus's views about the situatedness of human behavior form one major part of his theory about human intelligence; the other major part consists of his view that intelligence is embodied; that is, requires a human body (Dreyfus 1967, 1972, 1996). This view, which is not

entirely independent of the first, derives mainly from the philosophy of Merleau-Ponty. However, Dreyfus's explication of this view is, much like his explication of his ideas about the situatedness of intelligence, often unclear and schematic. It is unclear, for instance, whether Dreyfus means that intelligence is something that is distributed throughout the entire body and thus cannot be spoken of as localised in the brain or the mind, or whether he means that intelligence can exist without a body, but can only be developed with with the aid of a body. Let's consider the plausibility of each in turn:

1. *Does intelligence require a body?*

Scientists as well as nonscientists often assume that intelligence is localized in the brain. In the case of at least one important type of intelligence, sensorimotor intelligence, however, this assumption is clearly disputable. Sensorimotor intelligence is the skill which human beings use in perceiving, recognizing, moving, and manipulating objects, as well as in coordinating and integrating perception and movement. The development of sensorimotor intelligence clearly requires a body, but this of itself does not mean that sensorimotor intelligence is also localized in the body; it is in principle possible that sensorimotor intelligence is exclusively a product of the brain in response to stimuli provided by the senses and carried out by the musculature. An alternate and equally defensible hypothesis, however, is that sensorimotor intelligence is localized in a complex feedback-system that comprises the nervous system, the senses, the glands, and the muscles. All these elements could be then be analyzed information processing systems, or parts thereof. Sensorimotor intelligence would then be a property of a fully developed body, in which not only brains but also other organs pass through a training process leading to the development of a total system able to carry out intelligent and fully coordinated perceptions and movements.

But even if this hypothesis is correct, it is unlikely that all human intelligence is distributed in the body. Especially abstract, 'higher' forms of intelligence, like abstract reasoning and calculation, do not appear to be dependent on a body. Human beings can have limbs and organs amputated or paralyzed and still not lose their ability to engage in abstract thought and it is at least a theoretical possibility that, as sometimes depicted in science fiction stories, a brain could be removed from its body and kept in laboratory conditions while still retaining the ability to think. Not all types of intelligence thus appear to require a body. So if this is what Dreyfus means when he says that intelligence is embodied, his position is implausible when it is supposed to apply to 'higher' forms of intelligence.

2. *Can intelligence only develop with the aid of a body?*

Even if a body is not required for the *possession* of intelligence, it could still be required for the *development* of intelligence. It is obvious that a body is required for the development of sensorimotor intelligence, but for more abstract forms of intelligence this assumption is less plausible. An alternative hypothesis, compatible with the psychological assumption of symbolic

AI, is that abstract intelligence is based on an innate symbol system in the brain, that in principle can develop independently of the body just as a computer does not require a body in order to extend its knowledge capabilities.

However, the first abstract thought processes which children develop appear to be closely integrated with their sensorimotor intelligence. Thus their first use of language is strongly tied to the world in which they perceive and behave, and their first use of numbers is related to concrete objects. Their imaginative abilities are also strongly associated with this sensorimotor world. An alternative hypothesis for the development of abstract intelligence is hence that it is not based on fundamentally new abilities but rather based on abilities which are already involved in the development of sensorimotor intelligence.

Sensorimotor intelligence includes abilities such as pattern recognition, the mental grouping and manual manipulation of objects, the assessment of the impacts of forces on things, the visual taking apart and transformation of spatial structures and the mental anticipation of the effects of actions. A developing abstract intelligence might be directly built up out of such abilities through their application to abstract domains. Thus even the manipulation of abstract symbols, as in mathematics and formal logic, ultimately would lead back to our ability to manipulate material objects in space and time. This is the view Dreyfus appears to lean towards, and he refers in his recent work to the studies of Mark Johnson (1987), who has tried to demonstrate that abstract concepts and abstract logic ultimately can be reduced to concrete, sensorimotor structures.

If intelligence is indeed situated and embodied, then it does not appear possible for digital computers to possess the broad scope of human intelligence, for they are not embodied and do not have a full human world at their disposal. The intelligence of computers appears to be limited to the performance of tasks in well-defined, formal domains and will fail in a complex human world.

