Seymour Papert

One AI or Many?

Is there one AI or are there many? A dramatic shift in the tone of discussion about artificial intelligence has brought about a suddenly increased awareness of the presence of divergent ways of thinking in what has generally been presented as a unified field. Readers of this issue of *Dædalus* who have not kept abreast of recent developments may be astonished to see how many of its authors have chosen to focus on divergences in the field, and particularly on one trend in AI that has come to be known as connectionism. They would not be alone in their surprise. Late in 1985 I participated in a planning meeting to discuss an issue of *Dædalus* on AI. At that time I knew (and I assume that most people at that meeting knew) that research activity on "connectionist" themes was growing. But I would have expressed disbelief had anyone at the meeting suggested (no one did) that these themes would soon burst out of the technical journals into such publications as the *New York Times Book Review*—where connectionism is characterized as cognitive counter-revolution[1]—and become the central talking point wherever AI or cognitive science is discussed. The contents of this issue of *Dædalus* reflect this movement more than any deliberate plan: something intriguing and dramatic had taken place on a larger scale than the planning of a journal. So when Stephen Graubard invited me to contribute a piece of my own, I could not resist using the connectionist brouhaha as the occasion to discuss some larger issues about

Seymour Papert is professor of media technology and director of the Learning and Epistemology Group in the Media Laboratory at MIT.

the nature of artificial intelligence and its appeal to people more interested in the human mind than in building robots.

The field of artificial intelligence is currently divided into what seem to be several competing paradigms. The present contenders differ over the form of mechanisms needed to capture all forms of intelligence. They are each engaged in a search for mechanisms with a universal application. Allen Newell, dean of information processing, believes that he is close, that all knowledge can be formulated as the rules behind a special kind of program known as a "production system." The authors of the current connectionist manifesto, *Parallel Distributed Processing,*[2] do not think they are as close, but speak with confidence that their way—relying not on programs but on networked neuronlike entities—will provide universal mechanisms.

I do not foresee the future in terms of an ultimate victory for any of the present contenders. What I do foresee is a change of frame, away from the search for universal mechanisms. I believe that we have much more to learn from studying the differences, rather than the sameness, of kinds of knowing. And just because knowing takes place in one brain is not a reason to argue, as both connectionists and programmers do, that there is one privileged and universal mechanism on any psychologically relevant level.

An analogy dramatizes what I mean by psychologically relevant. An evolutionary biologist might try to understand how tigers came to have stripes. And a molecular biologist might try to understand the origin of life in some primeval soup. But how life started gives you no information about how a tiger looks. Yet this fallacy pervades the intellectual discourse of connectionists and programmers. The connectionists talk about experiments on the level of small groups of simulated neurons and then, almost in the same breath, talk about how one can walk and think at the same time. Multiprocessing is assumed to be the same kind of enterprise in both cases. Information processing experts display rule systems that match the behavior of people and computers solving logical problems, and jump from there to statements like Allen Newell's: "Psychology has arrived at the possibility of a unified theory of cognition."

There is the same mistake on both sides: the category error of supposing that the existence of a common mechanism provides both an explanation and a unification of all systems, however complex, in which this mechanism might play a central role. My thesis here is that

AI needs to be defined in a way that does not put it in jeopardy of making this category error. As it matures, I see AI developing the conceptual frameworks that will enable us to obtain a rigorous understanding not only of what is the same in such activities as falling in love and playing chess, but of what is different between them. Artificial intelligence should become the methodology for thinking about ways of knowing.

In this essay I use an incident in the development of connectionism to illustrate the current resistance of the field to this way of thinking about its intellectual identity.

I

I do not come to the discussion of connectionism as a neutral observer. In fact, the standard version of its history assigns me a role in a romantic story whose fairytale resonances surely contribute at least a little to connectionism's aura of excitement.

Once upon a time two daughter sciences were born to the new science of cybernetics. One sister was natural, with features inherited from the study of the brain, from the way nature does things. The other was artificial, related from the beginning to the use of computers. Each of the sister sciences tried to build models of intelligence, but from very different materials. The natural sister built models (called neural networks) out of mathematically purified neurones. The artificial sister built her models out of computer programs.

