STANFORD ARTIFICIAL INTELLIGENCE PROJECT
MEMO AIM-181

STAN-CS-72-325

REVIEW OF HUBERT DREYFUS' WHAT COMPUTERS CAN'T DO: A CRITIQUE OF ARTIFICIAL REASON

(Harper & Row, New York, 1972)

BY

BRUCE G. BUCHANAN

SUPPORTED BY
National Institutes of Health Career Development Award

November 1972

COMPUTER SCIENCE D'EPARTMENT
School of Humanities and Sciences
STANFORD UN IVERS ITY

STANFORD ARTIFICIAL INTELLIGENCE PROJECT
MEMO AIM-181

NOVEMBER, 1972

COMPUTER SCIENCE DEPARTMENT REPORT
NO. STAN-CS-72-325

REVIEW OF HUBERT DREYFUS' WHAT COMPUTERS CAN'T DO: A CRITIQUE OF ARTIFICIAL REASON (Harper & Row, New York, 1972)

by

Bruce G. Buchanan

ABSTRACT: The recent book What Computers Can't Do by Hubert Dreyfus is an attack on artificial intelligence research. This review takes the position that the philosophical content of the book is interesting, but that the attack on artificial intelligence is not well reasoned.

This work was supported by a National Institutes of Health Career Development Award and partially by SD-183 of ARPA.

Review of Hubert Dreyfus' WHAT COMPUTERS CAN'T DO:
A Critique of Artificial Reason. Harper & Row, New York, 1972.

Bruce G. Buchanan
Computer Science Department
Stanford University

Under the guise of a reasoned criticism of artificial intelligence, Dreyfus juxtaposes the framework of current artificial intelligence research with a current philosophical position known as phenomenology. Because his phenomenological view of man leads him to believe that it is impossible to pick apart the essence of man in a cold, analytical way, he is eager to show that artificial intelligence (AI) and cognitive simulation (CS) computer programs cannot capture the entire range of intelligent behavior exhibited by humans. Unfortunately, his hostility toward computer scientists working in artificial intelligence mars the credibility of the book.

Phenomenology is a philosphical viewpoint much like

existentialism in its emphasis on the value of human experience -- including so-called "subjective" aspects of experience such as emotions -- for interpreting the world. In both views the understanding to be gained is not a scientific, analytical understanding *of* objects in terms of simple entities, but something better described as awareness or intuition of objects in terms *of* their interconnectedness with ourselves and other objects. Its method is often contrasted with the method of linguistic analysis: it stresses the presentation of objects to consciousness as a means of understanding them, as opposed to clearing up misunderstandings by analyzing the ways we talk about things.

Written from a phenomenological point of view, the goals of this book are (1) criticism of AI assumptions, ends, and methods, and (2) suggestion of alternatives, Under the current AI assumptions, the author argues, the goal of producing a digital computer capable of all forms of human intelligent behavior is impossible. The main theme is that AI work has reached a plateau (in terms of what types *of* problems can be successfully solved) and that A? assumptfans, ends, and methods, therefore, should be replaced. This theme is repeated over and over, for example (p. 99): MAIN THEME:

> "The answer to the question whether man can make [an intelligent] machine must rest on the evidence of work being done. And on the basis of actual achievements and current stagnation, the most plausible answer seems to be, No."

It is lamentable that the critique of AI in this book has taken the form of a popular-press attack on At work. The author's phrases are damning but his arguments are not convincing. As the author mentions, the popular press has often given over-enthusiastic impressions of AI work (pp. xxvii - xxvix); this book is wrftten in the same vein but with a negative sign.

As with popular articles, the book gains strength from the reader's association of well-defined concepts with the author's claims. The rigorous concept of impossibility, as used by Godel and Turing, for example, is the concept readers are supposed to associate with the title or with the term 'impossible' used throughout the book. Yet no demonstration of impossibility accompanies the claim. The author substitutes little more than the implausibility argument quo ted above (Main Theme).