The new paradigm of neural networks

Not only have the shortcomings of symbolic AI become ever more apparent in recent years, but a rival AI paradigm has also arisen, called *neural networks* or *connectionism*. Neural network AI, which began to be developed in the beginning of the 1980s, is viewed by most researchers as a radical alternative to symbolic AI, rejecting from the start the idea that intelligent behavior springs from the manipulation of symbols according to formal rules. The neural network approach derives its inspiration for the modeling of intelligent processes not from the digital computer, but from the structure and operation of the human brain. What his approach still has in common with symbolic AI is that intelligence is regarded as something that consists of information processing.

The structure and operations of neural networks are built to resemble those of the human nervous system, specifically the brain. The nervous system is built up out of nerve cells

(neurons). Neurons can be conceived as tiny information processing systems: they receive stimuli from other nerve cells or sometimes directly from sense organs, and react by delivering electrochemical stimuli to other nerve cells, or sometimes to muscles and glands, where they are taken up by receptors. Whether a neuron delivers such stimuli and how strongly depends on a physiologically determined "program" in the neuron which responds to the way the impulses it receives reinforce or interfere with each other. When these impulses are above a certain threshhold the nerve cell reacts, and it delivers an impulse in turn to its surroundings. Neurons can therefore be conceived as processors with relatively simple input/output functions.

Researchers think that the difference between human nervous systems with that differ in their intelligent capabilities is chiefly determined by the way the neurons in them are connected with each other and with the rest of the body. Intelligence is therefore mainly a product of the connections which the neurons enter into -- hence the name "connectionism." Neurons develop through entering into or breaking off, or strengthening or weakening, connections with their surroundings, depending on the way they are stimulated. At birth the connections which the nerve cells enter into are to some extent arbitrary, but as the infant interacts with the surroundings the nerve cells adapt in such a way that the behavior they instigate becomes progressively more intelligent and successful. To say that a nervous system learns therefore means that the connections between nerve cells become modified by experience.

Neural network AI tries to create artificial intelligence by trying to simulate the operation of the nervous system, by constructing a system of simple information processors with input/output functions resembling those of nerve cells. The number of processors can range from a few dozen into the thousands, and the strength or 'weight' of their connections changes depending on the stimuli they receive. They typically consist of an input-layer through which information is entered, one or more intermediate layers, and an output-layer. In practice, thus far at least, neural networks are not true physical constructions but are simulated in ordinary digital computers. In recent years, however, these computers have become extraordinarily powerful, consisting sometimes of tens of thousands of parallel processing computers, and research is underway on parallel computers based on optical fibres.

Existing neural networks turn out to be astoundingly good at carrying out certain intelligent tasks, as pattern recognition, categorization, and the coordination of behavior. Neural networks, for instance, have been built which are able to recognize human faces from different angles, and which can vocalize words on the basis of a written text. Neural networks give their best performance with tasks which require "lower" forms of intelligence, such as pattern recognition and categorization of perceptual stimuli. However, thus far neural networks have been unable to tackle tasks requiring the application of higher intelligence, such as mathematical

or logical problems -- which are precisely the problems with which symbolic AI has scored its best successes.[2]

Dreyfus has asserted that the basic assumptions of neural network AI are compatible with his own vision of intelligence (Dreyfus & Dreyfus 1988; Dreyfus 1992). Neural networks relinquish the rationalistic idea that intelligence is a matter of symbol manipulation and rule application. Knowledge in neural networks is not a matter of possessing explicit representations, but rather of the appropriate connections (ultimately) between nerves and muscles. Knowledge involves possession of an ability: it is more knowing how to do something than knowing that an assertion is true. In neural network AI, intelligent processes are frequently holistic and intuitive. Moreover, neural network AI is fully compatible with the assumption that intelligence requires a body and is situated: higher processes are often built up out of lower ones and intelligence is conceived as something that develops through interaction with the environment. Thus by Dreyfus's own criteria neural network AI appears to have more of what it takes to manufacture artificial intelligence.

