Also by Sherry Turkle

The Second Self: Computers and the Human Spirit
Psychoanalytic Politics: Jacques Lacan and Freud's French Revolution

LIFE ON THE SCREEN

Identity in the Age of the Internet

SHERRY TURKLE

SIMON & SCHUSTER
NEW YORK LONDON TORONTO SYDNEY TOKYO SINGAPORE

Chapter 5

THE QUALITY OF EMERGENCE

The field of artificial intelligence has a complex identity. It is an engineering discipline. Its researchers make smart artifacts—industrial robots that assemble cars, expert systems that analyze the stock market, computer agents that sort electronic mail. It also has a theoretical side. AI researchers try to use ideas about computer intelligence to think more generally about human minds. But there is not a clear division between these two sides of AI. Even "engineering AI" is more than a purely technical discipline. Its objects as well as its theories offer themselves as a mirror for contemplating the nature of human identity. Only a few years ago, it was primarily those who inhabited the rather small world of AI researchers who gazed into this mirror. Today, that mirror is starting to turn toward the face of popular culture.

Marvin Minsky, one of AI's founders, once characterized it as "trying to get computers to do things that would be considered intelligent if done by people." Minsky's ironic definition has remained in circulation for nearly a quarter of a century because it captures an enduring tension in the human response to "thinking machines." When confronted by a machine that exhibits some aspect of intelligence, many people both concede the program's competency and insist that their own human intelligence is precisely the kind the computer does not have. Or they insist that the type of intelligence the computer has is not the kind that makes people special. This response to the computer presence is sometimes provoked by an actual program and sometimes by the mere suggestion of one. It occurs not only on the boundary between minds and machines, but on the boundary between ideas about minds and ideas about machines. We have seen that it is not a simple manifestation of resistance to the idea of machine intelligence. It is also a part of how

people come to accept the idea. In this complex story, disavowal and appropriation are each tied up with the other.

This chapter traces a pattern of disavowal and appropriation in response to a major change in the philosophy of artificial intelligence research. From the late 1960s to the mid-1980s mainstream AI researchers conceived of computer intelligence as being made up of a complex set of rules programmed in advance. By the late 1980s, the field was more identified with theories of intelligence as emergent. Earlier we saw how both real and fictive images of emergent and "neural" AI were able to undermine long-standing resistance to computer psychotherapy in particular and machine intelligence in general. Now the story moves a step further. We will see how emergent AI has recently promoted the idea of a fundamental kinship between human and machine minds.

Information Processing in the Age of Calculation

In the tradition of romantic and magical thought, life is breathed into dead or inanimate matter by a person with special powers. In the early 1950s, there was a growing belief among a diverse group of engineers, mathematicians, and psychologists that this fantasy could be brought down to earth. During those early years, the atmosphere in AI laboratories was heady. Researchers were thinking about the ultimate nature of intelligence, and they were sure it could be captured in machines. The goal, mythic in proportion, was to use computers to generate a fragment of mind. AI researchers combined intellectual fervor with academic imperialism. They aspired to use computational principles to reshape the disciplines of philosophy, psychology, and linguistics.

These early AI researchers divided into two camps, each supporting one of the two primary competing models for how AI should be done. One group considered intelligence entirely formal and logical and pinned its hopes on giving computers detailed rules they could follow. The other envisioned machines whose underlying mathematical structures would allow them to learn from experience. The proponents of the second vision imagined a system of independent agents within a computer from whose simultaneous interactions intelligence would emerge.[1] From the perspective of these researchers, a rule was not something you gave to a computer but a pattern you inferred when you observed the machine's behavior.

In the mid-1960s, the early emergent models seemed as promising as the rule-driven, information processing approach. However, by the end of that decade, the emergent models had been largely swept aside. One problem was that the emergent models relied on the results of the simultaneous interactions of multiple independent agents, but the computers of the era could only handle one computation at a time. Additionally, simple emergent systems were shown to have significant theoretical limitations[2] and more sophisticated mathematical techniques for hooking up programs that would operate in parallel were not well developed. Rule-based AI came to dominate the field. It dominated efforts to create general models of intelligence and it dominated the burgeoning subdiscipline of expert systems. Expert systems were literally built out of rules. They were created by debriefing human experts to determine the rules they follow and trying to embody these in a computer.

