COMPUTERS IN BEHAVIORAL SCIENCE

MACHINE TRACES AND PROTOCOL STATEMENTS

by Daniel C. Dennett

Department of Philosophy, University of California at Irvine

In this paper the author examines some assumptions underlying efforts in the computer simulation of cognitive processes. It is argued that the validity of simulations as confirmations of theories or models depends on there being logically analogous relations between machine trace and computer operations on the one hand, and human protocols and human operations on the other. This assumption is shown to be highly speculative and considerations are raised for and against it.

My purpose is to examine some presuppositions underlying current interpretations of results in the area of computer simulation of cognitive processes, such as Newell and Simon's General Problem Solver (1961). I hope to show that once these presuppositions are made explicit, recent criticisms of these endeavors can be shown to be confused, but the same considerations that sink the critics will point to the need for a change in the goals and defenses of the investigators.

Feigenbaum and Feldman, in the introduction to their anthology, *Computers and Thought* (1963), present the immediate goal of efforts in computer simulation as they see it: "to construct computer programs which exhibit behavior that we call 'intelligent behavior' when we observe it in human beings" (p. 3). They distinguish between research in artificial intelligence, in which the aim is to produce programs that solve problems in any way whatever, and simulation, which has the further aim of producing programs that solve problems in the same way people do. Typical procedure in the latter activity is to construct one's program so that it prints out a play-by-play of its operations in the course of searching for a solution, and this "machine trace" or "program trace" is then compared with the "protocol statements" of a human subject or subjects describing their own efforts to solve the same problem. The burden of simulation, as contrasted with artificial intelligence, is to show that the programs arrived at do work in the same way people do, and it is held that this is adequately demonstrated if there is a high degree of correspondence between the machine trace and the protocol statements. It is this central tenet regarding confirmation in simulation that needs examination.

In order to justify the delicate business of using first person "introspective" accounts of human problem solvers, so much abhorred by behaviorists, the simulators have devised a rationale to the effect that they are simply trying to construct computer programs that will simulate the "stream of behavior" (for example, Newell and Simon, 1961), including that which issues from the mouth, of the human problem solver. But what counts as the person's stream of behavior to be simulated? Presumably one is to ignore fidgets, pencil chewing, and so forth, and concentrate instead on what he writes down and says. But can one give a purely behavioral criterion (one involving only statements about physical motions, with no reference to goals, meanings, or other teleological or mentalistic notions) that will exclude from consideration the "Um's" and "Ah me's" while allowing the "compare's" and "conclude's?" This possibility is very dubious, and Charles Taylor (1964) presents an excellent general case for the impossibility of providing adequate "physical motion" descriptions of any higher behavior, from maze running on up. Even if it were granted that somehow this culling out of the essential

behavior is possible within the restrictions of behaviorism, there remains the question of what is to count as the relevant behavior in the machine. Apparently the printout, rather than it together with the internal machinations of the computer, is to be the machine's behavior. But then how does the printout in any way resemble the behavior of the person being modeled? To take an example from Newell and Simon, in what physical ways could the typing of the marks "GOAL 2 DELETE R FROM LI" resemble any sound waves or lip motions the subject could make? There is no more physical similarity between the typing of the marks "DELETE" (and to remain behaviorally pure, series of ink patterns cannot be called words) and the uttering of the sound "delete" than there is between an elephant and the word "elephant." One might suppose that this difficulty could be overcome by having the subject type out his protocols on a computer terminal, and use the same computer language in the bargain, but here, clearly, we have lost track of which is the model and which the modeled. The point of comparisons is lost if the subject is asked, in effect, to attempt to simulate the behavior of the computer.

Only allegiance to ingrained dogmas can obscure the fact that it is only by treating the computer printout and the human protocol statements not as semantically neutral productions of motions, but as significant, referring reports, that the desired comparisons between them can be made. This can be brought out by the fact that the hoped for similarities can exist between the machine printout and the person's speech whether he speaks English or French or uses sign language; and in addition the similarities no doubt exist between a French speaking person and an English speaking person. It would be ridiculous to maintain that French people solve problems differently from English people because they use a different language. Whether or not there are important differences in the way the French and English solve problems is an empirical question that does not begin to be answered by the observation that they use different languages. So, in spite of disclaimers, the simulator seeks a semantic correspondence between machine trace and protocol statements, as he makes quite clear by making no attempt to simulate idiomatic English, being quite content with an artificial computer language.

