WHAT COMPUTERS CAN'T DO

A CRITIQUE

NEW YORK, EVANSTON, SAN FRANCISCO, LONDON

WHAT COMPUTERS CAN'T DO

OF ARTIFICIAL REASON

By Hubert L. Dreyfus

HARPER & ROW, PUBLISHERS

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.

The Future of Artificial Intelligence

But these difficulties give us no idea of the future of artificial intelligence. Even if the attempt to program isolated intelligent activities always ultimately requires the programming of the whole mature human form of life, and even if an Athene-like digital computer is impossible in principle—that is, even if mature human intelligence is organized in terms of a field which is reciprocally determined by the objects in it and capable of radical revision—the question still remains to what extent workers in artificial intelligence can use their piecemeal techniques to approximate intelligent human behavior. In order to complete our analysis of the scope and limits of artificial reason we must now draw out the practical implications of the foregoing arguments.

Before drawing our practical conclusions, however, it will be helpful to distinguish four areas of intelligent activity. We can then determine to what extent intelligent behavior in each area presupposes the four human forms of "information processing" we distinguished in Part I. This will enable us to account for what success has been attained and predict what further progress can be expected.

One can distinguish four types of intelligent activity (see Table 1). We have seen that the first two types are amenable to digital computer simulation, while the third is only partially programmable and the fourth is totally intractable.

Area I is where the S–R psychologists are most at home. It includes all forms of elementary associationistic behavior where meaning and context are irrelevant to the activity concerned. Rote learning of nonsense syllables is the most perfect example of such behavior so far programmed, although any form of conditioned reflex would serve as well.

Table 1

CLASSIFICATION OF INTELLIGENT ACTIVITIES

I. Associationistic	II. Simple Formal	III. Complex Formal	IV. Nonformal
Characteristics of Activity			
Irrelevance of meaning and situation.	Meanings completely explicit and situation independent.	In principle, same as II; in practice, internally situation-dependent, independent of external situation.	Dependent on meaning and situation which are not explicit.
Innate or learned by repetition.	Learned by rule.	Learned by rule and practice.	Learned by perspicuous examples.
Field of Activity (and Appropriate Procedure)			
Memory games, e.g., "Geography" (association).	Computable or quasi-computable games, e.g., nim or tic-tac-toe (seek algorithm or count out).	Uncomputable games, e.g., chess or go (global intuition and detailed counting out).	Ill-defined games, e.g., riddles (perceptive guess).
Maze problems (trial and error).	Combinatorial problems (nonheuristic means/ends analysis).	Complex combinatorial problems (planning and maze calculation).	Open-structured problems (insight).
Word-by-word translation (mechanical dictionary).	Proof of theorems using mechanical proof procedures (seek algorithm).	Proof of theorems where no mechanical proof procedure exists (intuition and calculation).	Translating a natural language (understanding in context of use).
Response to rigid patterns (innate releasers and classical conditioning).	Recognition of simple rigid patterns, e.g., reading typed page (search for traits whose conjunction defines class membership).	Recognition of complex patterns in noise (search for regularities).	Recognition of varied and distorted patterns (recognition of generic or use of paradigm case).
Kinds of Program			
Decision tree, list search, template.	Algorithm.	Search-pruning heuristics.	None.

Also some games, such as the game sometimes called Geography (which simply consists of finding a country whose name begins with the last letter of the previously named country), belong in this area. In language translating, this is the level of the mechanical dictionary; in problem solving, that of pure trial-and-error search routines; in pattern recognition, matching pattern against fixed templates.

Area II is the domain of Pascal's *esprit de géométrie*—the terrain most favorable for artificial intelligence. It encompasses the conceptual rather than the perceptual world. Problems are completely formalized and completely calculable. For this reason, it might best be called the area of the simple-formal. Here artificial intelligence is possible in principle and in fact.