Douglas Hofstadter, author of *Gödel, Escher, Bach: The Eternal Golden Braid,* called the 1970s the era of AI's Boolean dream.[3] George Boole, the nineteenth-century mathematician, had formalized a set of algebraic rules for the transformation of logical propositions. Apparently not one for understatement, he called these rules the Laws of Thought.[4] Boole's laws were far from an all-inclusive model of mind. For one thing, they needed an external agent to operate them. However, computers were able to breathe life into Boole's equations by placing an operator in the form of a computer program right into the system. Once there, the operator and the laws could be seen as a functioning model, if not of the mind, at least of part of the mind.

Information processing AI gives active shape to formal propositions and creates an embodiment of intelligence as rules and reason. Boole would have felt an intellectual kinship with Allen Newell and Herbert Simon, pioneers of information processing AI, who saw brain and computer as different examples of a single species of information processing device.

In the late 1950s, in the spirit of "The Laws of Thought," Newell and Simon wrote a program called the General Problem Solver (GPS) that attempted to capture human reasoning and recode it as computational rules. Questions about GPS's "reasoning" could be answered by referring to whatever rules it had been given, even though the interaction of the rules might produce unpredictable results.

As the GPS became well known in academic circles, some psychologists began to wonder why it should not be possible to ask similar questions about how *people* solve logical problems. In the intellectual atmosphere of the time, this train of thought was countercultural. American academic psychology was dominated by behaviorism, which rigidly excluded the discussion of internal mental states. Orthodox behaviorists insisted that the study of mind be expressed in terms of stimulus and response. What lay between was a black box that could not be opened. So, for example, behaviorist psychologists would not refer to memory, only to the behavior of remembering.

By the end of the 1960s, however, behaviorism was in retreat. Some psychologists were willing to open the black box of the human mind and talk about the processes taking place inside it. The computer had an important metaphorical role to play in the demise of behaviorism. The very *existence* of the computer and the language surrounding it supported a way of thinking about mind that undermined behaviorism. Computer scientists had of necessity developed a vocabulary for talking about what was happening inside their machines, the internal states of their systems. And AI researchers freely used mentalistic language to refer to their programs—referring to their "thoughts," "intentions," and "goals." If the new machine minds had internal states, common sense suggested that people must have them too. The psychologist George Miller, who was at Harvard during the heyday of behaviorism, has described how psychologists began to feel uneasy about not being allowed to discuss human memory now that computers were said to have one:

> The engineers showed us how to build a machine that has memory, a machine that has purpose, a machine that plays chess, a machine that can detect signals in the presence of noise, and so on. If they can do that, then the kind of things they say about the machines, a psychologist should be permitted to say about a human being.[5]

In this way, the computer presence legitimated the study of memory and inner states within psychology. "Suddenly," said Miller, "engineers were using the mentalistic terms that soft-hearted psychologists had wanted to use but had been told were unscientific."[6] The machines supported an intellectual climate in which it was permissible to talk about aspects of the mind that had been banned by behaviorism.

That these ideas came from a hard-edged engineering discipline raised their status in a community of psychologists that still tended to see science as an objective arbiter of truth. Although information processing ideas challenged behaviorism, their mechanistic qualities also had a certain resonance with it. This shared sensibility eased the way for the appropriation of computational models by psychologists.

This new psychology for describing inner states in terms of logic and rules came to be known as cognitive science and the computer presence served as its sustaining myth. Cognitive science was in harmony with what I have called the modernist intellectual aesthetic of the culture of calculation. Mechanism and at least the fantasy of transparency was at its heart.

When I began my studies of the computer culture in the mid-1970s, artificial intelligence was closely identified with information processing and the rule-based approaches of cognitive science.[7] Cognitive science

may have provided psychology with a welcome respite from behaviorist orthodoxy, and rule-based expert systems had considerable worldly success in business and medicine, but the spread of information processing ideas about the human mind met with significant resistance in the broader culture.

During the 1970s to the mid-1980s, many people I interviewed responded to advances in information processing AI by agreeing with the premise that human minds are some kind of computer but then found ways to think of themselves as something more than that. Their sense of personal identity often became focused on whatever they defined as "not cognition" or "beyond information." People commonly referred to spontaneity, feelings, intentionality, and sensuality in describing what made them special. They conceded to the rule-based computer some power of reason and then turned their attention to the soul and spirit in the human machine.

For some, the appropriation and disavowal of computational images of mind took the form of a pendulum swing. In 1982, a thirty-two-year-old nurse said: "I'm programmed to fall for the same kind of man every time. I'm like a damned computer stuck in a loop. . . . I guess my cards are punched out the same way." But a few minutes later, she described her emotional life in terms of what the computer was not: "When people fall in love or their passions for their children, it's like a blinding emotion. Computers don't have anything to do with that." Others split the self. One student spoke of his "technology self" and his "feeling self," another of her "machine part" and her "animal part." When talking about family life, people might insist there was nothing machine-like about their emotions. When talking about business decisions, they thought they might be working like a computer program. Thus, for many people, competing views of the self existed simultaneously. There was no victory of one model over another; there was only ambivalence.