But Dreyfus is ultimately pessimistic about the possibility that neural networks will ever realize this lofty aim. The problem lies not in the basic assumptions of neural networks, but in the incredible complexity of human intelligence. The fundamental problem with neural network AI is that, once again, the problem of "common sense" crops up, though in a somewhat different form. The intelligence of neural networks to a large extent depends on experience, for it is based on the connections which have been cultivated in order for the network to deal effectively with those situations that it happened to encounter in the past. The ability to deal intelligently with new situations depends on the ability to generalize intelligently from these past experiences to new ones. But here it seems as though intelligent criteria are first required in order to make intelligent generalizations: which past knowledge is relevant to the new situation, and which adaptations to it are needed in order to apply the knowledge to the new situation? If in the past I have only eaten applies inside, does this allow me to conclude that applies are also edible outside?

In principle, all the knowledge that a person possesses can be relevant in generalizing to new situations, and one cannot determine in advance what is and is not relevant. One must therefore have at one's disposal all the knowledge stored in one's brains in order to be able to generalize successfully. The same would be true of the generalizing ability of neural networks. But this suggests that a network that is able to generalize as successfully as a human brain would have to consist, not of dozens or hundreds, but of millions of processors.

Aside from the fact that such networks are currently impractical, there is still the question of how such a network could possibly acquire all the relevant knowledge possessed by

[2] For an introduction to neural networks see Anderson (1995).

a normal human being. To acquire this knowledge would seem to require that the network pass through the same learning trajectory that an adult human being has -- but this would require that the neural network be embodied. It therefore appears that a neural network which could generalize as intelligently as a human being could only happen if it were built with the complexity of a human brain, supplied with an artificial body resembling a human one that pass through a developmental trajectory similar to that of human beings when they mature. But so far the creation of such an android form of life belongs to science fiction.

From Medical Specialist to Teacher: Intelligent Computer Systems in Society

Although the promise of eventual practical applications may have played a role in the early enthusiasm of AI researchers and their supporters, few such applications had turned up by the end of the 1970s. But during the 1980s, things began to change and AI began to take on the character of a technology, as opposed to a theoretical science. The ultimate goal of most recent AI research is concerned with interesting technological applications. AI research has largely lost the ambition, as it had in the Cognitive Simulation approach, to make a fundamental contribution to the scientific explanation of intelligence. Indeed, many AI researchers have stopped referring to themselves as scientists and now call themselves knowledge engineers.

AI technology has become a multibillion dollar industry and ever since the late 1970s has delivered a stream of interesting products such as chess computers and expert systems. And ever since the 1990s there has been an upsurge in conventional devices that have been equipped with artificial intelligence, such as 'intelligent' vacuum cleaners, washing machines, and video cameras; control systems in industry, and 'intelligent' computer software like the more sophisticated internet search engines and operating systems that adapt their behavior to the user's habits. These kinds of applications mean that the boundary between AI research and other technological research -- especially in computer science and electrical engineering -- is rapidly disappearing.

The widespread use of intelligent computer systems in society has brought about new philosophical -- and especially ethical -- issues. Intelligent computer systems make choices and decisions according to criteria of which the users generally have little or no understanding. In effect, the computer systems take over responsibility for such choices and decisions from human beings. Handing over some decisions to computers -- such as those involved in intelligent video cameras or in chess programs -- are ethically unproblematic. But ethical issues are clearly raised when decisions are handed over to computer systems about issues like criminal punishment or whether to admit someone seeking political asylum.

The most important ethical problems related to intelligent computer systems are associated with expert systems. Expert systems, the first of which were developed in the middle of the 1970s, are computer systems which are intented to take over tasks from experts in a particular specialized domain, and examples have been developed in medicine, law, industry,

mathematics, science, financial planning and accounting. Thus expert systems have been created to diagnose illnesses and recommend treatments, to track down flaws in airplane engines, to identify geological sites where valuable minerals might be mined, to put together investment portfolios, to establish whether a person deserves unemployment compensation, and to determine punishments for convicted lawbreakers.

Expert systems are mainly built according to the assumptions of symbolic AI. Their designers try to provide these systems with the required knowledge by interviewing experts and seeking to make explicit their often unverbalized and intuitive knowledge. This results in a list of often thousands or ten thousands of facts and heuristics (rules that experts are thought to follow in reasoning) which are then translated into a computer program. The performance of the system is then compared with the performance of a human expert. If the system appears to perform satisfactorily, it can be put to use.