In Area II, natural language is replaced by a formal language, of which the best example is logic. Games have precise rules and can be calculated out completely, as in the case of nim or tic-tac-toe. Pattern recognition on this level takes place according to determinate types, which are defined by a list of traits characterizing the individuals which belong to the class in question. Problem solving takes the form of reducing the distance between means and ends by repeated application of formal rules. The formal systems in this area are simple enough to be manipulated by algorithms which require no search procedure at all (for example, Wang's logic program). Heuristics are not of unnecessary here, they are a positive handicap, as the superiority of Wang's algorithmic logic program over Newell, Shaw, and Simon's heuristic logic program demonstrates. In this area, artificial intelligence has had its only unqualified successes.

Area III, complex-formal systems, is the most difficult to define and has generated most of the misunderstandings and difficulties in the field. It contains behavior which is in principle reproducible but in fact intractable. As the number of elements increases, the number of transformations required grows exponentially with the number of elements involved. As used here, "complex-formal" includes those systems which in practice cannot be dealt with by exhaustive enumeration algorithms (chess, go, etc.), and thus require heuristic programs.[1]*

Area IV might be called the area of nonformal behavior. This includes

all those everyday activities in our human world which are regular but not rule governed. The most striking example of this controlled imprecision is our disambiguation of natural languages. This area also includes games in which the rules are not definite, such as guessing riddles. Pattern recognition in this domain is based on recognition of the generic, or of the typical, by means of a paradigm case. Problems on this level are open-structured, requiring a determination of what is relevant and insight into which operations are essential, before the problem can be attacked.[2]* Techniques on this level are usually taught by generalizing from examples and are followed intuitively without appeal to rules. We might adopt Pascal's terminology and call Area IV the home of the *esprit de finesse.* Since in this area a sense of the global situation is necessary to avoid storing an infinity of facts, it is impossible in principle to use discrete techniques to reproduce directly adult behavior. Even to order the four as in Table 1 is misleadingly encouraging, since it suggests that Area IV differs from Area III simply by introducing a further level of complexity, whereas Area IV is of an entirely different order than Area III. Far from being more complex, it is really more primitive, being evolutionarily, ontogenetically, and phenomenologically prior to Areas II and III, just as natural language is prior to mathematics.

The literature of artificial intelligence generally fails to distinguish these four areas. For example, Newell, Shaw, and Simon announce that their logic theorist "was devised to learn how it is possible to solve difficult problems such as proving mathematical theorems [II or III], discovering scientific laws from data [III and IV], playing chess [III], or understanding the meaning of English prose [IV]."[3] The assumption, made explicitly by Paul Armer of the RAND Corporation, that all intelligent behavior is of the same general type, has encouraged workers to generalize from success in the two promising areas to unfounded expectation of success in the other two.

This confusion has two dangerous consequences. First there is the tendency, typified by Simon, to think that heuristics discovered in one field of intelligent activity, such as theorem proving, must tell us something about the "information processing" in another area, such as the understanding of a natural language. Thus, certain simple forms of infor-

mation processing applicable to Areas I and II are imposed on Area IV, while the unique form of "information processing" in this area, namely that "data" are not being "processed" at all, is overlooked. The result is that the same problem of exponential growth that causes trouble when the techniques of Areas I and II are extended to Area III shows up in attempts to reproduce the behavior characteristic of Area IV.[4*]

Second, there is the converse danger. The success of artificial intelligence in Area II depends upon avoiding anything but discrete, determinate, situation-free operations. The fact that, like the simple systems in Area II, the complex systems in Area III are formalizable leads the simulator to suppose the activities in Area III can be reproduced on a digital computer. When the difference in degree between simple and complex systems turns out in practice, however, to be a difference in kind —exponential growth becoming a serious problem—the programmer, unaware of the differences between the two areas, tries to introduce procedures borrowed from the observation of how human beings perform the activities in Area IV—for example, position evaluation in chess, means-ends analysis in problem solving, semantic considerations in theorem proving—into Area III. These procedures, however, when used by human beings depend upon one or more of the specifically human forms of "information processing"—for human beings at least, the use of chess heuristics presupposes fringe consciousness of a field of strength and weakness; the introduction of means-ends analysis eventually requires planning and thus a distinction between essential and inessential operations; semantic considerations require a sense of the context.