Everyday expressions of reluctance about the idea of intelligent machines had counterparts in the philosophical community's responses to AI. In the 1960s, Hubert Dreyfus argued that there was a fundamental difference between human and computer intelligence. For Dreyfus, human intelligence was not reducible to propositions or rules that could be specified in advance; it arose through having a body and experiencing a changing world. Dreyfus held that without embodied knowledge computers "could not do" intellectual tasks that required intuition and experience.[8] He further held that beating him at chess was one of those tasks. This turned out not to be the case, since by 1966, a chess program was able to triumph over Dreyfus and other, even more skilled, players.

Dreyfus had set a trap for himself by defining human uniqueness in terms of machine performance, a definition that had to remain one step

ahead of what engineers could come up with next. In 1980, John Searle's Chinese Room thought experiment took a different tack,[9] by making the point that real intelligence was not about what computers could do, but whether they could really be said to understand.

Searle had no argument with what he called "weak AI," artificial intelligence research that tries to use the study of machine intelligence to generate potentially useful insights about human processes. Rather, Searle concentrated his attack on "strong AI," which contends that intelligent machines actually demonstrate how people think. The Chinese Room dealt a blow to this school of thought, because Searle described the inner workings of a computer program in terms so alien to how most people experience their own minds that they felt a shock of nonrecognition.

Searle's paper appeared at a time of general disappointment with progress in information processing AI. During more optimistic times, Marvin Minsky had frequently been quoted as saying that almost any apparently complex aspect of human intelligence "could probably be described by three algorithms." By the mid-1980s, such absolute faith was sorely tested. It was becoming clear that vast realms of mind could not be easily grasped by information processing or expert-system formalisms.

The Chinese Room served as something of a cultural watershed. It defused a sense of threat from information processing, but it left the door open to a startling rejoinder: Although the man inside the room did not understand Chinese, perhaps the entire room could be said to understand Chinese! Similarly, no one part of the brain understands Chinese, but the brain as a whole does. In other words, intelligence was distributed; it existed within the system as a whole, not within any particular agent in the system. Intelligence did not reside in an isolated thinking subject, but in the interaction of multiple fragments of mind, figuratively speaking, in a society of mind. This rejoinder was an indication of where the computer culture was going in the 1980s. It was going back to some long abandoned images from emergent AI. The images were biological and social.

Emergent AI

The renaissance of emergent AI took up a research tradition from the 1960s that was based on a simple emergent system known as the perceptron. A perceptron is a computer program made up of smaller programs called agents, each of which has a narrow set of rules it can follow and a small amount of data on which to base its decisions. All agents "vote" on a question posed to the perceptron, but the system weights

their votes differently depending on the individual agent's past record of success. Those agents who guess right more often end up having more of a voice in subsequent decision-making. In this sense, the perceptron learns from its experiences. On a metaphorical level, the perceptron's intelligence is not programmed into it, but grows out of the agents' competing voices.

To get a sense of how this works, imagine trying to design a system for predicting rain. One would begin by accessing the opinions of, say, a thousand simple-minded meteorologists, analogous to the agents of the perceptron, each of whom has a different imperfect method of forecasting rain. Each meteorologist bases his or her judgment on a fragment of evidence that may or may not be related to predicting rain. One possibility would be simply to identify the meteorologist who has the best track record for rain prediction and always go with that meteorologist's vote. Another strategy would be to let the majority of the voting meteorologists decide. The perceptron refines this strategy by weighting each vote according to individual meteorologists' records.

In an information processing model, the concept "rain" would be explicitly represented in the system. In the perceptron, the prediction "it will rain" is born from interactions among agents, none of which has a formal concept of rain. Information processing begins with formal symbols. Perceptrons operate on a subsymbolic and subformal level. The analogy with the neurons in the brain is evident.

If you applied the information processing method to the rain-forecasting example you would have complete breakdown if your chosen meteorologist became incapacitated. But in the brain, damage seldom leads to complete breakdown. More often it produces a gradual degradation of performance. When things go wrong, the system still works, but just not as well as before. Information processing systems lost credibility as models of mind because they lacked this feature. The perceptron showed the gradual degradation of performance that characterizes the brain. Even when injured, with some disabled meteorologists on board, the perceptron still can produce weather forecasts.

