Why the Law of Effect will not Go Away*

D. C. DENNETT

The poet Paul Valéry said: 'It takes two to invent anything.' He was not referring to collaborative partnerships between people but to a bifurcation in the individual inventor. 'The one', he says, 'makes up combinations; the other one chooses, recognizes what he wishes and what is important to him in the mass of the things which the former has imparted to him. What we call genius is much less the work of the first one than the readiness of the second one to grasp the value of what has been laid before him and to choose it.'[1] This is a plausible claim. Why? Is it true? If it is, what kind of truth is it? An empirical generalization for which there is wide scale confirmation? Or a 'conceptual truth' derivable from our concept of invention? Or something else?

Herbert Simon, in *The Sciences of the Artificial*, makes a related claim: 'human problem solving, from the most blundering to the most insightful, involves nothing more than varying mixtures of trial and error and selectivity.'[2] This claim is also plausible, I think, but less so. Simon presents it as if it were the conclusion of an inductive investigation, but *that*, I think, is not plausible at all. An extensive survey of human problem solving may have driven home this thesis to Simon, but its claim to our assent comes from a different quarter.

I want to show that these claims owe their plausibility to the fact that they are implications of an abstract principle whose 'necessity' (such as it is) consists in this: we can know independently of empirical research in psychology that any adequate and complete psychological theory must exploit some version or other of the principle. The most familiar version of the principle I have in mind is the derided darling of the behaviorists: the

* Read to the first meeting of the Society for the Philosophy of Psychology, October 26, 1974, at M.I.T.

[1] Quoted by Jacques Hadamard, in *The Psychology of Inventing in the Mathematical Field*, Princeton University Press, 1949, p. 30.

[2] Herbert Simon, *The Sciences of the Artificial*, M.I.T., p. 97.

Law of Effect. 'The rough idea', Broadbent observes,[1] 'that actions followed by reward are repeated, is one which is likely to occur to most intelligent people who think about possible explanations of behavior.' This rough idea, refined, is the Law of Effect, and my claim is that it is not just part of *a* possible explanation of behaviour, but of *any* possible adequate explanation of behaviour.

In order to establish this condition of adequacy for psychological theories, we must first be clear about the burden of psychology. Consider the way the rest of the social sciences depend on the more basic science of psychology. Economics, or at any rate classical economics, assumes at the outset an ontology of rational, self-interested agents, and then proposes to discover generalizations about how such agents, the 'atoms' of economics, will behave in the market-place. This assumption of intelligence and self-interest in agents is not idle; it is needed to ground and explain the generalizations. Consider the law of supply and demand. There is no mystery about why the law holds as reliably as it does: *people are not fools*; they want as much as they can get, they know what they want and how much they want it, and they know enough to charge what the market will bear and buy as cheap as they can. If that didn't explain why the law of supply and demand works, we would be utterly baffled or incredulous on learning that it did. Political science, sociology, anthropology and social psychology are similarly content to *assume* capacities of discrimination, perception, reason and action based on reason and then seek interesting generalizations about the exploitation of these capacities in particular circumstances. One way of alluding to this shared feature of these social sciences is to note that they are all *Intentional*: they utilize the Intentional or 'mentalistic' or 'cognitive' vocabulary—they speak of belief, desire, expectation, recognition, action, etc.—and they permit explanations to come to an end, at least on occasion, with the citation of a stretch of practical reasoning (usually drastically enthymematic). The voters elected the Democrat *because* they were working men and believed the Republican candidate to be anti-labour; the stock market dropped *because* investors believed other havens for their money were safer. These sciences leave to psychology the task of explaining *how there come to be* entities—organisms, human beings—that can be so usefully assumed to be self-interested, knowledgeable and rational. A fundamental task of psychology then is to explain intelligence. For the super-abstemious behaviourist who will not permit himself to speak even of intelligence (that being too 'mentalistic' for him) we can say, with Hull, that a primary task of psychology 'is to understand . . . why . . . behavior . . . is so generally

[1] D. E. Broadbent, *Behaviour*, 1961 (University Paperbacks edn., p. 75).

adaptive, i.e., successful in the sense of reducing needs and facilitating survival . . .'.[1] The account of intelligence required of psychology must not of course be question-begging. It must not explain intelligence in terms of intelligence, for instance by assigning responsibility for the existence of intelligence in creatures to the munificence of an intelligent Creator, or by putting clever homunculi at the control panels of the nervous system.[2] If that were the best psychology could do, then psychology could not do the job assigned it.

We already have a model of a theory that admirably discharges just this *sort* of burden in the Darwinian theory of evolution by natural selection, and as many commentators have pointed out, the Law of Effect is closely analogous to the principle of natural selection. The Law of Effect presumes there to be a 'population' of stimulus-response pairs, more or less randomly or in any case arbitrarily mated, and from this large and varied pool reinforcers *select* the well-designed, the adaptive, the fortuitously appropriate pairs in an entirely mechanical way: their recurrence is made more probable, while their maladaptive or merely neutral brethren suffer 'extinction', not by being *killed* (all particular stimulus-response pairs come to swift ends), but by *failing to reproduce*. The analogy is very strong, very satisfying, and very familiar.