The programmer confidently notes that Area III is in principle formalizable just like Area II. He is not aware that in transplanting the techniques of Area IV into Area III he is introducing into the continuity between Areas II and III the discontinuity which exists between Areas III and IV and thus introducing all the difficulties confronting the formalization of nonformal behavior. Thus the problems which in principle should only arise in trying to program the "ill-structured," that is, open-ended activities of daily life, arise in practice for complex-formal systems. Since what counts as relevant data in Area III is completely explicit, heuristics can work to some extent (as in Samuel's Checker

Program), but since Area IV is just that area of intelligent behavior in which the attempt to program digital computers to exhibit fully formed adult intelligence must fail, the unavoidable recourse in Area III to heuristics which presuppose the abilities of Area IV is bound, sooner or later, to run into difficulties. Just how far heuristic programming can go in Area III before it runs up against the need for fringe consciousness, ambiguity tolerance, essential/inessential discrimination, and so forth, is an empirical question. However, we have seen ample evidence of trouble in the failure to produce a chess champion, to prove any interesting theorems, to translate languages, and in the abandonment of GPS.

Still there are some techniques for approximating some of the Area IV short-cuts necessary for progress in Area III, without presupposing the foregoing human forms of "information processing" which cannot be reproduced in any Athena-like program.

To surmount present stagnation in Area III the following improved techniques seem to be required:

1. Since current computers, even primitive hand-eye coordinating robots, do not have bodies in the sense described in Chapter 7, and since no one understands or has any idea how to program the global organization and indeterminacy which is characteristic of perception and embodied skills, the best that can be hoped for at this time is some sort of crude, wholistic, first-level processing, which approximates the human ability to zero in on a segment of a field of experience before beginning explicit rule-governed manipulation or counting out. This cannot mean adding still further explicit ways of picking out what area is worth exploring further. In chess programs, for example, it is beginning to be clear that adding more and more specific bits of chess knowledge to plausible move generators, finally bogs down in too many *ad hoc* subroutines. (Samuel thinks this is why there has been no further progress reported for the Greenblatt chess program.[5]) What is needed is something which corresponds to the master's way of seeing the board as having promising and threatening areas.

Just what such wholistic processing could be is hard to determine, given the discrete nature of all computer calculations. There seem to be two different claims in the air. When Minsky and Papert talk of finding

"global features," they seem to mean finding certain isolable, and determinate, features of a pattern (for example, certain angles of intersection of two lines) which allow the program to make reliable guesses about the whole. This just introduces further heuristics and is not wholistic in any interesting sense. Neisser, however, in discussing the problem of segmenting shapes for pattern recognition before analyzing them in detail makes a more ambitious proposal.

> Since the processes of focal attention cannot operate on the whole visual field simultaneously, they can come into play only after preliminary operations have already segregated the figural units involved. These preliminary operations are of great interest in their own right. They correspond in part to what the Gestalt psychologists called "autochthonous forces," and they produce what Hebb called "primitive unity." I will call them the *preattentive processes* to emphasize that they produce the objects which later mechanisms are to flesh out and interpret.
>
> The requirements of this task mean that the preattentive processes must be genuinely "global" and "wholistic." Each figure or object must be separated from the others in its entirety, as a potential framework for the subsequent and more detailed analyses of attention.[6]

But Neisser is disappointing when it comes to explaining how this crude, first approximation is to be accomplished by a digital computer. He seems to have in mind simply cleaning-up heuristics which, as Neisser implicitly admits, only work where the patterns are already fairly clearly demarcated. "Very simple operations can separate units, *provided they have continuous contours or empty spaces between them.* Computer programs which follow lines or detect gaps, for example, are as easily written as those which fill holes and wipe out local irregularities."[7] But such techniques fail, for example, in the case of cursive script.

Of course, it is hard to propose anything else. What is being asked for is a way of dealing with the field of experience before it has been broken up into determinate objects, but such preobjective experience is, by definition, out of bounds for a digital computer. Computers must apply specific rules to determinate data; if the problem is one of first carving out the determinate data, the programmer is left with the problem of applying determinate rules to a blur.

The best that can be hoped in trying to circumvent the techniques of

Area IV, therefore, may well be the sort of clever heuristics Minsky and Papert propose to enable a first-pass program to pick out certain specific features which will be useful in directing the program in filling in more details. But such *ad hoc* techniques risk becoming unmanageable and in any case can never provide the generality and flexibility of a partially determinate global response.