This analogy with brain performance was decisive for connectionists, the group of emergent AI researchers who most seriously challenged the information processing approach in the mid-1980s.[10] The connectionists used programs known as learning algorithms that are intellectual cousins to the perceptron. They spoke of artificial neurons and neural nets and claimed that the best way to build intelligent systems was to simulate the natural processes of the brain as closely as possible.[11] A system modeled after the brain would not be guided by top-down procedures. It would make connections from the bottom up, as the brain's neurons are thought to do. So the system could learn by a large number of different con-

nections. In this sense, the system would be unpredictable and nondeterministic. In a manner of speaking, when connectionists spoke of unpredictable and nondeterministic AI, they met the romantic reaction to artificial intelligence with their own romantic machines.

Some of the connectionists described themselves as working at a subsymbolic level: They didn't want to program symbols directly, they wanted symbols (and their associated meanings) to emerge. The connectionists were still writing programs, but they were operating on a lower level of objects within the computer. By working at a lower level, they hoped to achieve systems of greater flexibility and adaptability.[12]

In the mid-1980s, such connectionist images began to capture popular as well as professional attention. The idea that computers would not have to be taught all necessary knowledge in advance but could learn from experience was appealing at a time when it was increasingly clear that it was easier to teach a computer to play chess than to build a mudpie. AI researchers had succeeded in getting computers to play excellent chess but had stumbled on such feats as recognizing human faces. The connectionist models suggested another way to approach the problem. Instead of searching for the rules that would permit a computer to recognize faces, one should "train" a network of artificial neurons. The network could be shown a certain number of faces and be "rewarded" when it recognized one. The network would be woven through with a learning algorithm that could give feedback to the system, establishing the appropriate connections and weights to its elements. Unlike information processing AI, which looked to programs and specific locations for information storage, the connectionists did not see information as being stored anywhere in particular. Rather, it was inherent everywhere. The system's information, like information in the brain, would be evoked rather than found.[13]

The resurgence of models that attempted to simulate brain processes was indissociable from a new enthusiasm about parallel-processing computers.[14] Although in the brain, millions of things might be happening at once, the standard computers available through the 1980s performed only one operation at a time. By the mid-1980s, two developments made the reality of massive parallel computing seem closer, at least as a research tool. First, computers with high parallel-processing capacity (such as the Connection Machine with its 64,000 processors) were becoming economically and technically feasible. Second, it became possible to simulate parallel-processing computers on powerful serial ones. Although this resulted not in real but in virtual parallel-processing machines, they turned out to be real enough to legitimate connectionist theories of mind. Parallel computation was established as a new sustaining myth for cognitive science.[15]

The 1980s saw researchers from many different fields writing papers that emphasized both parallel processing and intelligence emerging from the interaction of computational objects. These papers came from engineers enthusiastic about building parallel machines, computer scientists eager to try new mathematical ideas for machine learning, and psychologists looking for computer models with a neurological resonance. As the decade progressed, cognitive psychology, neurobiology, and connectionism developed a sense of themselves as more than sister disciplines; these diverse areas of study were starting to think of themselves as branches of the same discipline, united by the study of emergent, parallel phenomena in the sciences of mind, separated only by the domains in which they looked for them.

By the mid-1980s, it was clear that emergent AI had not died but had only gone underground. Now that the emergent tradition had resurfaced, it did so with a vengeance. In 1988, the computer scientist Paul Smolensky summed up the situation with the comment: "In the past half-decade the connectionist approach to cognitive modeling has grown from an obscure cult claiming a few true believers to a movement so vigorous that recent meetings of the Cognitive Science Society have begun to look like connectionist pep rallies."[16] By the 1990s, emergent AI had done more than enter the mainstream; it had become the mainstream.

With the resurgence of emergent AI, the story of romantic reactions to the computer presence came full circle. In the popular culture, people had been trying to establish human uniqueness in contrast to computers while in the research community the proponents of emergent AI were linking computers to the world of humans through biological and social metaphors. Now both people and computers were said to be "nondeterministic," "spontaneous," and "nonprogrammed." The story of romantic reactions to the computer presence was no longer simply about people responding to their reflection in the mirror of the machine. Now computer designers were explicitly trying to mirror the brain. There had been a passage through the looking glass.