But there has been some misinterpretation of the nature of its appeal. Broadbent observes that

> The attraction both of natural selection and of the Law of Effect, to certain types of mind, is that they do not call on explanatory principles of a quite separate order from those used in the physical sciences. It is not surprising therefore that the Law of Effect had been seized on, not merely as a generalization which is true of animals under certain conditions, but also as a fundamental principle which would explain all adaptive behaviour.[3]

It is certainly true that these analogous principles appeal to physicalists or materialists because they are mechanistically explicable, but there is a more fundamental reason for favouring them: they both can provide clearly non-question-begging accounts of explicanda for which it is very hard to devise non-question-begging accounts. Darwin explains a world of final causes and teleological laws with a principle that is (to be sure) mechanistic but—more fundamentally—utterly independent of 'meaning' or 'purpose'.

[1] Clark Hull, *Principles of Behaviour*, 1943, p. 19.

[2] Cf. B. F. Skinner, 'Behaviorism at Fifty', in T. W. Wann, ed., *Behaviorism and Phenomenology*, 1969, University of Chicago Press, p. 80; and my 'Skinner Skinned' (unpublished).

[3] Broadbent, *op. cit.*, p. 56.

It assumes a world that is *absurd* in the existentialist's sense of the term: not ludicrous but pointless, and this assumption is a necessary condition of any non-question-begging account of *purpose*. Whether we can imagine a *non*-mechanistic but also non-question-begging principle for explaining design in the biological world is doubtful; it is tempting to see the commitment to non-question-begging accounts here as tantamount to a commitment to mechanistic materialism, but the priority of these commitments is clear. It is not that one's prior prejudice in favour of materialism gives one reason to accept Darwin's principle because it is materialistic, but rather that one's prior acknowledgment of the constraint against begging the question gives one *reason to adopt materialism* once one sees that Darwin's non-question-begging account of design or purpose in nature is materialistic. One argues: Darwin's materialistic theory may not be the only non-question-begging theory of these matters, but it is one such theory, and the only one we have found, which is quite a good reason for espousing materialism.

A precisely parallel argument might occur to the psychologist trying to decide whether to throw in with the behaviourists: theories based on the Law of Effect may not be the only psychological theories that do not beg the question of intelligence, but they *are* clearly non-question-begging in this regard, and their rivals are not, which is quite a good reason for joining the austere and demanding brotherhood of behaviourists. But all is not well in that camp, and has not been for some time. Contrary to the claims of the more optimistic apologists, the Law of Effect has not been knit into any theory with anything remotely like the proven power of the theory of natural selection. The Law of Effect has appeared in several guises since Thorndike introduced it as a principle of learning; most influentially, it it assumed centrality in Hull's behaviourism as the 'law of primary reinforcement' and in Skinner's as the 'principle of operant conditioning',[1] but the history of these attempts is the history of ever more sophisticated failures to get the Law of Effect to *do enough work*. It may account for a *lot* of learning, but it can't seem to account for it all. Why, then, not look for another fundamental principle of more power to explain the balance? It is not just mulishness or proprietary pride that has kept behaviourists from following this suggestion, but rather something like the conviction that the Law of Effect is not just *a* good idea, but the only possible good idea for this job. There is something right in this conviction, I want to maintain, but what is wrong in it has had an ironic result: allegiance to the Law of Effect in its behaviouristic or peripheralistic versions has forced psychologists to

[1] Skinner explicitly identifies his principle with the Law of Effect in *Science and Human Behavior*, 1953, p. 87.

beg small questions left and right in order to keep from begging the big question. One 'saves' the Law of Effect from persistent counterinstances by the *ad hoc* postulation of reinforcers and stimulus histories for which one has not the slightest grounds except the demands of the theory. For instance one postulates curiosity drives the reduction of which is reinforcing in order to explain 'latent' learning, or presumes that when one exhibits an apparently novel bit of intelligent behaviour, there *must have been* some 'relevantly similar' responses in one's past for which one was reinforced. These strategies are not altogether bad; they parallel the evolutionist's speculative hypothetical ancestries of species, which are similarly made up out of whole cloth to begin with, but which differ usually in being clearly confirmable or disconfirmable. These criticisms of behaviourism are not new,[1] and not universally fair in application either. I am convinced, nevertheless, that no behaviourism, however sophisticated, can elude all versions of these familiar objections, but that is not a claim to be supported in short compass. It will be more constructive to turn to what I claim is right about the Law of Effect, and to suggest another way a version of it can be introduced to take up where behaviourism leaves off.

The first thing to note is that the Law of Effect and the principle of natural selection are not just analogues; they are designed to work together. There is a kind of intelligence, or pseudo-intelligence, for which the principle of natural selection itself provides the complete explanation, and that is the 'intelligence' manifest in tropistic, 'instinctual' behaviour control. The environmental appropriateness, the biological and strategic wisdom, evident in bird's-nest-building, spider-web-making and less intricate 'innate' behavioural dispositions is to be explained by the same principle that explains the well-designedness of the bird's wings or the spider's eyes. We are to understand that creatures so 'wired' as to exhibit useful tropistic behaviour in their environmental niches will have a survival advantage over creatures not so wired, and hence will gradually be selected by the vicissitudes of nature. Tropistic behaviour is not plastic in the individual, however, and it is evident that solely tropistically controlled creatures would not be evolution's final solution to the needs-versus-environment problem. *If* creatures with some plasticity in their input-output relations were to appear, *some* of them might have an advantage over even the most sophisticated of their tropistic cousins. Which ones? Those that are able to distinguish good results of plasticity from bad, and preserve