The simulator must therefore abandon the strict behavioristic stance if he is to get on with his work. He is not examining two streams of behavior, but two reports of streams of behavior—hidden behavior. We can have other access to the computer's hidden behavior (through some form of monitoring that might be devised were there any point to it), but as yet we have no detailed access to the inner workings of the brain. The crucial question now becomes: Is the programmed printout an adequate analogue of the person's protocol? That is, since both are being treated as reports—the former on the grounds that we know how it is produced according to the program, and the latter on the grounds of fallible common sense—it must be shown that the reports bear the same relation to the presumed subject matter, what is being reported, if comparisons are to be of any value in confirming cases of successful simulation.

To bring this out, consider what would be amiss in the following situation. The computer prints out its running report, and at the same time the human problem solver stays mute. He is observed however by a self-styled telepathist, who reports "Now he is wondering whether to divide or subtract." Or in place of the telepathist one could have a psychologist armed with a theory about correlations between facial expressions and problem solving, "deducing" the same protocol as the telepathist, perhaps with the aid of EEG readings of the subject's brain. In these cases it is clear that the reports of the second person do not bear the same relation to their subject matter as do the machine traces to theirs, and hence no comparisons between the two can be treated as confirming or disconfirming the theory behind the model.

What, then, must be presumed to be the relation between a person's protocol and its subject matter, the "inner" operations, if a comparison of protocol and machine trace is to be used for theory confirmation? Using the programmed machine trace as a stand-

ard, certain possible relations between protocol and subject matter must be ruled out right away, some for the better no doubt. A particularly eccentric possibility would be that the person has an inner eye (a real, seeing eye, not a "figurative" eye) which observes the microscopic crossings of synapses and so forth, and on the basis of a God-given theory of how these basic processes add up to information processing, is able to say "And now I am adding, and now comparing two results..." This is ruled out because it is not the way the computer arrives at its printout. It does not observe, in this sense at least, its own workings and then draw inferences about their interpretation. We can also rule out more plausible mechanical scanning systems—"wire tap" systems monitoring individual neuronal functions—for this is also not the way of the computer. And since no one is prepared to account for computer printouts in terms of the activity of a "ghost in the machine," traditional Cartesian dualisms must also be ruled out as acceptable views of the human capacity of introspection. This should make it clear that when confirmation of theory is claimed, simulation of cognitive processes involves presuppositions that make it far from theory-pure when it comes to the nature of the mind and introspection. Admittedly, however, the views that have been eliminated so far would not be mourned by many. The exact view of the nature of human introspection that must be presupposed by the simulators can be revealed by examining the relation between the computer's operations and its machine trace.

The computer does not produce its printout on the basis of a physical scan of its own workings, but as a programmed by-product of the very operations it is performing. That this differs in an important respect from the use of a physical scan can be brought out by examining what might be called epistemic differences between the two sorts of results.

Operations in a computer are defined by means of a program or stack of programs, not in physical terms but in functional terms. That is, whereas a door can be (physically) open whether or not anything recognizes it as open, walks through it, or mistakes it for being closed, a logic gate is open by definition for the program, if and only if it functions in a particular way within the computations. This is not to say that it is often the case that a logic gate is physically open but functionally closed, for if this were so, computers could not be made to be reliable; but to say that the criterion to be used when talking of logic gates in the context of information processing rather than in the context of engineering or construction is the functional criterion.

Now suppose we have a physical criterion, measured by some sensing device, for a particular flip-flop in a computer being on, and suppose further that this physical criterion is logically independent of the flip-flop's operating properly. That is, suppose the criterion involves a discovered contingent relationship between the flip-flop's being functionally on and, say, its temperature. We then would have two different criteria for the state *flip-flop A is on*, a functional criterion and an independent physical criterion. Under these circumstances it would be possible for a sensing or scanning mechanism to determine and print out without malfunction on its part that flip-flop A was on when functionally it was off. The fact that this might contingently never occur does not change the epistemic situation. A provision for printout, on the other hand, that was initiated by the functional criterion could not logically fall into this kind of error. Barring malfunction in printout, it simply cannot have misdetermined that flip-flop A is on. Put anthropomorphically, in the scanning case it could "seem" to the scanning mechanism that flip-flop A was on, when in fact it was off, while in the function-relative case, flip-flop A's seeming to be on would always necessarily be a case of flip-flop A's actually being on. In this latter case, what we have is not a scan that determines whether flip-flop A is on, but a programmed direction for printout if flip-flop A is on. Thus in the case of the programmed printout there is a certain sort of infallibility barring self-correctable "typographical errors" that is lacking in the case of a scanning mechanism, and this epistemic distinction generalizes to cover reports of operations at all levels of complexity. Thus the reports of a printout capacity involving no extra scan-