From the beginning, the language of emergent AI borrowed freely from the languages of biology and of parenting. Not only did it refer to associations of networked computational objects as neural nets, but it presented programs as though they were white mice that might or might not learn to run their mazes, or children who might or might not learn their lessons. This way of talking was picked up by the users of the new connectionist programs and by the media.[17] Dave, forty years old, a high school English teacher and baseball coach, uses small connectionist programs to help him figure out what team to field. When I talk to him about his work, he speaks about his programs with something akin to fatherly pride. "I love to watch my programs do their thing. They get better right

in front of me. When you watch a little creature improve session by session, you think of it as a child even if it is a computer." While developers of information processing AI had been popularly depicted as knowledge engineers, hungry for rules, debriefing human experts so as to embody their methods in theorems and hardware, a computer scientist working in the new tradition of emergent AI was portrayed as a creator of life, "his young features rebelling, slipping into a grin not unlike that of a father watching his child's first performance on the violin," running his computer system overnight so that the agents within the machine would create intelligence by morning.[18]

In the romantic reaction to the computer presence during the late 1970s and early 1980s, it had become commonplace to paraphrase the famous remark of Lady Ada Lovelace, who in 1842 said, "The analytical engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform." In other words, computers only do what you tell them to do, nothing more, nothing less, or more colloquially, "garbage in, garbage out." The Lovelace objection to a computer model of mind was essentially that people don't follow rules. People learn and grow. And they make new connections that "mysteriously" emerge. The Lovelace objection worked fairly well for criticizing information processing models of the mind.[19] But emergent AI was characterized by explicitly "anti-Lovelace" representations of the computer. It implied a continuity between computers and people. Connectionism suggested that it was an experimental science and that there was mystery and unpredictably inside its machines.

W. Daniel Hillis, the inventor of the Connection Machine, refers to this mysterious quality as the appealing inscrutability of emergent systems. For Hillis, there was an enchantment in opacity. Inscrutable systems are the most anti-Lovelace and thus the most appealing thing a computer could aspire to be. They are as close as a computer could come to overcoming romantic objections to information processing. And they are as close as a computer could come to renouncing the modernist idea of understanding through the analysis of underlying mechanism. For Hillis, inscrutability is "seductive because it allows for the possibility of constructing intelligence without first understanding it. . . . The apparent inscrutability of the idea of intelligence as an emergent behavior accounts for much of its continuing popularity."[20] For Hillis, emergence "offers a way to believe in physical causality while simultaneously maintaining the impossibility of a reductionist explanation of thought. For those who fear mechanistic explanations of the human mind, our ignorance of how local interactions produce emergent behavior offers a reassuring fog in which to hide the soul."[21]

In a similar spirit, the AI researcher Terry Winograd once commented

that people are drawn to connectionism because its opaque systems allow for a high percentage of wishful thinking.[22] The remark was meant to be critical, but connectionists didn't have to take it that way. In nature, intelligence does not depend on the ability to specify process. Why should it when we build "second natures" in machines? If computers, like brains, are closed boxes, why should this interfere with their functioning as minds?

The movement from information processing to emergent AI marks a critical change in how AI approaches its central scientific problem. You can't get to connectionism by making incremental improvements to information processing systems. It requires a fundamental change in approach. In the history of science, such changes of approach stir up strong emotion.[23] In the 1980s, the confrontation in the research community between emergent and information processing AI was tense and highly charged. While Douglas Hofstadter tried to capture the spirit of emergent AI in the phrase "waking up from the Boolean dream,"[24] the champions of connectionism had found that dream to be more like a nightmare. To them it seemed obvious that since human intelligence was more than a set of rules, the computers that modeled it should not be about rules either. Like nineteenth-century Romantics, connectionists sought to liberate themselves from a rationalism they experienced as constraining and wrong-headed.

In the mid- to late 1980s, the cultural appeal of connectionism was in part that it could describe computers in much the same way that personal computer owners were being encouraged to see them: as opaque systems in which emergent processes occur. There was a certain irony here. A quarter of a century before, the presence of the computer had challenged the behaviorist insistence on the mind as black box. Now, in some ways, emergent AI was closing the box that information processing had opened.

Information processing AI had opened the black box of the mind and filled it with rules. Connectionism replaced the idea that intelligence was based in logical understanding with a new emphasis on experience as the bedrock for learning. It postulated the emergence of intelligence from "fuzzy" processes, so opening up the box did not reveal a crisply defined mechanism that a critic could isolate and ridicule.[25] Information processing had provided an excuse for experimental psychology to return to the consideration of inner process. Now, emergent models invited philosophers, humanists, and a wider range of psychologists to compare machines to humans.

In the 1980s, connectionism became part of a complex web of intellectual alliances. Its way of talking about opacity made it resonant with the aesthetic of depthlessness that Fredric Jameson had classified as postmodern. Its commitment to neurological metaphors created a link to brain

scientists who tried to visualize the mind through sophisticated computer imaging.[26] Its assertion that mind could not be represented as rules made it interesting to humanists and post-positivist philosophers.