[1] Cf., e.g., Charles Taylor, *The Explanation of Behavior*, 1964, Chomsky's reviews of Skinner's *Verbal Behavior*, and *Beyond Freedom and Dignity*, Broadbent, *op. cit.*

the good. The problem of selection reappears and points to its own solution: let some class of events in the organisms be genetically endowed with the capacity to increase the likelihood of the recurrence of behaviour-controlling events upon which they act. Call them reinforcers. Some mutations, we can then speculate, appeared with inappropriate reinforcers, others with neutral reinforcers, and a lucky few with appropriate reinforcers. Those lucky few survive, of course, and their progeny are endowed genetically with a capacity to *learn*, where learning is understood to be nothing more than a change (in the environmentally appropriate direction) in stimulus-response probability relations. The obviously adaptive positive reinforcers will be events normally caused by the presence of food or water, by sexual contact, and by bodily well-being, while the normal effects of injury and deprivation will be the obvious negative reinforcers, though there could be many more than these.[1]

The picture so far is of creatures well endowed by natural selection with tropistic *hard-wiring*, including the hard-wiring of some reinforcers. These reinforcers, in turn, permit the further selection and establishment of adaptive soft-wiring, such selection to be drawn from a pool of essentially arbitrary, *undesigned* temporary interconnections. Whenever a creature is fortunate enough to have one of its interconnections be followed by an environmental effect that in turn produces a reinforcer as 'feedback', that interconnection will be favoured. Skinner is quite explicit about all this. In *Science and Human Behavior* he notes that 'The process of conditioning has survival value', but of course what he means is that the *capacity* to be conditioned has survival value. 'Where inherited behaviour leaves off, the inherited modifiability of the process of conditioning takes over.'[2] So let us use the term 'Skinnerian creatures' for all creatures that are susceptible to operant conditioning, all creatures whose learning can be explained by the Law of Effect. Skinnerian creatures clearly have it over merely tropistic creatures, but it seems that there are other creatures, e.g., at least ourselves and many other mammals, that have it over merely Skinnerian creatures.

The trouble, intuitively, with Skinnerian creatures is that they can learn only by actual behavioural trial and error in the environment. A good bit of

[1] Cf. Skinner, *Science and Human Behavior*, p. 83. Skinner speaks of food and water *themselves* being the reinforcers, but commenting on this difference would entail entering the familiar and arid 'more peripheral than thou' controversy. A point of Skinner's that is always worth reiterating, though, is that negative reinforcers are not *punishments*; they are events the cessation of which is positively reinforcing, that is, their cessation *increases* the probability of recurrence of the behaviour followed by cessation.

[2] *Science and Human Behavior*, p. 55.

soft-wiring cannot get selected until it has had an opportunity to provoke some reinforcing feedback from the environment, and the problem seems to be that merely *potential*, as yet *unutilized* behavioural controls can *ex hypothesi* have no environmental effects which could lead to their being reinforced. And yet experience seems to show that we, and even monkeys, often think out and select an adaptive course of action without benefit of prior external feedback and reinforcement. Faced with this dilemma, we might indulge in a little wishful thinking: if only the Law of Effect could provide for the reinforcement of merely potential, unutilized bits of behaviour control wiring! If only such unutilized controls could have some subtle effect on the environment (i.e., if only merely 'thinking about the solution' could have some environmental effect) and if only the environment were benign enough to bounce back the appropriate feedback in response! But that, it seems, would be miraculous.

Not so. We can have all that and more by simply positing that creatures have *two* environments, the outer environment in which they live, and an 'inner' environment they carry around with them. The inner environment is just to be conceived as an input-output box for providing 'feedback' for events in the brain.[1] Now we can run just the same speculative argument on Skinnerian creatures that we earlier ran on tropistic creatures. Suppose there appear among the Skinnerian creatures of the world mutations that have inner environments of the sort just mentioned. Some, we can assume, will have maladaptive inner environments (the environments will make environmentally inappropriate behaviour more likely); others will have neutral inner environments; but a lucky few will have inner environments that happen to reinforce, by and large, only adaptive *potential* behavioural controls. In a way we are turning the principle of natural selection on its head: we are talking of the evolution of (inner) environments to suit the organism, of environments that would have survival value in an organism. Mutations equipped with such benign inner environments would have a distinct survival advantage over merely Skinnerian creatures in any exiguous environment, since they could learn faster and *more safely* (for trial and error learning is not only tedious; it can be dangerous). The advantage provided by such a benign inner environment has been elegantly expressed in a phrase of Karl Popper's: it 'permits our hypotheses to die in our stead'.

[1] This is not Simon's distinction between inner and outer environment in *The Sciences of the Artificial*, but a more restrictive notion. It also has *nothing whatever* to do with any distinction between the 'subjective' or 'phenomenal' world and the objective, public world.