The implausibility of AI could itself make an intriguing book, if it were well argued. The reader certainly should not expect impossibility arguments. Unfortunately, the case for the implausibility of AI, also, is supported more by suggestions than by arguments. The reader should be prepared for a diatribe agafnst At which, though colorful in language and example, adds little real support to the implausibility claim.

The support for the implausibility of AI comes from the entire philosophical framework in which this book is written. And because

this is an important current of contemporary philosophy, the book may broaden the viewpoint of readers who are not disposed to read Husserl, Merleau-Panty, Heidegger, Sartre or Wittgenstein on their own. But it must be read with the above cautions in mind.

The Author's Assumptions:

Some of the mistaken fundamental assumptions found in the book are: (D1) AI research aims at one common goal; (D2) The goal of AI is to "program [digital] computers with fully formed Athene-like Intel ligence" (pp. 202-03); (D3) AI work is floundering; (D4) There are limitations on the capabilities of digital computers not shared by analog computers.

Assumption D1 is false. The diverse backgrounds of people working in AI should alert anyone to the diversity *of* goals. Among those goals are representing problem-solving knowledge and problem statements efficaciously (representation theory), designing efficient and appropriate problem-solving strategies for computers (heuristic programming), controlling instruments or machines in complex environments (robotics), communicating with computers in natural language, and applying the current tools to complex problems of science and industry. Any imaginable single common goal woul d be vague, such as "pushing the state of the art of computer science", or "trying to put more intelligence in computer programs".

Because D1 is false, assumption D2 is at least misstated. Even if there were individuals actively pursuing the goal of producing total human intelligence in a computer, calling this the goal of AI research distorts the picture. AI (and CS) workers also aim for understanding human cognitive processes, or producing programs to enhance and complement human thinking, for example, by relieving humans of repetitive tasks. But these are less sensational goals than producing humanoids, and thus are not mentioned.

Throughout the book one is led to believe that AI work is never aimed at the lesser goal of programming only some aspects of intelligent behavior. For example, the author approvingly cites a recent very bad article in LIFE magazine warning of machines in the near future "with the general Intelligence of an average human being" (p. xxviii). Continually, the author seems to be attacking AI workers for failing to reproduce all aspects of intelligent behavior. Yet he correctly notes:

> "No one expects the resulting robot to reproduce everything that counts as intelligent behavior in human beings." (p. xxvi)

Assumption D3 questions whether the state of the art of AI can change from the 1967 position at all. AI 1 of Part I is devoted to a review of AI and CS work in the 1957-1967 period. Vehement criticism of fledgling programs accompanies the review -- most of it missing the point of the early research, that point being to delineate problems and test various computing strategies. Performance within limited

domains has been, and still is, one criterion by which AI strategies are Judged -- but the limited domain should not itself be a reason for criticism.

One can only speculate why the author fails to acknowledge recent AI work. To this reviewer, and other persons doing AI research, programs developed in the last five years seem to outperform programs written in the tool-building period of 1957- 1967. The reader is advised to compare articles in the Journal of Artificial Intelligence, recent proceedings of the International Joint Conference on Artificial Intelligence or recent volumes of Machine Intelligence (Edinburgh University Press) with the descriptions of early programs in COMPUTERS AND THOUGHT. For example, it is dishonest to entitle the book a "critique" of AI when it dwells on the failure of early language translation programs (based primarily on syntactical analysis) without analyzing the recent work on understanding natural language (based on syntax, semantics, and context). Of particular significance are the natural language understanding programs of Woods and Winograd; perception programs from the MIT, Stqnford and SRI laboratories; Colby's simulation of paranoid behavior; the complex reasoning programs of Kling and Feigenbaum, et.al. The author simply would not be willing to call these programs "significant progress" (p. 197). He merely tells the reader that AI results "are sufficiently disappointing to be self-incriminating" (p. 217). One would hope that a criticism of a growing discipline would mention work in the most recent one-third of the years of work. But his discussion of why the

results are disappointing mostly points to problems which AI programs have not attempted to solve or to prior claims which the programs did not meet.