ning mechanisms but built in at some level of the programming stands in quite a different epistemic relation to its subject matter, the functioning of the computer, than would the reports of an attached scanning device.

This distinction, developed along somewhat different lines by Hilary Putnam (1960), is very suggestive when applied to the case of human protocol statements. Does the human brain when it functions to produce a "machine trace" operate like a computer? Is the printout programmed in, somehow allowing for inhibition of printout under most circumstances? This might go some way in explaining the commonly acknowledged infallibility, barring correctable slips of the tongue, of such introspective reports as "I seem to see something red." The answer may well be yes, but it is important to note that the positive answer is already presupposed by the simulators making comparisons between machine traces and protocol statements, and the answer is far from obvious on the basis of present neurological data. It is at least possible that the brain operates on quite different principles, and these may rule out telling comparisons between protocol statements and machine traces.

Getting straight on the necessity of this large presupposition is not a merely pro forma girding of the methodolgical loins, but has quite direct application to interpretations and criticisms of current results in the field. The machine trace of the solution of a particular problem may report many steps, particularly of the mindless trial and error sort of series known as "brute force" computing, that do not appear in the subject's protocol, and are in fact explicitly denied by the subject, such as "I certainly didn't methodically check each piece on the chessboard before concluding my rook was unguarded." What can be concluded from this discrepancy? Some critics have tried to argue that this is evidence that the computer and the subject are using very different methods, but this does not follow at all. In the case of the computer, there is a certain limit to the depth of analysis of the printout, determined by the language of the printout. Ordinarily the printout is in a high order language rather than in basic machine language, and hence the computer is incapable of reporting the atomic steps of it computations. Is there a similar limit to the depth of analysis in the human protocol? Here again, any answer to this question presupposes a decision regarding the way in which the protocol statements are produced in the brain, but the fact that we can ask the question disarms the critic on this point. It is tempting to suppose that when the subject finds on introspection that addition of single digits or some other simple operation is quite unanalyzable for him, an atomic operation lacking introspectible parts, so to speak, he has reached the limit of analysis imposed on him by the "language" in which he is programmed for these particular tasks. On this view, it would not follow that addition of digits really is, for human beings, what it is introspected to be: a basic, unanalyzable operation. And hence it would not follow on the basis of a comparison of machine trace and protocol statement that a digital computer, for which we know addition of digits is a complex, analyzable operation, operates differently from a human being. The human printout capacity in this case just might not go as deep as the computer's. This defense of the simulator's position against the critic requires opting for a specific view of the operation of the brain, a view for which there is at present scant evidence. And until this specific analogy between machine trace and protocol statement is borne out by evidence, not only do discrepancies of the sort described not count against the simulator's claims, but similarities between machine trace and protocol statement do not count for his claims.

A more glaring example of shaky arguments about the use of protocols is the case of "intuition." Intuition is often contrasted by the workers in the field and their critics to brute force methods of solution, and the simulators are somewhat in the dark about how one could even begin to build "genuine" intuition into a program. (Here "genuine intuition" would be contrasted to the "mere appearance of intuition" which might be produced by the imaginative use of algorithms and heuristics in devising programs.) Workers in the clutches of this way of looking at things tend to see a subject's

protocol to the effect that he "just caught on in a flash" as a stymying indication, but such a protocol could be a case of "printout" in a language far removed from the actual operations. A quixotic but illuminating exercise would be to program a computer to solve certain problems without providing any printout capacity, except for the standard phrase accompanying each solution: "It just came to me, that's all." Intuition, after all, is not a species of deduction or induction; to speak of intuition is to deny that one knows how one arrived at the answer, and the truth of this denial is compatible with one's having arrived at the answer by any method at all, including "unconscious" brute force computing. "Intuition" is not the name of a known or recognized means of processing information, and we can fairly safely assume that all possible means of processing information are now known to programmers and computer theoreticians; new methods will be no more than ingenious combinations of old methods, and whatever new methods are arrived at, none of them will be discovered to have the characteristic hallmarks of human intuition, just because intuition has no hallmarks. The same can be said for the term "insight," long a bone of contention in this area and other areas in psychology.