The behaviourist, faced with the shortcomings of the Law of Effect, insisted that all we needed was more of the same (that only more of the same could explain what had to be explained), and that is what we have given him. He was just construing 'the same' too narrowly. The *peripheralism* of behaviourist versions of the Law of Effect turns out to be not so essential as they had thought. For instance, our talk of an inner *environment* is merely vestigal peripheralism; the inner environment is just an inner something that selects. Ultimately of course it is environmental effects that are the measure of adaptivity and the mainspring of learning, but the environment can delegate its selective function to something in the organism (just as death had earlier delegated its selective function to pain), and if this occurs, a more intelligent, flexible, organism is the result.

It might be asked if behaviourists haven't already, in fact long ago, taken this step to inner reinforcement or selection. I think the fairest answer is that some have and some haven't, and even those that have have not been clear about what they are doing. On the one hand there are the neo-Skinnerians who have no qualms about talking about the operant conditioning that results in the subject who *imagines* courses of action followed by reinforcing results, and on the other hand you have the neo-Skinnerians that still rail against the use of such mentalistic terms as 'imagine'. Skinner himself falls into both camps, often within the compass of a single page.[1] 'The skin', says Skinner, 'is not that important as a boundary'[2] but it is hard to believe he sees the implications of this observation. In any event it will be clearer here to suppose that behaviourists are 'classical' periphalists who do not envisage such a reapplication of the Law of Effect via an inner environment.

At this point it is important to ask whether this proposed principle of selection by inner environment hasn't smuggled in some incoherency or impossibility, for if it has not, we can argue that since our hypothesized mutations would clearly have the edge over merely Skinnerian creatures, there is no reason to believe that operant conditioning was evolution's final solution to the learning or intelligence problem, and we could then safely 'predict' the appearance and establishment of such mutations. Here we are, we could add. We could then go on to ask how powerful our new principle was, and whether there was learning or intelligence *it* couldn't explain. And we could afford to be more open-minded about this question than the behaviourist was, since if we thought there *was* learning it couldn't

[1] See my 'Skinner Skinned' for detailed support of this and similar vacillation in Skinner.

[2] 'Behaviorism at Fifty', in Wann, p. 84.

handle, we'd know where to look for yet a stronger principle; yet a *fourth* incarnation of our basic principle of natural selection (or, otherwise viewed, yet a *third* incarnation of our basic psychological principle of the Law of Effect). In fact we can already see just what it will be. Nothing requires the inner environment to be entirely genetically hard-wired. A more versatile capacity would be one in which the inner environment *itself* could evolve in the individual as a result of—for starters—operant conditioning. We not only learn; we learn better how to learn, and learn better how to learn better how to learn.[1]

So is there anything incoherent about the supposition of inner environments that can select adaptive features of *potential* behaviour control systems (and favour their incorporation into *actual* behaviour controls—for that is what reinforcement amounts to in this application)? Is anything miraculous or question-begging being assumed here? The notion of an inner environment was *introduced* in explicitly non-Intentional language: the inner environment is simply any internal region that can affect and be affected by features of potential behavioural control systems. The benign and hence selected inner environments are simply those in which the result of these causal interactions is the increased conditional probability of the actualization of those potential controls that would be adaptive under the conditions in which they are probable. The way the notion is introduced is thus uncontaminated by covert appeal to intelligence, but it is still not obvious that an inner environment could 'work'.

What conditions must we put on features of bits of brain design to ensure that their selection by an optimally designed selector-mechanism would yield a better than chance improvement in ultimate performance? Since selection by inner environment is ultimately a mechanical sorting, which can key only on physical features of what is sorted, at the very least there would have to be a *normal* or *systematic* correlation between the physical event types selected and what we may call a *functional role* in some control program. A physically characterized type of wiring could not consist in the main of reliably adaptive tokens unless those tokens normally played a particular function.[2] This is the same condition, raised one level,

[1] At a glance it seems that ultimately we want one-shot learning to change the inner environment. In ordinary perspective, we want to account for the fact that if I am trying to solve a problem, *someone can tell me*, once, what won't work and I can take this lesson to heart immediately.

[2] See Simon, *op. cit.*, p. 73, also pp. 90–2. He argues that *efficient* evolution of design also requires a hierarchical organization of design elements. My treatment of these issues is (obviously) heavily indebted to Simon's illuminating and lucid account.

that we find on operant conditioning: if physically characterized *response* classes do not produce a normally uniform environmental effect, reinforcement cannot be adaptive. So if and when this principle works, it works to establish high probabilities that particular appropriate functional roles will be filled at the appropriate times in control programs. Functional roles will be *discriminated*, and thereby control programs will become well designed.

It is hard to keep track of these purported functions and effects while speaking in the sterilized vocabulary of the behaviourist, but there is an easier way of talking: we can say that physical event tokens of a selected type have—in virtue of their normally playing a certain role in a well-designed functional organization—a *meaning* or *content*. We have many familiar examples of *adaptive potential behaviour control elements*: accurate *maps* are adaptive potential behaviour control elements, and so are true *beliefs*, warranted *expectations*, clear *concepts*, well-ordered *preferences*, sound *plans of action*, in short all the favourite tools of the cognitivist psychologist. As Popper says, it is *hypotheses*—events or states endowed with an Intentional characterization—that die in our stead. Is *cognitivist* psychology then bound ultimately to versions of the Law of Effect? That it is, I hope to show by looking at artificial intelligence (AI) research.