Finally, D4 -- the digital-analog distinction -- keeps coming back every time Dreyfus contrasts information processing on digital computers with human information processing. Also, his final long-run solution is for "nondigital automata" with processing powers different from digital computers. He frequently mentions AI work on digital computers (not Just on computers) to give the impression that the problems he sees are fundamental to digital machines. This is a myth which he seems to recognize as such in the beginning (p. xx), but obscures in the rest of the book.

AI Assumptions:

Part II of the book is an examination of four assumptions the author ascribes to AI workers, with off-hand criticisms of AI methods and goals.

Biological Assumption: Human brains operate digitally at some level.

Psychological Assumption: Human minds can be viewed as digital devices.

Epistemological Assumption: All Knowledge can be formalized and formalized rules can reproduce all intelligent behavior.

Ontological Assumption: There exist Independent determinate elements in terms of which all human Intelligence can be described.

Discussion of the first assumption is mercifully short. Discussion of the second assumption is more of the same kind of tedium found in Part I, The author's discussions of assumptions three and four, however, bring in all the suggestions of the phenomenological point of view which are the topic of Part I I I.

The Biological Assumption:

From a 1961 Newell and Simon statement that computers may be programmed to execute information processes which are functionally much like those carried out by the brain, Dreyfus claims the authors are stating the biological assumption. The first part of the quotation, however, is that "...this [information processing] approach makes no assumption that the 'hardware' of computers and brains are similar." Ye obviously believes that equivalence of function implies equivalence of structure, in spite of the apparent absurdity in arguing from functional equivalence of bird and airplane tails, for example, to structural equivalence.

The Dreyfus argument, as reconstructed from pp. 67-68 is:

(1) Newell & Simon: Both brains and computers are general-purpose symbol manipulating devices;

(2) Dreyfus: "Digital computers...are the only general-purpose symbol manipulating devices which we know how to design";

(3) Dreyfus conclusion: Thus (1) "amounts to a biological

assumption that on some level of operation...the brain processes information in discrete operations by way of some biological equivalent of on/off switches".

The reader may judge this argument for himself. Unfortunately, the biological assumption is used later to support the impossibility of AI as well as to ridicule the supposed naivte of all AI workers.

The Psychological Assumption:

The information processing (IP) model of intelligent behavior is a framework for viewing the mind as a symbol-manipulating device. Without having to say whether the device is digital or analague, most AI researchers (especially those *most* interested in psychology) would affirm (or already have affirmed) the usefulness of the IP model. The IP model is not exactly the same as the psychological assumption because the IP emphasis is on symbol manipulation -- however performed -- and not on discrete operations. Dreyfus cites human processes "such as zeroing in and essential/inessentialdiscrimination" (p. 99) as beyond the scope of the IP model, and notes the failure of AI programs (in the 1957-1967 period) to cope with such processes. The failure is mostly attributed to the digital nature of the machines on which the AI work was programmed.

No CS or AI researcher actually does assume that human cognitve processes are identical with computer processes. Viewing intelligent behavior in an information processing framework does

not imply that the mind's processes are exactly the same, The model serves as an analogy, and it is a model of "macroscopic" behavior, not primitive mechanisms. The nature of the underlying mechanisms is not known; but the model, as any scientific theory, allows us to make inferences which can be useful.

The Epistemological Assumption:

Dreyfus sees AI workers as retreating from the stronger psychological assumption of CS to the weaker epistemological assumption -- that formalizable rules can reproduce Intelligent behavior in a computer, even though these may not be the same rules as humans follow (p.102). He attacks the epistemological assumption by arguing that (a) some intelligent behavior cannot be formalized and (b) formal rules cannot capture the richness of behavior in the human situation. Both (a) and (b)(if they are not identical) depend on the phenomenological framework for credence. Because of that, this is an interesting chapter. However, it would be more interesting if the reader were told what the author means by 'formalization' so the claims of the chapter could be evaluated.

The clash between formalists and non-formalists is always difficult for formalists to understand. The precision of scientific explanation is a useful tool for understanding many aspects of our world. The non-formalist's claim, that aspects of human behavior are too dependent on the whole experience of the individual to be formalizable in terms of discrete parts, thus

appears to the formalist to he sheer romantic nonsense. The difficulty is much more fundamental than merely a dispute over the use *of* terms: it is the result of totally different frameworks for understanding human experience.