Another problem facing the simulators reveals a somewhat different difficulty with the use of protocols. The introspective evidence, both from the protocols of subjects and from the introspection of the investigators themselves, suggests that human beings in many cognitive activities "ignore" certain possibilities or certain data. In discussing simulation experiments in pattern recognition, Dreyfus (1965), a recent critic of simulation, says that in people, unlike computers, "noise is not tested and excluded; it is ignored as inessential" (p. 38); and in discussing language translation by computer he points out that inappropriate parsings of technically ambiguous sentences are "ignored" by human readers. He also concludes that "Game playing [reveals] the necessity of processing information which is not explicitly considered or rejected, i.e., information on the fringes of consciousness." These descriptions of human information processing verge on the self-contradictory. How does one ignore something? Ignoring, like intuiting, is not a type of information processing unless it is just what Dreyfus says it is not—a process of testing and excluding. One cannot ignore what is not there, and hence any information that is ignored has arrived and has been excluded on the basis of some test, however rudimentary or even inappropriate. Of course a particular possibility may never "occur" to a person, or a computer, but in this case it would be misleading to say that the possibility was ignored. If ignoring is doing something, then it is a process of excluding; if ignoring is failing to do something, it is well within the capacities of both men and machines, and for that matter within the capacities of anything whatever.

What is the distinction Dreyfus is looking for? I suggest that what he means is that we do not consciously test and exclude this information, which may only mean that no account of this operation can occur in the protocol because of the way the brain is "programmed."

The relation between what information is processed in us and what our protocol statements report is still more tenuous. Consider the following brace of examples. In case A, I walk into the kitchen, pick up an apple, and bite into it. When asked why, I remark with surprise "Oh! I wasn't really aware that I had picked up the apple at all. I don't know why I did it." In case B, I walk into the kitchen, see the apple, say to myself: "That is a nice apple I have there, and it won't spoil my supper, and I like apples, so I think I'll just pick it up and eat it." Here, when I am asked about my action I have quite an elaborate protocol to present. But in both cases it seems quite likely that approximately the same information processing went on, including a great deal that did not enter into my protocol in case B. In both cases I would not have picked up the apple had I been in someone else's house, nor would I have bitten into a raw egg, nor would I have eaten the apple had I known it was time for dinner. It follows that either the appropriateness of my behavior is an immense coincidence, or a great deal of information must have been processed of which I can give

no account in the protocol, for example, that apples are not poisonous, that it is socially acceptable to eat apples before dark, and so on virtually ad infinitum. In fact, if one were to interpret this activity as an example of information processing by the human brain, and write out in sentences the content of the information processed, the list of information might well be endless. This is not to say that all this information need have been processed at this moment, but that earlier processing has prepared me for the appropriate processing I now perform.

Suppose that in case A, I attempt to give a protocol anyway, in spite of the fact that I was not even aware of picking up the apple. I might say that I had realized it was my apple, and so on, but here the status of my protocol would be clearly that of speculation about inner operations, which is far from the status of a machine trace. How could the simulators exclude this sort of speculation in their experiments with problem solving? By telling the subject not to speculate? Is this an order the subject will know how to obey? Or can the simulators cull out the speculation after the fact? This would require independent knowledge about human proclivities to speculate or behavioral manifestations of speculation, and this is clearly lacking.

As if these difficulties were not enough, it also seems prudent to withhold judgment about the common-sense and Freud-founded views of self-deception. Perhaps the protocol I give sincerely and with no intention to speculate is nothing but rationalization; perhaps in case B what actually leads to my action and controls it is the processing of information to the effect that the apple represents my father, and eating it will satisfy a hidden desire to destroy him. A printout provision with the capacity to rationalize would occupy a substantially different place in a program or model than do the provisions devised so far, and if it is supposed that we can rationalized our decisions concerning everyday affairs, why not our decisions concerning artificial puzzles?