Connectionism began to present the computer as though it were an evolving biological organism. The neurons and pathways of connectionism were designed on the template of biology. Connectionism opened the way for new ideas of nature as a computer and of the computer as part of nature. And it thus suggested that traditional distinctions between the natural and artificial, the real and simulated, might dissolve.

A Postmodern Convergence

By the late 1980s it was clear that many of those who had been most critical of information processing AI were disarmed by connectionism's romantic, postmodern allure and by its new emphasis on learning through experience, sometimes referred to as situated learning.[27] Even Hubert Dreyfus cautiously expressed his interest in connectionism. Dreyfus's critique of information processing had drawn on the writing of Martin Heidegger and the later works of Ludwig Wittgenstein. "Both these thinkers," said Dreyfus, "had called into question the very tradition on which symbolic information processing was based. Both were holists, both were struck by the importance of everyday practices, and both held that one could not have a theory of the everyday world."[28] But Dreyfus was sympathetic to connectionism because he saw it as consistent with such views:

> If multilayered networks succeed in fulfilling their promise, researchers will have to give up the conviction of Descartes, Husserl, and early Wittgenstein that the only way to produce intelligent behavior is to mirror the world with a formal theory of mind. . . . Neural networks may show that Heidegger, later Wittgenstein and Rosenblatt [an early neural net theorist] were right in thinking that we behave intelligently in the world without having a theory of that world.[29]

Philosophers like Dreyfus were joined in their enthusiasm for connectionism by cultural critics who had long been skeptical about the impact of technology on humanistic values. For example, the literary scholar Leo Marx found the "contextual, gestaltist, or holistic theory of knowledge implicit in the connectionist research program" to be "particularly conducive to acquiring complex cultural understanding, a vital form of liberal knowledge."[30] Although Marx's real sympathy was less for connectionism than for its metaphors, his comments illustrate how the new approach

opened possibilities for intellectual alliances that had been closed to information processing. In general, connectionism received good press from both professionals and the lay public as a more humanistic form of AI endeavor. By this, they usually meant that connectionism left room for mind to have complexity and mystery.

Marvin Minsky had long justified the AI enterprise with the quip, "The mind is a meat machine." The remark was frequently cited during the late 1970s and early 1980s as an example of what was wrong with artificial intelligence. Minsky's comment provoked irritation, even disgust. Much of what seemed unacceptable about Minsky's words had to do with the prevailing images of what kind of meat machine the mind might be. Those images were mechanistic and deterministic. Connectionism's appeal was that it proposed an artificial meat machine made up of biologically resonant components. With a changed image of what machines could be, the idea that the mind could be one became far less problematic. Edith, a thirty-four-year-old physician whose residency in psychiatry included readings on connectionist neuroscience, was enthusiastic about its prospects for modeling mind. "The mind may be a machine, but it's not just any old machine," she said. "Connectionism fits the picture because it's scientific, but not deterministic."

In the 1980s, Minsky, long associated with information processing, became sympathetic to a form of emergent AI. There was considerable irony in this. In the late 1960s, Minsky and the mathematician Seymour Papert had coauthored *Perceptrons,* a book that had helped put early emergent AI into eclipse. And yet in his 1985 book, *The Society of Mind,* Minsky describes an emergent system, an inner world of highly anthropomorphized agents. Each agent has a limited point of view. Complexity of behavior, emotion, and thought emerge from the interplay of their opposing views, from their interactions and negotiations.[31] Minsky's society theory differs from connectionism in that it implies a greater degree of programming of the inner agents. However, it may be seen as a variant of emergent theory, because in it intelligence does not follow from programmed rules but emerges from the associations and connections of objects within a system. One MIT student is extravagant in his description of Minsky's new model. "With the idea of mind as society," he says, "Minsky is trying to create a computer complex enough, indeed beautiful enough, that a soul might want to live in it." Emergent AI appears to soften the boundaries between machines and people, making it easier to see the machine as akin to the human and the human as akin to the machine.

Decentered Psychology

Emergent AI depends on the way local interactions among decentralized components can lead to overall patterns. So does the working of ant colonies and the immune system, the pile-up of cars in a traffic jam, and the motion of a flock of birds. The result is a perfectly coordinated and graceful dance. Mitchel Resnick, an educational researcher at MIT, has noted new cultural interest in such emergent models. He calls it the "decentralized mindset."[32]

Decentralized models have appeared in economics, ecology, biology, political science, medicine, and psychology. In the latter, psychoanalytic theory has been an important actor in the development of decentralized or decentered views of the self.