2. A second difficulty shows up in connection with *representing* the problem in a problem-solving system. It reflects the need for essential/inessential discrimination. Feigenbaum, in discussing problems facing artificial intelligence research in the second decade, calls this problem "the most important though not the most immediately tractable."[8] He explains the problem as follows:

> In heuristic problem solving programs, the search for solutions within a problem space is conducted and controlled by heuristic rules. The representation that defines the problem space is the problem solver's "way of looking at" the problem and also specifies the form of solutions. Choosing a representation that is right for a problem can improve spectacularly the efficiency of the solution-finding process. The choice of problem representation is the job of the human programmer and is a creative act.[9]

This is the activity we called finding the deep structure or insight. Since current computers, even current primitive robots, do not have needs in the sense we have discussed in Chapter 9, and since no one has any idea how to program needs into a machine, there is no present hope of dispensing with this "creative act." The best that can be expected at this time is the development of programs with specific objectives which take an active part in organizing data rather than passively receiving them. Programmers have noticed that, in the analysis of complex scenes, it is useful to have the program formulate an hypothesis about what it would expect to find on the basis of data it already has, and look for that. This should not be confused with the way the human being organizes *what counts as data* in terms of his field of purposes. All that can be expected is fixed rules to apply to fixed data; that is, there will be a programmed set of alternatives, and the program can, on the basis of present data, select one of these alternatives as the most probable and look for further data on the basis of this prediction.

Thus, specific long-range objectives or a set of alternative long-range objectives might be built into game-playing and problem-solving programs, so that in certain situations certain strategies would be tried by the computer (and predicted for the opponent). This technique, of course, would not remove the restriction that all these alternatives must be explicitly stored beforehand and explicitly consulted at certain points in the program, whereas human purposes implicitly organize and direct human activity moment by moment. Thus even with these breakthroughs the computer could not exhibit the flexibility of a human being solving an open-structured problem (Area IV), but these techniques could help with complex-formal problems such as strategy in games and long-range planning in organizing means-ends analysis.

3. Since computers are not in a situation, and since no one understands how to begin to program primitive robots, even those which move around, to have a world, computer workers are faced with a final problem: how to program a representation of the computer's environment. We have seen that the present attempt to store all the facts about the environment in an internal model of the world runs up against the problem of how to store and access this very large, perhaps infinite amount of data. This is sometimes called the large data base problem. Minsky's book, as we have seen, presents several *ad hoc* ways of trying to get around this problem, but so far none has proved to be generalizable.

In spite of Minsky's claims to have made a first step in solving the problem, C. A. Rosen in discussing current robot projects after the work reported in Minsky's book acknowledges new techniques are still required:

> We can foresee an ultimate capability of storing an encyclopedic quantity of facts about specific environments of interest, but new methods of organization are badly needed which permit both rapid search and logical deductions to be made efficiently.[10]

In Feigenbaum's report, there is at last a recognition of the seriousness of this problem and even a suggestion of a different way to proceed. In discussing the mobile robot project at the Stanford Research Institute, Feigenbaum notes:

It is felt by the SRI group that the most unsatisfactory part of their simulation effort was the simulation of the environment. Yet, they say that 90% of the effort of the simulation team went into this part of the simulation. It turned out to be very difficult to reproduce in an internal representation for a computer the necessary richness of environment that would give rise to interesting behavior by the highly adaptive robot.[11]

We have seen that this problem is avoided by human beings because their model of the world is the world itself. It is interesting to find work at SRI moving in this direction.

It is easier and cheaper to build a hardware robot to extract what information it needs from the real world than to organize and store a useful model. Crudely put, the SRI group's argument is that the most economic and efficient store of information about the real world is the real world itself.[12]

This attempt to get around the large data base problem by recalculating much of the data when needed is an interesting idea, although how far it can go is not yet clear. It presupposes some solution to the wholistic problem discussed in 1 above, so that it can segment areas to be recognized. It also would require some way to distinguish essential from inessential facts. Most fundamentally, it is of course limited by having to treat the real world, whether stored in the robot memory or read off a TV screen, as a set of facts; whereas human beings organize the world in terms of their interests so that facts need be made explicit only insofar as they are relevant.