The foregoing considerations require a reevaluation of the practice of comparing protocol statements with machine traces. The initial presumption or hope behind this use of comparisons is that the human protocol gives us valuable clues as to how human beings process information. Provided protocols are used only as a source of clues, no difficulties arise and impressive gains in the generation of hypotheses concerning brain function can be expected, for there can be little doubt that there are things to be learned using protocols. It would be a strange world if there were no relation between protocols and actual problem solving. It is only when such comparisons are made to take on the burden of directly supporting a theory that the difficulties arise. For such a move must involve the presupposition that a person's ability to say what he is doing is strongly analogous to the printout capacity of some programs, and while this is a very attractive suggestion, it needs a great deal of backing up, and there are some troubling counterindications. Human protocol statements exhibit variations in depth of analysis, the interpolation of nonessential "let me think's" and "What was I saying's," and the use of such terms as "ignore" and "intuit," which have their own strange logic. Also, it seems that often our introspective accounts are speculative or rationalizing, and we ourselves cannot be sure when. These phenomena suggest that protocol statements are at a rather greater remove from fundamental operations with information than are machine traces.

Newell and Simon present a relatively cautious account of what should be looked for in comparisons:

> Hence, we should expect to find every feature of the protocol that concerns the task mirrored in an essential way in the program trace. The converse is not true, since many things concerning the task surely occurred without the subject's commenting on them (or even being aware of them). (1961, p. 288.)

Even this will not do. We have seen that there are good reasons for ignoring as uninformative or misleading such protocol statements as "It just came to me. I know I ignored the left hand side of the board." And how are these to be distinguished from informative features about the task? The only elements in protocols that turn out to be reliably informative are those that reasonably well mirror the accounts of operations known to computer science. If

what a person says does not jibe with the sorts of things the computer can say, what is said can be taken to be an expression not about the actual information processing at all. Here the model and the modeled have changed places, and human beings can be said to be processing information only to the extent that they do the sorts of things computers do! This is not absurd because human beings are the paradigm information processors, for in fact computers are better understood in this capacity than human beings are. The absurdity lies in worrying about the discrepancy between machine trace and protocol. Of course human beings are information processors, and eventually we will find out how, but not by slavishly trying to reproduce protocol statements. Where discrepancies do not count against you, verisimilitude is a dubious goal.

The idea that causes the trouble is the idea that models of human information processing can be proved or disproved to be adequate on the grounds of strict comparison of protocol statement and machine trace. This notion is attractive, no doubt, because it allows the simulators to accept the challenge of the critics who argue a priori that machines cannot be minds. The difficulties raised here suggest that the simulators should not be seduced into premature battle on this point, and should resist the desire for an immediate acid test in the form of a confirmation vindicating their efforts. Any defense of their efforts that depends on such comparisons as proofs of successful simulation involves suppositions that cannot yet be backed up. The value of research in simulation would not be diminished by a less ambitious interpretation of goals and results, nor would reinterpretation effect any real changes in experimental method. The distinction between clues and data does not arise at the level of experimentation, but only when one is defending an inflated interpretation of results. If the goal of verisimilitude is dropped, there will no longer remain a clear difference in burden of proof between the simulators of cognitive processes and the investigators in artificial intelligence, so this distinction might well lapse. At most, there would be a difference in the emphasis put on the use of the available clues in protocol statements.

REFERENCES

Dreyfus, H. L. *Alchemy and artificial intelligence.* Santa Monica: the RAND Corporation, 1965.

Feigenbaum, E. A., & Feldman, J. *Computers and thought.* New York: McGraw-Hill, 1963.

Newell, A., & Simon, H. A. GPS, a program that simulates human thought. In H. Billing (Ed.), *Lernende automaten.* Munich: Oldenbourg, 1961. Reprinted in E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill, 1963, Pp. 279–292.

Putnam, Hilary. Minds and machines. In Sidney Hook (Ed.), *Dimensions of mind.* New York: New York University Press, 1960, Pp. 148–179.

Taylor, Charles. *The explanation of behaviour.* New York: Humanities Press, 1964.

(Manuscript received Aug. 29, 1966)

Human history becomes more and more a race between education and catastrophe.

H. G. WELLS