Early psychoanalytic theory was built around the idea of drive: a centralized demand that is generated by the body and that provides the energy and goals for all mental activity. But later, when Freud turned his attention to the ego's relations to the external world, he began to describe a process by which we internalize important people in our lives to form inner "objects."[33] Freud proposed this kind of process as the mechanism for the development of the superego, what most people think of as the conscience. The superego was formed by taking in, or introjecting, the ideal parent.

In Freud's work, the concept of inner objects coexisted with drive theory; we internalize objects because our instincts impel us to. But many theorists who followed Freud were less committed to the notion of drive than to the idea that the mind was built up of inner objects, each with its own history. Whereas Freud had focused his attention on a single, internalized object—the superego—a group of later psychoanalysts, collectively known as object-relations theorists, widened the scope of the inquiry about the people and things that each of us is able to bring inside.[34] They described the mind as a society of inner agents—"unconscious suborganizations of the ego capable of generating meaning and experience, i.e. capable of thought, feeling, and perception."[35] In the work of the psychoanalyst Melanie Klein, these inner agents can be seen as loving, hating, greedy, or envious. The psychoanalyst W. R. D. Fairbairn envisioned independent agencies within the mind that think, wish, and generate meaning in interaction with one another.[36] What we think of as the self emerges from their negotiations and interactions.

Thus, while Freud believed that a few powerful inner structures like the superego act on memories, thoughts, and wishes, in object-relations theory the self becomes a dynamic system in which the distinction between processor and processed breaks down. A French school of psychoanalytic theory, inspired by Jacques Lacan, went even further. Lacan viewed the idea of a centralized ego as an illusion. For him, only the *sense* of an ego emerges from chains of linguistic associations that reach no endpoint. There is no core self. What we experience as the "I" can be likened to something we create with smoke and mirrors.

The parallel between the historical development of psychoanalysis and the historical development of artificial intelligence is striking. In both fields there has been movement away from a model in which a few structures act on more passive substance. Psychoanalysis began with drive and artificial intelligence began with logic. Both moved from a centralized to a decentered model of mind. Both moved toward a metatheory based on objects and emergence. Both began with an aesthetic of modernist understanding. Both have developed in directions that come close to shattering the idea that modernist understanding is possible. In the case of psychoanalysis, which developed as one of the great metanarratives of modernism, both the object-relations and Lacanian traditions have substantially weakened its modernist core. Psychoanalysis is a survivor discourse, finding a voice in both modernist and postmodern times. AI, too, may be such a survivor discourse.

Psychoanalysts were almost universally hostile to information processing AI, because they felt it reduced the Freudian search for meaning to a search for mechanism, as, for example, when AI researchers and computer science students would reinterpret Freudian slips as information processing errors.[37] But psychoanalysts have shown considerable interest in emergent AI.[38]

Consider the images in Minsky's *The Society of Mind.* There, he describes how in a microworld of toy blocks, agents that at first seem like simple computational subroutines work together to perform well-defined tasks like building towers and tearing them down. Minsky speculates how, in a child's mind, the agents responsible for "Building" and "Wrecking" might become versatile enough to offer support for one another's goals. In Minsky's text, they utter sentences like, "Please Wrecker, wait a moment more till Builder adds just one more block: it's worth it for a louder crash."[39] It quickly becomes clear that what Minsky has in mind are not mere computational subroutines but a society of subminds that collaborate to produce complex behavior.[40] The kind of emergence implicit in Minsky's society model has a natural affinity with object-relations psychoanalysis. Indeed, Minsky's language evokes the world of the psychoanalyst Fairbairn. And connectionism's language of links and associations evokes the radically decentered theories of Lacan and is appealing to analysts eager to reconcile Freudian ideas with neurobiology.

A 1992 paper by the psychoanalyst David Olds explicitly tries to recruit psychoanalysts to a connectionist view of the mind.[41] Olds argues that

psychoanalysts need connectionist theory because it presents them with a plausible link to biology; analysts can use its models to provide an account of the ego in terms of the brain. Connectionism can also help psychoanalysis undermine centralized and unitary views of the ego and support the notion of a decentered self. Historically, theories of a decentered self have needed to be supported.