In their first bloom of youth the two were equally successful and equally pursued by suitors from other fields of knowledge. They got on very well together. Their relationship changed in the early sixties when a new monarch appeared, one with the largest coffers ever seen in the kingdom of the sciences: Lord DARPA, the Defense Department's Advanced Research Projects Agency. The artificial sister grew jealous and was determined to keep for herself the access to Lord DARPA's research funds. The natural sister would have to be slain.

The bloody work was attempted by two staunch followers of the artificial sister, Marvin Minsky and Seymour Papert, cast in the role of the huntsman sent to slay Snow White and bring back her heart as proof of the deed. Their weapon was not the dagger but the mightier pen, from which came a book—*Perceptrons*[3]—purporting to prove that neural nets could never fill their promise of building models of mind: *only computer programs could do this.* Victory seemed assured for the artificial sister. And indeed, for the next decade all the rewards of the kingdom came to her progeny, of which the family of expert systems did best in fame and fortune.

But Snow White was not dead. What Minsky and Papert had shown the world as proof was not the heart of the princess; it was the heart of a pig. To be more literal: their book was read as proving that the neural net approach to building models of mind was dead. But a closer look reveals that they really demonstrated something much less than this. The book did indeed point out very serious limitations of a certain class of nets (nowadays known as one-layer perceptrons) but was misleading in its suggestion that this class of nets was the heart of connectionism. *Parallel Distributed Processing,* allowing that the suggestion could have been an honest mistake, lapses into a fairy-tale tone in talking about how things were back in "Minsky and Papert's day." In that far-off time and place, the technical discoveries were still to be made that would open the vision—model connectionism's sustaining myth—of much more powerful neural nets than could then be imagined.

Connectionist writings present the story as having a happy ending. The natural sister was quietly nurtured in the laboratories of a few ardent researchers who kept the faith, even when the world at large let itself be convinced that the enterprise was futile. Who (or what) should be cast in the role of Prince Charming is a problem I shall take up later: Who are the parties to the present-day connectionist love affair? Who woke connectionism? And why now? And what next? But for the moment suffice it to note that the princess has emerged from relative rags and obscurity to win the admiration of all except a few of her sister's disgruntled hangers-on.

II

The story seems to call for a plea of guilty or innocent: Did Minsky and I try to kill connectionism, and how do we feel now about its resurrection? Something more complex than a plea is needed. Yes, there was *some* hostility in the energy behind the research reported in *Perceptrons,* and there is *some* degree of annoyance at the way the new movement has developed; part of our drive came, as we quite plainly acknowledged in our book, from the fact that funding and research energy were being dissipated on what still appear to me (since the story of new, powerful network mechanisms is seriously exaggerated) to be misleading attempts to use connectionist methods in practical applications. But most of the motivation for *Perceptrons* came from more fundamental concerns, many of which cut cleanly across the division between networkers and programmers.

One of these concerns had to do with finding an appropriate balance between romanticism and rigor in the pursuit of artificial intelligence. Many serious endeavors would never get off the ground if pioneers were limited to discussing in public only what they could demonstrate rigorously. Think, for example, of the development of flying machines. The excitement generated when the Wright brothers made their first flight had a large element of the romantic. And rightly so: it is hard to work up respect for those critics who complained that a short hop on a beach did not prove the feasibility of useful air transportation. When final success cannot be taken as a criterion for judging initial steps, the problem of developing a sensible critical methodology is an essential and often delicate part of any very out-of-the-ordinary endeavor. In the case of artificial intelligence, the problem of critical judgment of partial results is compounded by the fact that a little intelligence is not easily recognized as intelligence. Indeed, in English we have a special word for it: although a short flight is still counted as a flight, a little intelligence is counted as stupidity, and in AI's early stages (where it still is), this is all that can be expected. How, then, does one decide whether the latest "stupidity" of a machine should be counted as a step toward intelligence? The methodology Minsky and I used in *Perceptrons* is best explained through an example.