AI program designers work backwards on the same task behaviourists work forwards on. We have just traced the behaviourists' cautious and self-denying efforts to build from mechanistic principles towards the levels of complexity at which it becomes apt and illuminating to speak in Intentional terms about what they claim is going on. The AI researcher *starts* with an Intentionally characterized problem (e.g., how can I get a computer to *understand* questions of English?), breaks it down into sub-problems that are also Intentionally characterized (e.g., how do I get the computer to *recognize* questions, *distinguish* subjects from predicates, *ignore* irrelevant parsings?) and then breaks these problems down still further until finally he reaches problem or task descriptions that are obviously mechanistic. Here is a way of looking at the process. The AI programmer begins with an Intentionally characterized problem, and thus frankly views the computer anthropomorphically: if he *solves* the problem he will say he has designed a computer that can understand questions in English. His first and highest level of design breaks the computer down into subsystems, each of which is given Intentionally characterized tasks; he composes a flow chart of evaluators, rememberers, discriminators, overseers and the like. These are *homunculi* with a vengeance; the highest level design breaks the computer down into a committee or army of intelligent homunculi with purposes,

information and strategies. Each homunculus in turn is analysed into *smaller* homunculi, but more important into *less clever* homunculi. When the level is reached where the homunculi are no more than adders and subtractors, by the time they need only the intelligence to pick the larger of two numbers when directed to, they have been reduced to functionaries 'who can be replaced by a machine'. The aid to comprehension of anthropomorphizing the elements just about lapses at this point, and a mechanistic view of the proceedings becomes workable and comprehensible. The AI programmer uses Intentional language fearlessly because he *knows* that if he succeeds in getting his program to run, any questions he has been begging provisionally will have been paid back. The computer is more unforgiving than any human critic; if the program works then we can be certain that all homunculi have been discharged from the theory.[1]

Working backwards in this way has proved to be a remarkably fruitful research strategy, for powerful principles of design have been developed and tested, so it is interesting to note that the overall shape of AI models is strikingly similar to the organization proposed for our post-Skinnerian mutations, and the problems encountered echo the problems faced by the behaviourist. A ubiquitous strategy in AI programming is known as *generate-and-test*, and our opening quotation of Paul Valéry perfectly describes it. The problem solver (or inventor) is broken down at some point or points into a generator and a tester. The generator spews up candidates for solutions or elements of solutions to the problems, and the tester accepts or rejects then on the basis of stored criteria. Simon points out the analogy, once again, to natural selection.[2]

The tester of a generate-and-test subroutine is none other than a part of the inner environment of our post-Skinnerian mutations, so if we want to know how well the principle of selection by inner environment can work, the answer is that it can work as well as generate-and-test methods can work in AI programs, which is hearteningly well.[3] Simon, as we saw at the outset, was prepared to go so far as to conclude that *all* 'human problem solving, from the most blundering to the most insightful' can be captured in

[1] Cf. my 'Intentional Systems', *J. Phil.*, 1971, and 'Why You Can't Make a Computer Feel Pain' (unpublished). In *Content and Consciousness* (1969) I scorned theories that replaced the little man in the brain with a committee. This was a big mistake, for this is just how one gets to 'pay back' the 'intelligence loans' of Intentionalist theories.

[2] Simon, *op. cit.*, pp. 95–8.

[3] Hubert Dreyfus would disagree. (See *What Computers Can't Do: A Critique of Artificial Reason*, Harper & Row, 1973.) But Dreyfus has not succeeded in demonstrating any *a priori* limits to generate-and-test systems hierarchically organized, so his contribution to date is salutary scepticism, not refutation.

the net of generate-and-test programming: 'varying mixtures of trial and error and selectivity'. This claim is exactly analogous to the behaviourists' creed that the Law of Effect could explain all learning, and again we may ask whether this is short-sighted allegiance to an idea that is good, but not the only good idea. Generate-and-test programs can simulate, and hence account for (in one important sense)[1] a lot of problem-solving and invention; what grounds have we for supposing it is powerful enough to handle it all? The behaviourist was in no position to defend his creed, but the AI researcher is in better shape.

Some AI researchers have taken their task to be the *simulation* of particular cognitive capacities 'found in nature'—even the capacities and styles of particular human individuals[2]—and such research is known as CS or 'cognitive simulation' research, but others take their task to be, not simulation, but the construction of intelligent programs *by any means whatever*. The only constraint on design principles in AI thus viewed is that they should *work*, and hence any boundaries the AI programmer keeps running into are arguably boundaries that restrict *all possible* modes of intelligence and learning. Thus if AI is truly the study of all possible modes of intelligence, and if generate-and-test is truly a necessary feature of AI learning programs, then generate-and-test is a necessary feature of all modes of learning, and hence a necessary principle in any adequate psychological theory.

Both premises in that argument need further support. The first premise was proposed on the grounds that AI's guiding principle is that *anything is permitted that works*, but isn't AI really more restrictive than that principle suggests? Isn't it really that AI is the investigation of all possible *mechanistically realizable* modes of intelligence? Doesn't AI's claim to cover all possible modes beg the question against the vitalist or dualist who is looking for a non-question-begging but also non-mechanistic psychology? The AI researcher is a mechanist, to be sure, but a mechanist-*malgré-lui*. He typically does not know or care what the hardware realizations of his designs will be, and often even relinquishes control and authorship of his

[1] There is a tradition of overstating the import of successful AI or CS (cognitive simulation) programs (e.g., 'programs are theories and successful programs are confirmed theories'). For the moment all we need accept is the minimal claim that a successful program proves a particular sort of capacity to be in principle mechanistically realizable and hence mechanistically explicable. Obviously much more can be inferred from successful programs, but it takes some detailed work to say what, where and why.