What can we expect while waiting for the development and application of these improved techniques? Progress can evidently be expected in Area II. As Wang points out, "we are in possession of slaves which are . . . persistent plodders."[13] We can make good use of them in the area of simple-formal systems. Moreover, the protocols collected by Newell, Shaw, and Simon suggest that human beings sometimes operate like digital computers, within the context of more global processes. Since digital machines have symbol-manipulating powers superior to those of humans, they should, so far as possible, take over the digital aspects of human "information processing."

To use computers in Areas III and IV we must couple their capacity for fast and accurate calculation with the short-cut processing made possible by fringe-consciousness, insight, and ambiguity tolerance. Leibniz already claimed that a computer "could enhance the capabilities of the mind to a far greater extent than optical instruments strengthen the eyes." But microscopes and telescopes are useless without the selecting and interpreting eye itself. Thus a chess player who could call on a machine to count out alternatives once he had zeroed in on an interesting area would be a formidable opponent. Likewise, in problem solving, once the problem is structured and an attack planned, a machine could take over to work out the details (as in the case of machine-shop allocation or investment banking). A mechanical dictionary which could display meanings on a scope ranked as to their probable relevance would be useful in translation. In pattern recognition, machines are able to recognize certain complex patterns that the natural prominences in our experience lead us to ignore. Bar-Hillel, Oettinger, and John Pierce have each proposed that work be done on systems which promote a symbiosis between computers and human beings. As Walter Rosenblith put it at a recent symposium, "Man *and* computer is capable of accomplishing things that neither of them can do alone."[14]

Indeed, the first successful use of computers to augment rather than replace human intelligence has recently been reported. A theorem-proving program called SAM (Semi-Automated Mathematics) has solved an open problem in lattice theory. According to its developers:

> Semi-automated mathematics is an approach to theorem-proving which seeks to combine automatic logic routines with ordinary proof procedures in such a manner that the resulting procedure is both efficient and subject to human intervention in the form of control and guidance. Because it makes the mathematician an essential factor in the quest to establish theorems, this approach is a departure from the usual theorem-proving attempts in which the computer *unaided* seeks to establish proofs.[15]

One would expect the mathematician, with his sense of relevance, to assist the computer in zeroing in on an area worth counting out. And this is exactly what happens.

The user may intervene in the process of proof in a number of ways. His selection of the initial formulas is of course an important factor in determining the course AUTO-LOGIC will take. Overly large or ill-chosen sets of initial formulas tend to divert AUTO-LOGIC to the proving of trivial and uninteresting results so that it never gets to the interesting formulas. Provided with a good set of initial formulas, however, AUTO-LOGIC will produce useful and interesting results. As the user sees that AUTO-LOGIC is running out of useful ways in which to use the original formulas, he can halt the process and insert additional axioms or other material. He can also guide the process by deleting formulas which seem unimportant or distracting. This real-time interplay between man and machine has been found to be an exciting and rewarding mode of operation.[16]

Instead of trying to make use of the special capacities of computers, however, workers in artificial intelligence—blinded by their early success and hypnotized by the assumption that thinking is a continuum—will settle for nothing short of unaided intelligence. Feigenbaum and Feldman's anthology opens with the baldest statement of this dubious principle:

In terms of the continuum of intelligence suggested by Armer, the computer programs we have been able to construct are still at the low end. What is important is that we continue to strike out in the direction of the milestone that represents the capabilities of human intelligence. Is there any reason to suppose that we shall never get there? None whatever. Not a single piece of evidence, no logical argument, no proof or theorem has ever been advanced which demonstrates an insurmountable hurdle along the continuum.[17]

Armer prudently suggests a boundary, but he is still optimistic:

It is irrelevant whether or not there may exist some upper bound above which machines cannot go in this continuum. Even if such a boundary exists, there is no evidence that it is located close to the position occupied by today's machines.[18]

Current difficulties, once they are interpreted independently of optimistic *a priori* assumptions, however, suggest that the areas of intelligent behavior are *discontinuous* and that *the boundary is near.* The stagnation of each of the specific efforts in artificial intelligence suggests that there can be no piecemeal breakthrough to fully formed adult intelligent behavior for any isolated kind of human performance. Game playing, language translation, problem solving, and pattern recognition, each

depends on specific forms of human "information processing," which are in turn based on the human way of being in the world. And this way of being-in-a-situation turns out to be unprogrammable in principle using presently conceivable techniques.