Parallel Distributed Processing reports an experiment in which a simulated machine (I'll call it Exor) learned to tell whether two inputs, each of which must be either a one or a zero, are different.* Exor's learning process consumed 2,232 repetitions of a training cycle; in each repetition the machine was presented with one of the four possible combinations of inputs (one-one, zero-zero, zero-one, one-zero) and a feedback signal to indicate whether it had given the right response ("no" for the first two and "yes" for the others). Smart

*XOR, pronounced as if written *exor,* is a computerist abbreviation for "exclusive or" (i.e., "this or that but not both"). This makes it the perfect name for our simulated machine.

or stupid? Should one be more impressed by the fact that the thing "learned" at all, or by the fact that it learned so slowly and laboriously?

There was a time, in the early days of cybernetics, when a machine doing anything at all that resembled learning would have been impressive. Today something more is needed to give significance, and in this case the something more is closely related to our allegory. Exor is a neural net, and the task it learned to perform happens, for all its simplicity, to be one of those things a one-layer net cannot do. Knowing this turns the dilemma of judging Exor into an encapsulation of the larger dilemma of judging connectionism. If you want to believe, Exor allows you to proclaim, "Snow White lives." If you don't, Exor's retarded pace of learning allows you to whisper, "But barely." *Perceptrons* set out on a very different tack: instead of asking whether nets are good, we asked what they are *good for*. The focus of enquiry shifted from generalities about kinds of machines to specifics about kinds of tasks. From this point of view, Exor raises such questions as: Which tasks would be learned faster and which would be learned even more slowly by this machine? Can we make a theory of tasks that will explain why 2,232 repetitions were needed in this particular act of learning? The shift in perspective is sharp: interest has moved from making a judgment of the machine to using the performance of the machine on particular tasks as a way to learn more about the nature of the tasks. This shift is reflected in the subtitle of our book—*Perceptrons: An Introduction to Computational Geometry*. We approached our study of neural networks by looking carefully at the kinds of tasks for which their use was being advocated at the time. Since most of these were in the area of visual pattern recognition, our methodology led us into building theories about such patterns. To our surprise, we found ourselves working a new problem area for geometric research, concerned with understanding why some recognition tasks could easily be performed by a given recognition mechanism, while other computations were extremely costly as measured by the number of repetitions needed for a task or the amount of machinery required. For example, a small single-layer perceptron can easily distinguish triangles from squares, but a very large network is needed to learn whether what is put in front of it is a single connected object or is made up of several parts.

Our surprise at finding ourselves working in geometry was a pleasant one. It reinforced our sense that we were opening a new field, not closing an old one. But although the shift from judging perceptrons abstractly to judging the tasks they perform might seem like plain common sense, it took us a long time to make it. So long, in fact, that we are now only mildly surprised to observe the resistance today's connectionists show to recognizing the nature of our work—and the nature of the problem area into which their own investigations must eventually lead.

III

Artificial intelligence, like any other scientific enterprise, had built a scientific culture. The way of working we used in *Perceptrons* ran against the grain of this culture, in whose development we ourselves had participated.

The quest for universality of mechanism is obscured as a pervasive trait of the AI culture by the circumstance that all successful AI demonstrations, whether by programmers or connectionists, perform quite specific tasks in quite narrow domains. Indeed, AI theorists sometimes claim as an important discovery the theory that domain specificity is not a limitation of machines but a characteristic of intelligence. However, the theoretical energy of AI has not gone into understanding differences between specific domains, but rather into finding general forms for the specific contents.

The universalist trait gains robustness from having numerous roots. Among the deepest may be the mythic nature of AI's original enterprise of mind building mind. The desire for universality was fed also by the legacy of the scientists, largely mathematicians, who created AI. And it was nurtured by the most mundane material circumstances of funding. By 1969, the date of the publication of *Perceptrons,* AI was not operating in an ivory-tower vacuum. Money was at stake. And while this pressured the field into a preference for short-term achievement, it also put a premium on claims that the sponsor's investment would bear fruits beyond the immediate product.