Despite his criticism of symbolic AI, Dreyfus was relatively optimistic in his early work about the prospects of expert systems. He had always claimed that computers could perform well in formalized domains which required little common sense. The sort of knowledge that experts like chess grandmasters and scientists acquire appears to be formalizable in rules, and appears to call for little common sense or everyday knowledge. Dreyfus claimed that computers in these specialized domains of knowledge might well be able to log striking successes.

But Dreyfus later reconsidered. What catalyzed this change of heart was a study he conducted together with his brother Stuart of the manner in which human expertise develops in a particular area (Dreyfus & Dreyfus 1986). This study seemed to show that humans employ rules in early stages of learning, but in later stages replace this with an intuitive and holistic manner of problem solving. A chess grandmaster, for instance, does not apply any rules to chess, as do beginners, but "sees" in a single glance the situation on the board, potential moves, and potential replies. The expertise consists not in a warehouse of facts and rules, but in the recollection of past situations which were successfully confronted. The simple rules which are taught beginners -- such as "first the knights, then the bishops," or "a rook is usually worth more than a bishop" -- are used as rules of thumb in the global context of thousands of perceived situations, plans, moves and countermoves.

Rules of thumb are an important learning tool for the novice and advanced beginner in a particular knowledge domain, providing a simplified vision of its structure and a handle for addressing specific situations. Because reality lacks a formal structure which can be grasped in rules (contra the ontological assumption), expertise ultimately consists in a knowledge of and ability to deal with countless separate situations. Expert systems, which are based on the assumption that the knowledge of experts can be formalized, can never reach this level of sheer expertise.

This gives rise to a number of limitations in the range of application of (symbolic) expert systems. Because expert systems cannot make decisions or form judgments at the level of an

expert, they cannot be entrusted with tasks that require expertise. However, Dreyfus is convinced that expert systems can often attain a certain degree of competence, being a performance level that surpasses that of a novice or advanced beginner and is comparable with that of an advanced student. Expert systems therefore might indeed prove useful in applications that do not call for performance at the expert level.

One question which Dreyfus does not broach, however, is that of deciding whether a particular task calls for expertise or only competence. The determination of the right punishment for a crime clearly calls for the expertise of a judge, who takes into account the circumstances of the crime and background of the lawbreaker in establishing the punishment. A legislative body, however, may decide that judges (or juries) henceforth must decide the punishment based on a certain list of formal principles, such as the type of crime, the criminal's record, and a set of other verifiable data. This would eliminate the role of the intuitive judgment of the judge and transform the judge's function into the application of a number of formal rules -- a task which could well be taken over by a competent expert system.

Whether the use of expert systems in particular domains can be justified thus largely depends on the legitimacy of the decision to formalize these domains and on the decision to exclude the role of intuitive judgment. Already in 1976, the AI researcher and critic J. Weizenbaum wrote an influential "critique of instrumental reason", which attacked the tendency to reduce human problems to calculable, logical problems. This phenomenon, of course, predated computers, but their use greatly stimulated the desire to make such attempts. Weizenbaum's conclusion, to which Dreyfus would clearly subscribe, is that the intuitive judgment of human beings is indispensable even in specialized domains.[3]

Besides expert systems, a second type of intelligent computer system that Dreyfus discusses (and which is closely allied with expert systems), consists of intelligent tutoring systems, or ITSes, which are employed in computer aided instruction. Intelligent teaching systems are computer programs which take over certain teaching roles. For the most part they are not intended actually to replace the teacher, but rather to supplement the instruction. An important distinction must be drawn between the use of a computer as an ITS and its use in other functions such as word processing, electronic whiteboards, or databases which use "unintelligent" computer programs. An ITS, by contrast, is a program that pretends to be intelligent, for it pretends to possess some of the abilities of a professional teacher.

Intelligent tutoring systems can help students in two ways. In its most simple form, an intelligent tutoring system can supply problems to which the student must find the correct answer; exercises in spelling or algebra, for instance. This kind of ITS has the ability to generate

[3] A more recent and exhaustive critique of expert systems is found in Collins (1990). For ethical discussions about expert systems see Forester & Morrison (1994, ch. 7).

new questions or problems and to evaluate the answers given by the students -- and through the repeated production of examples and exercises to help a student obtain knowledge and ability in a particular domain. Dreyfus has little problem with this kind of ITS, viewing it as a superlative application of the ability of computers to foster learning. The only danger with this application is that it works so well as to create the temptation to overuse it in the learning process, at the cost of other ways of learning.