Arguments over such fundamental issues are rarely resolved because this is much like a religious debate. In this case, the author points to aspects of human behavior (for example, zeroing in) which seem to resist formalization, and concludes that the epistemological assumption must be wrong. Formal ists, too, should be willing to keep an open mind about alternatives.

The Ontological Assumption:

The ontological assumption -- that the world is analyzable as a set of facts -- is rejected by Dreyfus for nearly the same reasons as the epistemological assumption. In the phenomenologist's view, humans and human situations are not merely collections of individual physical entities. Analyzable facts cannot account for the richness of experience within a human situation.

> "Computers can only deal with facts, but man -- the source *of* facts -- Is not a fact or a set of facts, but a being who creates himself and the world of facts in the process of living in the world. This human world with its recognizable objects is organized by human beings using their embodied capacities to satisfy their embodied needs. There is no reason to suppose that a world organized in terms of these fundamental human capacities should be accessible by any other means." (pp. 202-03)

This is an alternative to the analytical way of viewing the world. But again there is no good argument why it is better. Dreyfus' argument, again, is:

(1) AI work is predicated on the four assumptions discussed above.
(2) AI work is "up against a stone wall" (p.144).
(3) Therefore, these are faulty assumptions.
(4) Therefore, his own alternative assumption is better.

Even if (2) were true, which it is not, (3) does not follow from (1) and (2). And certainly (4) does not follow from (1)-(3).

The criticisms of means and ends of AI work are not nearly as extensive as the criticism of AI assumptions. Throughout the book Dreyfus attacks the method of heuristic search on the grounds that humans are not "unconsciously running with incredible speed through the enormous calculation which would be involved in programming a computer to perform a similar task" (p.165). But AI methods need not parallel human methods, as he has noted. Moreover, it is by no means obvious that humans do not themselves break up continuous aspects of the world into manageable discrete parts.

As mentioned earlier, the goal of AI which he attacks is the goal of programming the entire range of human intelligence in a digital computer. Although this is an interesting position to attack, it is wrong to suppose that AI work has been done with this goal in mind. A more modest goal of producing reasoned solutions to particular problems within constrained contexts has, in fact, guided most AI work to date.

Philosophical Discussion:

Part I I I should he the whole book. The rest of the material discussed in the book is out of the author's realm of expertise, so the philosophical discusston of Part iii gets lost in the noise. Making the criticism of Al central-to the book obscures the author's attempts to Juxtapose the explicit information processing methods of Al programs with the phenomenologist's description of human behavior.

The context in which the reader is forced to read Part ii I, unfortunately, is one in which the phenomenalogical view is seen as providing alternatives to the scientific/formalist Al assumptions. This should not ever have happened. The philosophy can stand on its own; the author's criticisms of Al diminish his effective presentation of the philosophy. Because Part III is supposed to provide alternatives on which to base research in Al (or something akin to Al), the author is forced to talk about a vague short-term "solution" of man-machine cooperative interaction -- which is already the subject of much Al research. And, still vaguer, the author posits future "nondigital automata" programmed (somehow) t o incorporate the phenomenological points of view to guide a11 future Al-11 ke research. Calling this an alternative on which to base future research is, to use the author's own term (p.217), self-incriminating.

Conclusion:

Al work is following the analytical and empiricist currents

in Western thought, as Dreyfus points out, and thus builds *from* many of the same assumptions. The unstated aim *of* the book seems to be to criticize the empiricist framework and substitute the . phenomenological way of viewing the world. But the criticisms of AI work are not valid. Also, his reasons for adopting the phenomenological point of view consist largely of examples of human behavior which seem to be difficult for AI programs to emulate.

Dreyfus admits that his alternative "is vaguer and less experimental than that of either the behaviorists or intellectualists which it is meant to supplant" (p.145). That is certainly true. Nevertheless, Part III is interesting reading, for here phenomenalogy is presented as a positive position. If there is any reason to read the book at all, it is to become acquainted with this current view of man and the world which is different from the traditional scientific view.