Its universalism made it almost inevitable for AI to appropriate our work as proof that neural nets were universally bad. We did not think of our work as killing Snow White; we saw it as a way to understand her. In fact, more than half of our book is devoted to "properceptron" findings about some very surprising and hitherto unknown things that perceptrons can do. But in a culture set up for global judgment of mechanisms, being understood can be a fate as bad as death. A real understanding of what a mechanism can do carries too much implication about what it can not do.

The same trait of universalism leads the new generation of connectionists to assess their own microlevel experiments, such as Exor, as a projective screen for looking at the largest macroissues in the philosophy of mind. The category error analogous to seeking explanations of the tiger's stripes in the structure of DNA is not an isolated error. It is solidly rooted in AI's culture.

IV

The conceit of using the story of Snow White as a metaphor has allowed me to talk about the connectionist counterrevolution without saying exactly what connectionism is or what it is revolting against. A little more technical detail is needed to situate connectionism in the larger field of sciences of mind.

The actual task of recognizing the sameness of the two binary inputs would be a trivial one for a programmer. The first of several remarkable features possessed by Exor is that no one programmed it; it was "trained" to do its task by a strictly behaviorist process of external association of stimuli with reinforcements. It could have been trained by someone who rigorously followed Watson's strictures against thinking about the innards of a system. But if this was its only merit as a model of mental process, the large number of repetitions would negate its interest: machines specifically designed to simulate conditioned reflexes have done so with a psychologically more plausible number of repetitions.

Exor's claim of universality is a stronger feature. Exor is small and limited in power, but it sustains the vision of larger machines that are built on the same principles and that will learn whatever is learnable with no innate disposition to acquire particular behaviors. The prospect of such performance becomes a vindication of something more than neural nets. It promises a vindication of behaviorism against Jean Piaget, Noam Chomsky, and all those students of mind who criticized the universalism inherent in behaviorism's tabula rasa.

Behaviorism has been beaten down in another version of the Snow White story, but the response of academic psychology to connectionism may turn out to be a classic example of the return of the repressed.

Connectionism does more than bring back old-fashioned behaviorism. It brings it back in a form that offers a reconciliation with biological thinking about the brain. The structure of the machine reflects, albeit in an abstract way, a certain model of how brains might conceivably be built out of neurons. Although the actual Exor experiments are, of course, performed by computer programs, these programs are meant to represent what would happen if one connected together networks of units that are held to be neuronlike in the following sense. Each unit in the network receives signals from the others or from sensor units connected to the outside world; at any given time, each unit has a certain level of activation that depends on the weighted sum of the states of activation of the units sending signals to it, and the signals sent out along the unit's "axon" reflect its state of activation. Learning takes place by a process that adjusts the weights (strengths of connections) between the units; when the weights are different, activation patterns produced by a given input will be different, and finally, the output (response) to an input (stimulus) will change. This feature gives machines in Exor's family a biological flavor that appeals strongly to the spirit of our times and yet takes very little away from the behaviorist simplicity: although one has to refer to the neuronlike structure in order to build the machine, one thinks only in terms of stimulus, response, and a feedback signal to operate it.

V

This presentation of connectionism as behaviorism in computer's clothing helps place *Perceptrons* in perspective: the questions it discusses are a modern form of an old debate originally couched as a humanistic and philosophical discussion of associations and taken up again more recently as a discussion of behaviorism. Such debates often turn around assertions of the form, "*Starting with nothing but* (associations, stimulus and response, or whatever), *you can never get to* (general ideas, language, or whatever)." Discussion of this form has been more or less compelling but seldom anywhere near conclusive to standards of rigor that seemed normal to people trained, as Minsky and I both were, as mathematicians. And indeed, how could the discussion even be formulated with any semblance of rigor in the absence of a tight theory of human thought? And how could one move seriously toward such a tight theory without knowing whether general ideas or whatever can be derived from associations or whatever?