A more advanced type of ITS attempts to take on a more active and participatory role in providing advice and instructions, in explaining what the student did wrong, and in selecting the problems and setting the pace for the individual student. This type of ITS is used in teaching complex knowledge and abilities; where the task, for instance, is to master certain theories and concepts and apply them to concrete situations. In order to do this, the ITS must have at its disposal a certain amount of didactic proficiency.

A first objection to this kind of ITS is that they are unsuited for helping students develop genuine expertise in a particular domain, for in order to do this the computer system itself needs to possess expertise. But as argued above, it is impossible to give expertise to computers that have been programmed according to symbolic AI. According to Dreyfus intelligent tutoring systems are well suited to teach a certain measure of competence in an area. They are especially well suited to teaching the early stages of learning, in which the acquisition of rules still plays a major role. But it would be disastrous if such systems were used in later stages of learning, for they only use rules. They would block the acquisition of expertise, which requires at a certain point giving up the use of rules.

Even when ITSes are used only with novices and apprentices, another more serious problem arises. In order to teach well, an ITS must possess not only a great store of specialized knowledge, but also the ability to connect that with the knowledge which the student already possesses, suitably adapting it in the process. A teacher of natural science, for instance, needs to have insight into the naive conceptions about the workings of nature which the students bring into class, and the ability to modify these to make way for the more advanced conceptions. Any ITS would have to possess similar insight and ability.

But the problem is that an ITS is able to express its knowledge and ability only in terms of a number of symbols and rules. The ITS assumes implicitly that the student is a rational, symbol-manipulating, rule-following being. In fact, however, the student is an embodied being that dwells in a human world, and the ITS needs to be able to put itself in the student's place in order to understand where the student is coming from. Because it is unable to do this, an ITS will be unable to help students in seeing the underlying connections which will enable them to master a new knowledge domain. In conclusion, the problem with ITSes used as full-blown teachers is that they cannot help advanced students because they do not themselves possess expertise in the relevant knowledge domain, and they cannot adequately help novices and advanced beginners because their didactic skills are lacking. Dreyfus concludes that existing

intelligent computer systems, especially expert systems and intelligent tutoring systems, foster the impression that the human mind works like a computer. They promote an ultimately erroneous conception of knowledge as something that can be formulated in explicit rules and principles. In the process, the intuitive ability and expertise of human beings, which cannot be grasped through formal rules, becomes devaluated, and students are encouraged to seek knowledge and skills according to the rationalistic model. Eventually this may alter the self-image of human beings to the point where they will begin describing themselves in rationalistic terms as abstract thinking machines. This is the tendency which Dreyfus fears and wants to change.[4]

Conclusion: The Validity and Influence of Dreyfus's Work

Already in 1965, Dreyfus prophesied that symbolic AI would end up failing to achieve a full and complete imitation of human intelligence. Over the years he has systematically criticized the predictions and expectations which have been projected onto new projects and approaches of symbolic AI. And in many respects Dreyfus has proven right. Although symbolic AI has certainly scored a number of successes, the results in many areas have been disappointing. Thus no computer programs have yet been developed which can understand natural language well and answer open-ended questions about a text, which can interpret the meaning of images, which can allow a robot to navigate successfully in a messy environment, and which can solve creative problems. Dreyfus's critique of neural network AI is harder to evaluate for the field is still too young, though the problem of generalization which he describes remains unsolved.