[2] See, for instance, the computer-copy of a *particular* stock-broker in E. A. Feigenbaum & J. Feldman, eds., *Computers and Thought*, 1963.

programs at a point where they are still replete with Intentionalistic constructions, still several levels away from machine language. He can do this because it is merely a clerical problem for compiler programs and the technicians that feed them to accomplish the ultimate 'reduction' to a mechanistic level. The constraints of mechanism do not loom large for the AI researcher, for he is confident that any design he can state *clearly* can be mechanized. The operative constraint for him, then, is something like clarity, and in practice clarity is ensured for anything expressible in a programming language of some level. Anything thus expressible is clear; what about the converse? Is anything clear thus expressible? The AI programmer believes it, but it is not something subject to proof; it is, or boils down to, some version of Church's thesis (e.g., anything computable is Turing-machine computable). But now we can see that the supposition that there might be a non-question-begging non-mechanistic psychology gets you nothing unless accompanied by the supposition that Church's thesis is false. For a non-question-begging psychology will be a psychology that makes no ultimate appeals to unexplained intelligence, and that condition can be reformulated as the condition that whatever functional parts a psychology breaks its subjects into, the smallest, or most fundamental, or least sophisticated parts must not be supposed to perform tasks or follow procedures requiring intelligence. That condition in turn is surely strong enough to ensure that any procedure admissible as an 'ultimate' procedure in a psychological theory falls well within the intuitive boundaries of the 'computable' or effective' as these terms are presumed to be used in Church's thesis. The intuitively computable functions mentioned in Church's thesis are those that 'any fool can do', while the admissible atomic functions of a psychological theory are those that 'presuppose *no* intelligence'. If Church's thesis is correct then the constraints of mechanism are no more severe than the constraint against begging the question in psychology, for any psychology that stipulated atomic tasks that were 'too difficult' to fall under Church's thesis would be a theory with undischarged homunculi.[1] So our first premise, that AI is the study of all pos-

[1] Note that this does *not* commit the AI researcher to the view that 'men are Turing machines'. The whole point of generate-and-test strategies in program design is to *permit* computers to *hit on* solutions to problems they cannot be *guaranteed* to solve either because we can prove there is no algorithm for getting the solution or because if there is an algorithm we don't know it or couldn't use it. Is there a decision procedure ensuring checkmate in chess? Few think so, and we don't know one way or the other. If there is, it would certainly take astronomically too much time and energy to use. Hence the utility of generate-and-test and heuristics in programming.

sible modes of intelligence, is supported as much as it could be, which is *not quite* total support, in two regards. The first premise depends on two unprovable but very reasonable assumptions: that Church's thesis is true, and that *there can be*, in principle, an adequate and complete psychology.

That leaves the second premise to defend: what reason is there to believe that generate-and-test is a necessary and not merely handy and ubiquitous feature of AI learning programs? First, it must be granted that many computer programs of great sophistication do not invoke any variety of generate-and-test. In these cases the correct or best steps to be taken by the computer are not selected but *given*; the program's procedures are completely designed and inflexible. These programs are the analogues of our merely tropistic creatures; their design is *fixed* by a prior design process. Sometimes there is a sequence of such programs, with the programmer making a series of changes in the program to improve its performance. Such genealogical developments do not so much represent problems solved as problems deferred, however, for the trick is to get the program to become self-designing, 'to get the teacher out of the learner'. As long as the programmer must, in effect, reach in and rewire the control system, the system is not *learning*. Learning can be viewed as *self-design*, and Simon suggests we 'think of the design process as involving first the generation of alternatives and then the testing of these alternatives against a whole array of requirements and constraints'.[1] Of course he would suggest this, and we can follow his suggestion, but are there any alternatives? Is there any way of thinking (coherently) about the design process that is incompatible with (and more powerful than) thinking of it as an evolution wrought by generate-and-test? It seems not, and here is an argument supposed to show why. I suspect this argument could be made to appear more rigorous (while also, perhaps, being revealed to be entirely unoriginal) by recasting it into the technical vocabulary of some version of 'information theory' or 'theory of self-organizing systems'. I would be interested to learn that this was so, but am content to let the argument, which is as intuitive as it is sketchy, rest on its own merits in the meantime.