In its narrowest sense, the intention of *Perceptrons* was to avoid for the study of "machine thinking" some of the chicken-and-egg difficulties that have plagued thinking about human thinking. The strategy was to study a class of computational machines that were sufficiently powerful to capture a significant slice of contemporary achievement in AI, yet sufficiently simple to make possible, with the limited analytic tools at our disposal, a rigorous mathematical analysis of their capacities. We chose the class of machines for which the book was named (in honor of Frank Rosenblatt): perceptrons are defined in the book to be a special and especially simple kind of neural net in the same family as Exor. Perceptrons are too simple to be interesting in their own right as models of mental process. But the most promising step toward developing tools powerful enough to analyze more complex systems, including the human mind, seemed to be achieving a thorough understanding of a single case as simple as a perceptron. Many readers, perhaps all except mathematicians, would be shocked to know how simple a machine can be and still elude full understanding of its capabilities. I find it quite awesome to think about how hard it was to confirm or reject our intuitions about the capacities of perceptrons.

Minsky and I both knew perceptrons extremely well. We had worked on them for many years before our joint project of understanding their limits was conceived; indeed, we originally met at a conference where we both coincidentally presented papers with an unlikely degree of overlap in content about what perceptronlike machines could do. With this background we should have been in an exceptional position to formulate strong conjectures about perceptrons. Yet when we challenged ourselves to prove our intuitions it sometimes took years of struggle to pin one down—to prove it true or to discover that it was seriously flawed.

I was left with a deep respect for the extraordinary difficulty of being sure of what a computational system can or cannot do. I wonder at people who seem so secure in their intuitive convictions, or their less-than-rigorous rhetorical arguments, about computers, neural nets, or human minds. One area in which intuition seems particularly in need of rigorous analysis is in dealing with the romantically attractive notion of holistic process.

VI

In the history of psychology, behaviorism and holism (or gestaltism) have been considered polar opposites. Behaviorism fragments the mind into a myriad of separate atoms of a much smaller size than common sense would allow. Holism and gestaltism insist that psychological atoms are bigger than common sense thinks. So it is quite remarkable that connectionism has facets that appeal to each of these schools of thought.

The title of the current bible of connectionism, *Parallel Distributed Processing*, juxtaposes two qualities that are taken in the connectionist movement as prime characteristics certainly of all natural, and probably of effective artificial, embodiments of intelligence. *Parallel* refers to the quality of having many processes go on at the same time: as people walk and talk at the same time, they very likely carry out large numbers of concurrent, mostly unconscious, mental processes. *Distributed* refers to the quality of not being localized: in traditional computers, items of information are stored in particular places, cleanly separated from one another; in neural nets, information is spread out (in principle, a new piece of learning might involve changes everywhere). Much of the sense that deep process is at work in the functioning of nets is related to the suggestion that what ordinary discourse and traditional cognitive theory misleadingly describe as atomistic items of information are holistically represented and yet appropriately evocable.

Parallel plus distributed *feels* right. But work with perceptrons made us acutely aware of ways in which the two qualities are in tension rather than sweet harmony. It is not hard to switch perceptions so as to make the juxtaposition feel intuitively problematic. In ordinary life, customs of separating activities into rooms and offices are founded on experience with the untidy consequences of having everything happening everywhere at the same time. But connectionism is built on the theory—what Sherry Turkle calls a sustaining myth—that a deeper understanding would reveal the naiveté of such everyday analogies. Just as modern physics teaches us not to project our sense of macroscopic events onto the subatomic world, so too deeper understanding of networks will teach us that our metaphors of macroscopic organization may be equally misleading.

Indeed, one can find analogies in physical science that go very strongly against uninformed intuitions about interference—how processes disturb one another. The vibrations of all radio and television waves pass through the same space at the same time, and yet tuning circuits can separate them. Even more incomprehensible, if not frankly shocking to common sense, is the hologram, which records a three-dimensional picture in a fully distributed way: if part of the holographic record is destroyed, no particular part of the picture is lost; there is only a uniform degradation of quality.

These examples plainly say that there is precedent in the physical world for distributed superposition. Enough in the universe is holistic so that the concept of distributed neural net cannot be rejected on general intuitive principles. But not everything is holistic, and commonsense (or even philosophical) opinion is of little use in spotting what is. Specific investigation, sometimes of a subtle and very technical mathematical nature, is needed to find out whether holistic representation is possible in any specific situation and whether (where it can be done) there is an exorbitant price to pay. The Exor machine illustrates, in a simple case, the concept of the cost of holism.