Alchemists were so successful in distilling quicksilver from what seemed to be dirt that, after several hundred years of fruitless efforts to convert lead into gold, they still refused to believe that on the chemical level one cannot transmute metals. They did, however, produce—as by-products—ovens, retorts, crucibles, and so forth, just as computer workers, while failing to produce artificial intelligence, have developed assembly programs, debugging programs, program-editing programs, and so on, and the M.I.T. robot project has built a very elegant mechanical arm.

To avoid the fate of the alchemists, it is time we asked where we stand. Now, before we invest more time and money on the information-processing level, we should ask whether the protocols of human subjects and the programs so far produced suggest that computer language is appropriate for analyzing human behavior: Is an exhaustive analysis of human reason into rule-governed operations on discrete, determinate, context-free elements possible? Is an approximation to this goal of artificial reason even probable? The answer to both these questions appears to be, No.

Does this mean that all the work and money put into artificial intelligence have been wasted? Not at all, if instead of trying to minimize our difficulties, we try to understand what they show. The success and subsequent stagnation of Cognitive Simulation and of AI, plus the omnipresent problems of pattern recognition and natural language understanding and their surprising difficulty, should lead us to focus research on the four human forms of "information processing" which they reveal and the situational character of embodied human reason which underlies them all. These human abilities are not necessary in those areas of intelligent activity in which artificial intelligence has had its early success, but they are essential in just those areas of intelligent behavior in which artificial intelligence has experienced consistent failure. We can then view recent work in artificial intelligence as a crucial experiment disconfirming the

traditional assumption that human reason can be analyzed into rule-governed operations on situation-free discrete elements—the most important disconfirmation of this metaphysical demand that has ever been produced. This technique of turning our philosophical assumptions into technology until they reveal their limits suggests fascinating new areas for basic research.

C. E. Shannon, the inventor of information theory, sees, to some extent, how different potentially intelligent machines would have to be. In his discussion of "What Computers Should be Doing," he observes:

> Efficient machines for such problems as pattern recognition, language translation, and so on, may require a different type of computer than any we have today. It is my feeling that this will be a computer whose natural operation is in terms of patterns, concepts, and vague similarities, rather than sequential operations on ten-digit numbers.[19]

We have seen that, as far as we can tell from the only being that can deal with such "vagueness," a "machine" which could use a natural language and recognize complex patterns would have to have a body so that it could be at home in the world.

But if robots for processing nonformal information must be, as Shannon suggests, entirely different from present digital computers, what can now be done? Nothing directly toward programming present machines to behave with human intelligence. We must think in the short run of cooperation between men and digital computers, and only in the long run of nondigital automata which, if they were in a situation, would exhibit the forms of "information processing" essential in dealing with our nonformal world. Artificial Intelligence workers who feel that some concrete results are better than none, and that we should not abandon work on artificial intelligence until the day we are in a position to construct such artificial men, cannot be refuted. The long reign of alchemy has shown that any research which has had an early success can always be justified and continued by those who prefer adventure to patience.[20]* If one insists on *a priori* proof of the impossibility of success, it is difficult, as we have seen, to show that such research is misguided except by denying very fundamental assumptions common to all science.

And one can always retort that at least the goal can be approached. If, however, one is willing to accept empirical evidence as to whether an effort has been misdirected, he has only to look at the predictions and the results. Even if there had been no predictions, only hopes, as in language translation, the results are sufficiently disappointing to be self-incriminating.

If the alchemist had stopped poring over his retorts and pentagrams and had spent his time looking for the deeper structure of the problem, as primitive man took his eyes off the moon, came out of the trees, and discovered fire and the wheel, things would have been set moving in a more promising direction. After all, three hundred years after the alchemists we did get gold from lead (and we have landed on the moon), but only after we abandoned work on the alchemic level, and worked to understand the chemical level and the even deeper nuclear level instead.