We are viewing learning as ultimately a *process* of self-design. That process is for the purposes of this argument defined only by its *product*, and the product is a *new* design. That is, as a result of the process something comes to have a design it previously did not have. This new design 'must come from somewhere'. That is, it takes *information* to distinguish the new design from all other designs, and that information must come from somewhere. Either all from outside the system, or all from inside, or a bit of

[1] Simon, *op. cit.*, p. 74.

both. If all from outside, then the system does not redesign itself; this is the case we just looked at, where the all-knowing programmer, who *has* the information, *imposes* the new design on the system from without. So the information must all come from inside, or from both inside and outside. Suppose it all comes from inside. Then either the information already exists inside or it is created inside. What I mean is this: either the new design *exists ready made* in the old design in the sense that its implementation at this time is already guaranteed by its old design, or the old design does not determine in this way what the new design will be. In the former case, the system has not really redesigned itself; it was designed all along to go into this phase at this time, and we must look to a prior design process to explain this. In the latter case, the new design is *underdetermined* by the old design. This is a feature shared with the one remaining possibility: that the information comes from both inside and outside. In both of these cases the new design is underdetermined by the old design by itself, and only in these cases is there 'genuine' learning (as opposed to the merely 'apparent' learning of the merely tropistic creature). In any such case of underdetermination, the new design is either underdetermined period—there is a truly random contribution here; nothing takes up all the slack left by the underdetermination of the old design—*or* the new design is determined by the combination of the old design and contributions (from either inside or outside or both) that are themselves *arbitrary*, that is, *undesigned* or *fortuitous*. But if the contribution of arbitrary elements is to yield a better than chance probability of the new design being an improvement over the old design, the old design must have the capacity to *reject* arbitrary contributions on the basis of design features —information—already present. In other words, there must be a *selection* from the fortuitous contributions, based on the old design. If the arbitrary or undesigned contribution comes from within, what we have is a non-deterministic automaton.[1] A non-deterministic automaton is one such that at some point or points its further operations must wait on the result of a procedure that is undetermined by its program and input. In other words, some tester must wait on some generator to produce a candidate for its inspection. If the undesigned contribution comes from the outside, the situation is much the same; the distinction between *input* and *random contribution* is just differently drawn. The automaton is now deterministic in that its next step is a determinate function of its program and its input,

[1] Cf. see above. Gilbert Harman points out in *Thought*, 1973, that nondeterministic automata can be physically deterministic (if what is random relative to the program is determined in the machine).

but what input it gets is a fortuitous matter. In either case the system can *protect itself* against merely fortuitous response to this merely fortuitous input only by *selecting* as a function of its old design from the fortuitous 'stimulation' presented. Learning must tread the fine line between the idiocy of pre-programmed tropism on the one hand and the idiocy of an over-plastic domination by fortuitous impingements on the other. In short, every process of genuine learning (or invention, which is just a special sort of learning) must invoke, at at least one but probably many levels, the principle of generate-and-test.

The moral of this story is that cognitivist theoreticians of all stamps may proceed merrily and *fruitfully* with temporarily question-begging theoretical formulations, but if they expect AI to *pay their debts* some day (and if anything can, AI can), they must acknowledge that the *processes* invoked will inevitably bear the analogy to natural selection exemplified by the Law of Effect. The moral is *not*, of course, that behaviourism is the road to truth in psychology; even our hypothesized first-generation mutations of Skinnerian creatures were too intelligent for behaviourism to account for, and we have every reason to believe actual higher organisms are much more complicated than that. The only solace for the behaviourist in this account is that his theoretical paralysis has been suffered in a Good Cause; he has not begged the question, and if the high-flying cognitivists ever achieve hie probity it will only be by relying on principles fundamentally analogous to his.

This leaves it open where these inevitable principles of selection will be invoked, and how often. Nothing requires generate-and-test formats to be simple and obviously mechanistic in any of their interesting realizations. On the contrary, *introspective* evidence, of a sort I will presently illustrate, seems to bear out the general claim that generate-and-test is a common and recognizable feature of human problem solving at the same time that it establishes that the generators and testers with which we are *introspectively* familiar are themselves highly sophisticated—highly intelligent homunculi. As Simon points out, generate-and-test is not an efficient or powerful process unless the *generator* is endowed with a high degree of selectivity (so that it generates only the most likely or most plausible candidates in a circumstance), and since, as he says, 'selectivity can always be equated with some kind of feedback of information from the environment' (p. 97), we must ask, of each sort and degree of selectivity in the generator, where did *it* come from—is it learned or innate, and at the end of any successful answer to that question will be a generate-and-test process, either of natural selection if the selectivity is innate, or of some variety of learning,

if it is not. A consequence of this is that we cannot tell by any simple inspection or introspection whether a particular stroke of genius we encounter is a bit of 'genuine' invention at all—that is, whether the invention occurred just *now*, or is the result of much earlier processes of invention that are now playing out their effects. Did Einstein's genetic endowment guarantee his creativity, or did his genetic endowment together with his nurture, his stimulus history, guarantee his creativity or did he genuinely create (during his own thought processes), his great insights? I hope it is clear how little hinges on knowing the answer to this question.

At this point I am prepared to say that the first part of Valéry's claim stands vindicated: it takes two to invent anything: the one makes up combinations; the other one chooses. What of the second part of this claim: 'What we call genius is much less the work of the first one than the readiness of the second one to grasp the value of what has been laid before him and to choose it.'? We have seen a way in which this must be true, in the strained sense that the *ultimate* generators must contain an element of randomness or arbitrariness. 'The original solution to a problem must lie in the category of luck.'[1] But it does not seem that Valéry's second claim is true on any ordinary interpretation. For instance, it does not seem to be true of all *inter-personal* collaborations that the choosers are more the geniuses than their 'idea-men' are. Some producers seldom offer poor suggestions; their choosers are virtual yes-men. Other producers are highly erratic in what they will propose, and require the censorship of severe and intelligent editors. There appears to be a trade-off here between, roughly, spontaneity or fertility of imagination on the one hand, and a critical eye on the other. A task of invention seems to require both, and it looks like a straightforwardly empirical question subject to continuous variation how much of each gets done by each collaborator.