The task that Exor learned can be seen as a superposition of two learnings in the same network: learning to say yes to one-zero and learning to say yes to zero-one. An important fact is that each of these tasks, taken separately, is much easier to learn than the combined task. And this is not an occasional phenomenon: Exor is a very mild case of incurred cost of distribution. One of the research results of *Perceptrons,* and one that required some mathematical labor, shows that in certain situations the degree of difficulty of superposed tasks can exceed the difficulty of each separate task by arbitrary, large factors.

The romantic stance is to make a new network that isn't quite a perceptron and to assume it innocent until proven guilty of the danger of superposition costs. On the whole, connectionist literature does so even when reporting experiments in which the new networks show empirical signs of such costs as those that Exor incurs in its mild way. The rigorous stance assumes the possibility of guilt until innocence can be established: the theorems proved about perceptrons are seen as showing what kind of phenomena need to be precluded before one can make assertions confidently.

VII

I said at the beginning that I would offer some thoughts about Prince Charming. Who woke connectionism? Why this surge of interest and activity? Why now? And I will use my speculations on these themes to comment on the important question, What next?

A purely technical account of Snow White's awakening goes like this: In the olden days of Minsky and Papert, neural networking models were hopelessly limited by the puniness of the computers available at the time and by the lack of ideas about how to make any but the simplest networks learn. Now things have changed. Powerful, massively parallel computers can implement very large nets, and new learning algorithms can make them learn. No romantic Prince Charming is needed for the story.

I don't believe it. The influential recent demonstrations of new networks all run on small computers and could have been done in 1970 with ease. Exor is a "toy problem" run for study and demonstration, but the examples discussed in the literature are still very small. Indeed, Minsky and I, in a more technical discussion of this history (added as a new chapter to a reissue of *Perceptrons*), suggest that the entire structure of recent connectionist theories might be built on quicksand: it is all based on toy-sized problems with no theoretical analysis to show that performance will be maintained when the models are scaled up to realistic size. The connectionist authors fail to read our work as a warning that networks, like "brute force" programs based on search procedures, scale very badly.

A more sociological explanation is needed. Massively parallel supercomputers do play an important role in the connectionist revival. But I see it as a cultural rather than a technical role, another example of a sustaining myth. Connectionism does not use the new computers as physical machines; it derives strength from the "computer in the mind," from its public's largely nontechnical awareness of supercomputers.

I see connectionism's relationship to biology in similar terms. Although its models use biological metaphors, they do not depend on technical findings in biology any more than they do on modern supercomputers. But here too there is a powerful, resonant phenomenon. Biology is increasingly the locus of the greatest excitement. And neurosciences are invading the territory of academic psychology just as psychopharmacology is invading the territory of clinical psychology.

I also see a more subtle, but not less relevant, cultural resonance. This is a generalized turn away from the hard-edged rationalism of the time connectionism last went into eclipse and a resurgent attraction to more holistic ways of thinking. The actual theoretical discussion in the connectionist literature may not be connected in any strict sense to such trends in intellectual fashion. But here again, the concepts of sustaining myth and cultural resonance are pertinent: this time, perhaps, in a two-way process of mutual support.

Voilà Prince Charming: a composite of cultural trends. Reductionist undertones in my discussion do not undermine my good wishes for a happy union with Snow White. The new sense of excitement that is already replacing a certain ho-hum tiredness in cognitive science will ensure the fertility of the union. But the impact of connectionism will come less from the ideas it engenders than from heightened awareness of the problems it avoids.

ENDNOTES

[1] James G. Greeno, "The Cognition Connection," *New York Times Book Review*, 4 Jan. 1987.

[2] David E. Rumelhart, James L. McClelland, and the PDP Research Group, *Parallel Distributed Processing* (Cambridge: MIT Press, 1986).

[3] Marvin Minsky and Seymour Papert, *Perceptrons: An Introduction to Computational Geometry* (Cambridge: MIT Press, 1969).

[4] Rumelhart, McClelland, and the PDP Research Group, *Parallel Distributed Processing,* p. 111.