Two recent projects in symbolic AI are worth considering as they have the potential to undermine Dreyfus's position. One recent project, the CYC project headed by AI researcher Douglas Lenat, had as its goal to develop a knowledge base of over one million assertions or 'rules,' that would codify most of the general background knowledge necessary for computer systems to intelligently engage in natural language communication: common sense knowledge including such facts as "People normally wear underwear" and "If an object is not supported by another object or surface, it will fall down." CYC was begun in 1984 as a ten-year project. If successful, it would disprove that the commonsense understanding problem of symbolic AI is unsurmountable, and that common sense can be programmed into a computer, with enough effort. The completed knowledge base of CYC is currently being marketed for application to various tasks, such as intelligent database retrieval, and improved machine translation and speech recognition. Indeed, applications based on CYC's knowledge base may perform better in these and other areas than systems that lack its extensive knowledge base. However, CYC is no longer expected to solve the problem of machine translation, or other major problems in natural

[4] The danger foreseen by Dreyfus that human beings will begin to see themselves as computers has to some extent already come true, as indicated by the psychological studies of Sherry Turkle (1984).

language understanding, which is what one would expect if CYC's common sense were to match human common sense. So the CYC project has not so far refuted the validity of the common sense understanding problem of symbolic AI.[5]

The victory of chess computer Deep Blue over world champion Garry Kasparov in their 1997 match seems to strike a more powerful blow against Dreyfus's position. Deep Blue is a system built upon the principles of symbolic AI. Dreyfus's position that symbolic expert systems cannot perform at the expert level seems to be refuted by the performance of Deep Blue. Dreyfus has always made an exception for expert systems operating in domains of knowledge that are completely formalizable, as some domains of mathematics and formal logic. Yet, whereas the rules of play of chess are fully formalizable, the best move in a chess game cannot be calculated, as it can be in a game like tic-tac-toe. This is because chess games can potentially drag on forever, and there is hence not a finite number of moves to consider that can be evaluated as part of a strategy for winning the game. Chess computers hence have to resort to heuristic rules that anticipate the myriad of possible strategies of the opponent.

Still, there are important differences between the knowledge domain of chess and other domains of expert knowledge that make it unlikely that the success of Deep Blue can be generalized. This is because the domain of chess can be made to resemble a completely formal domain that contains a finite search space in which an optimal solution can be found. If it is taken into account that most chess games are less than one hundred moves, then for practical purposes, the number of possible moves in a chess game is finite, although still too vast to have all possibilities considered. The challenge for chess computers is then to only perform calculations in the more promising parts of this search space. Deep Blue was able to do this with the aid of a vast database containing former chess games played by Kasparov that allowed it to direct its calculations to that part of the search space that Kasparov had occupied in past games. It is unlikely that this strategy of targeted calculation can be applied to other domains of expertise. The problem in other domains, like medicine or economics, is that they cannot be made to resemble a finite, formalized search space to which rules can then be applied, because the elements and regularities in these domains are not formal to begin with. Deep Blue's success hence does not undermine Dreyfus's general position on expert systems. Not only has Dreyfus been proven right in many of his predictions on the success of AI, but AI research has moved more and more in the direction of Dreyfus's alternative theory of intelligence. This is true, for instance, of the emergence of neural network AI, which as Dreyfus points out is fully compatible with his own ideas about intelligence. It is also true of the well-known work of Agre and Chapman at MIT (Agre 1988; Chapman 1991), which is sometimes called "Heideggerian AI" because it tries to implement in AI a number of views promoted by Heidegger and Dreyfus,

[5] See Dreyfus (1992) for a theoretical critique of the CYC project.

such as that intelligence is situated in a world and does not require rules, and that actions can be goal-oriented in the absence of explicitly represented goals.
The situatedness of intelligence is also a central point of departure of the work of the noted AI researcher Terry Winograd and his colleague Fernando Flores. They want to base not only AI research, but also the design of other computer systems on Heideggerian principles.[6]

Winograd and Flores argue that the design of computer systems as well as its internal logic needs to take into account and reflect the fact that these systems must function in a human world and communicate with human users. Computers must be prevented from imposing their own rationalistic logic on the surroundings in which they function.