Valéry seems to slight the contribution of the first, but perhaps that is just because he has in mind a collaboration at one end of the spectrum, where a relatively undiscriminating producer of combinations makes a lot of work for his editor. Of course, as said at the outset, Valéry is not talking about actual interpersonal collaboration, but of a bifurcation in the soul. He is perhaps thinking of his own case, which suggests that he is one of those who is *aware* of considering and rejecting many bad ideas. He does not credit *his* producer-homunculus with much genius, and is happy to

[1] Arthur Koestler, in *The Acts of Creation*, 1964, p. 559, quotes the behaviourist E. R. Guthrie to this effect, but it is a misquotation, sad to say, for had Guthrie said what Koestler says he said, he would have said something true and important. Perhaps he did say it, but not on the page, or in the book, where Koestler says he said it.

identify with the *responsible* partner, the chooser. Mozart, it seems, was of the same type: 'When I feel well and in a good humor, or when I am taking a drive or walking after a good meal, or in the night when I cannot sleep, thoughts crowd into my mind as easily as you would wish. Whence and how do they come? I do not know and *I have nothing to do with it*. Those which please me I keep in my head and hum them; at least others have told me that I do so'.[1] In such cases the producer-chooser bifurcation lines up with the unconscious and conscious selves bifurcation. One is conscious only of the *products* of the producer, which one then consciously tests and chooses.

Poincaré, in a famous lecture of 1908, offers an 'introspective' account of some mathematical inventing of his own that is more problematic: 'One evening, contrary to my custom, I drank black coffee and could not sleep. Ideas rose in crowds; I felt them collide until pairs interlocked, so to speak, making a stable combination.'[2] In this instance the chooser seems to have disappeared, but Poincaré has another, better interpretation of the incident. In this introspective experience he has been given a rare opportunity to glimpse the *processes* in the generator; what is normally accomplished out of sight of consciousness is witnessed on this occasion, and the ideas that form stable combinations are those few that would normally be presented to the conscious chooser for further evaluation. Poincaré supposes he has watched the selectivity within the generator at work. I am not a little sceptical about Poincaré's claimed *introspection* here (I think all introspection involves elements of rational reconstruction, and I smell a good deal of that in Poincaré's protocol), but I like his categories. In particular, Poincaré gives us, in his discussion of this experience, the key to another puzzling question.

For I have really had two burdens in this paper. The first, which I take to have discharged, is to explain why the Law of Effect is so popular in its various guises. The other is to explain why it is so *unpopular* in all its guises. There is no denying that the Law of Effect seems to be an affront to our self-esteem, and a lot of the resistance, even hatred encountered by behaviourists is surely due to this. Poincaré puts his finger on it. He was, if anyone ever has been, a creative and original thinker, and yet his own analysis of how he accomplished his inventions seemed to deny him *responsibility* for them. He saw only two alternatives, both disheartening. One was his unconscious self, the generator with whom he does not or cannot *identify* 'is capable of discernment; it has tact, delicacy; it knows how

[1] Quoted in Hadamard, *op. cit.*, p. 16, italics added.
[2] Quoted in Hadamard, *op. cit.*, p. 14.

to choose, to divine. What do I say? It knows better how to divine than the conscious self since it succeeds where that has failed. In a word, is not the subliminal self superior to the conscious self? I confess that, for my part, I should hate to accept this.'[1] The other is that the generator is an automaton, an ultimately absurd, blind trier of all possibilities. That is of course no more a homunculus with whom to identify oneself. One does not want to be the generator, then. As Mozart says of his musical ideas: 'Whence and how do they come? I do not know and I have nothing to do with it.' Nor does one want to be just the tester, for then one's chances of being creative depend on the luck one has with one's collaborator, the generator. The fundamental passivity of the testing role leaves no room for the 'creative self'.[2] But we couldn't have hoped for any other outcome. If we are to have any adequate *analysis* of creativity, invention, intelligence, it must be one in which intelligence is analysed into something none of whose parts is intelligence, and at that level of analysis, of course, no 'self' worth identifying with can survive.

The mistake in this pessimism lies in confusing explaining with explaining away. Giving a non-question-begging account of *how* creatures are intelligent can hardly prove that they aren't intelligent. If we want to catch a glimpse of a creative self, we should look, for instance, at M. Poincaré, for *he* (and not any of his proper parts) was certainly a genius.

Finally, I cannot resist passing on a wonderful bit of incidental intelligence reported by Hadamard: the Latin verb *cogito* is derived, as St. Augustine tell us, from Latin words meaning *to shake together*, while the verb *intelligo* means *to select among*. The Romans, it seems, knew what they were talking about.

Tufts University

[1] Quoted in Koestler, *op. cit.*, p. 164.

[2] This passivity is curiously evoked by Koestler in his account of 'underground games' in *The Act of Creation*. It is a tell-tale sign of the inescapability of the principle of selectivity discussed here that Koestler, the arch-enemy of behaviourism, can do no better, when he sets himself the task of composing a rival account of creativity, than to accept the generate-and-test format and then endow the generator with frankly mysterious effects of uncoincidental coincidence.