Even the idea that intelligence presupposes possession of a body has struck a responsive chord in AI research. A recent project at MIT that has drawn much international attention, for instance, is the Cog-project under the direction of Rodney Brooks. The principal assumption of this project is that human intelligence requires human interactions with the world, and therefore a body in which such interactions are possible (Brooks & Stein 1994). Cog is a robot that is equipped with artificial sensory organs (including sensors which keep track of the position of its own body), a voice, and steerable limbs. Cog's "mind" consists of a computer system that is to some extent distributed throughout his body. The aim is to have Cog acquire sensorimotor-intelligence thanks to its sensorimotor interactions with the environment, and perhaps develop "higher" forms of intelligence on top of these more basic abilities.[7]

Much of the inspiration for the development of such work can be traced back to the work of Dreyfus himself. Dreyfus was the one who introduced the ideas of thinkers like Heidegger and Merleau-Ponty into the AI world. The work of such AI researchers as Winograd and Flores, and Agre and Chapman was explicitly inspired by his ideas. But also many other AI researchers, even including followers of symbolic AI like Minsky and John McCarthy, admit that Dreyfus's critiques have influenced their own research. Dreyfus is living proof that philosophers can indeed play an extremely important role as critics of, and commentators on, science and technology in practice.

---

[6] See for instance Winograd and Flores (1986) and Winograd (1995); also the influential work of Suchman (1987).

[7] The notion of intelligence as a situated and embodied phenomenon has also gained ground in psychology and cognitive science. See, e.g., Clark (1996); Varela, Thompson & Rosch (1991); Johnson (1987); Lakoff (1987).

REFERENCES

Agre, Philip

1988, *The Dynamic Structure of Everyday Life*, MIT AI Lab Technical Report 1085.

Anderson, James

1995, *An Introduction to Neural Networks*, MIT Press, Cambridge, MA.

Brooks, Rodney, en Stein, Linda

1994, Building Brains for Bodies, *Autonomous Robotics* 1, 7-25.

Chapman, David

1991, *Vision, Instruction, and Action*, MIT Press, Cambridge, MA.

Clark, Andy

1996, *Being There. Putting Brain, Body, and World Together Again*, MIT Press, Cambridge, MA.

Collins, H. M.

1991, *Artificial Experts*, MIT Press, Cambridge, MA.

Crevier, Daniel

1993, *AI: The Tumultuous History of the Search for Artificial Intelligence*, Basic Books, New York.

Dreyfus, Hubert L.

1965, *Alchemy and Artificial Intelligence*, The RAND Corporation Paper P-3244.

1967, Why computers must have bodies in order to be intelligent, *Review of Metaphysics* 21, 13-32.

1972, *What Computers Can't Do: A Critique of Artificial Reason*, Harper and Row, New York.

1991, *Being-in-the-World: A Commentary on Heidegger's* Being and Time, MIT Press, Cambridge, MA.

1992, *What Computers* Still *Can't Do: A Critique of Artificial Reason*, MIT Press, Cambridge, Massachusetts.

1996, Response to my critics, *Artificial Intelligence* 80, 171-191.

Dreyfus, Hubert L. en Dreyfus, Stuart E.

1986, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York, Free Press.

1988, Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint, *Daedalus* 117, 15-43.

Forester, Tom en Morrison, Perry

1994, *Computer Ethics: Cautionary Tales and Ethical Dilemma's in Computing*, MIT Press,

Cambridge, MA.

Hoven, M. J. van den

1995, *Information Technology and Moral Philosophy*. Proefschrift Erasmus Universiteit, Rotterdam.

Johnson, Mark

1987, *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago, IL, University of Chicago Press.

Lakoff, George

1987, *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*, Chicago, IL, University of Chicago Press.

Meijering, Theo

1993, Neuraal vernuft en gedachteloze kennis. Het moderne pleidooi voor een niet-propositioneel kennismodel, *Algemeen Nederlands tijdschrift voor wijsbegeerte* 85, 24-48.

Suchman, Lucy

1987, *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, Cambridge University Press.

Turkle, Sherry

1984, *The second self: computers and the human spirit*, New York, Simon & Schuster.

Varela, Francisco, Thompson, Evan, en Rosch, Eleanor

1991, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, Cambridge, MA.

Weizenbaum, Joseph

1976, Computerkracht & Mensenmacht. Van Oordeel tot Berekening. (1984) Amsterdam, Contact. (*Computer Power and Human Reason*, Freeman, San Francisco.)

Winograd, Terry

1995, Heidegger and the Design of Computer Systems, in: Feenberg, Andrew en Hannay, Alastair, *Technology and the Politics of Knowledge*, Bloomington and Indianapolis, Indiana University Press.

Winograd, Terry, en Flores, Fernando

1986, *Understanding Computers and Cognition: A New Foundation for Design*, Ablex, Norwood, NJ.