Freud's notion of the unconscious had called into question the idea of the unitary self as an actor and agent. We don't know what we want, said Freud. Our wishes are hidden from us by complex processes of censorship and repression. Yet even as Freud decentered the ego, some of the theorists who followed him, collectively known as ego psychologists, sought to restore its central authority. They did so by focusing on the ego as a stable, objective platform from which to view the world. They began to see the ego as capable of integrating the psyche. To the psychoanalysts Anna Freud and Heinz Hartmann, the ego seemed almost a psychic hero, battling off id and superego at the same time as it tried to cope with the demands of the external world. Anna Freud wrote of the ego's powerful artillery, its "mechanisms of defense," and Hartmann argued that the ego had an aspect that was not tied up in the individual's neurotic conflicts; it had a conflict-free zone. This unhampered aspect of the ego was free to act and choose, independent of constraints. Hartmann's concept of a conflict-free zone was almost the site of a reborn notion of the will, the locus of moral responsibility. The intellectual historian Russell Jacoby, writing of ego psychology, described it as the "forgetting of psychoanalysis."[42]

For Olds, connectionism challenges ego psychology by providing a way to see the ego not as a central authority but as an emergent system. Through a connectionist lens, says Olds, the ego can be recast as a distributed system. Consciousness can be seen as a technical device by which the brain represents its own workings to itself. Olds likens it to "the monitor on a computer system," underscoring its passive quality. Even clinical practice can be interpreted in connectionist language: Interpretations that an analyst makes during a treatment session work when they correspond to a "well worn track in the brain, namely a set of connections among nets which generates a repetitive pattern of response and behavior."[43]

Olds acknowledges that really understanding connectionism requires "considerable mathematical sophistication." "[V]ery few people, including most psychologists, have even a sketchy understanding" of what the theory is actually saying. But he believes that connectionism will nevertheless be increasingly influential among psychoanalysts. Innocence of technical details has not kept psychology from mining scientific fields for their metaphors. Freud, for example, was not a master of hydraulic theory, but he borrowed many of his central images from it. Olds suggests that today's psychoanalysts should view connectionism in a similar spirit. What hydraulics was to Freud, emergent AI should be to today's analysts. In other words, Olds is explicitly advocating the use of connectionism as what I have called a sustaining myth:

> Many libido theorists probably did not know a great deal about steam engines; they made conceptual use of the properties which interested them. This is even more true with the early computer model; very few analogizers know a motherboard from a RAM, nor do they care. The way we *imagine* the machine handles information is what counts.
>
> The point is that what gets transferred from one realm to the other is a set of properties which we attribute to both entities.[44]

These remarks recall the way computers served as support for cognitive science in the 1950s and 1960s. There too, what the machines did was less important than how people thought about them. As Olds points out, although the theory of neural nets may be technically difficult, it is metaphorically evocative, presenting machine processes as the kinds of things that go on in the brain.[45]

THE APPROPRIABILITY OF EMERGENT AI

When the prevailing image of artificial intelligence was information processing, many who criticized the computer as a model of mind feared that it would lead people to view themselves as cold mechanism. When they looked at the computer, they had a "not me" response. Now we face an increasingly complex situation. These days when people look at emergent computer models they see reflected the idea that the "I" might be a bundle of neuron-like agents in communication. This sounds close enough to how people think about the brain to begin to make them feel comfortable. The not-me response turns into a like-me response.

I noted earlier that Freudian ideas became well known and gained wide acceptance for reasons that had little to do with their purported scientific validity. Ideas about the importance of slips of the tongue became part of the wider psychological culture not because they were rigorously proven but because slips were evocative objects-to-think-with. As people looked for slips and started to manipulate them, both seriously and playfully, the psychoanalytic concepts behind them began to feel more natural. Many of the ideas behind emergent AI are appropriable for the same reasons that slips were. For example, you can play with the agents of Minsky's society theory. You can imagine yourself in their place; acting out their

roles feels enough like acting out the theory to give a sense of understanding it. The language of "society" is helping to disseminate the idea that machines might be able to think like people and that people may have always thought like machines. As for connectionism, it too has been gleaned for appropriable images. Some people mentally translate the idea of connection strengths between neuron-like entities into the notion of moving things closer together and further apart. Other people translate connectionist ideas into social terms. One twenty-two-year-old laboratory technician transformed the neural networks into a network of friends:

> The neural nets, like friends, can join up in teams in many different combinations and degrees of closeness, depending on how gratifying their relationships are. If a group of neuron friends makes a good, strong combination, their associations are going to get stronger. They will increase their degree of association.

Clearly, what is involved here is not a weighing of scientific theory but an appropriation of images and metaphors. Although emergent AI is more opaque than information processing in terms of traditional, mechanical ways of understanding, it is simultaneously more graspable, since it builds intelligence out of simulated "stuff" as opposed to logic. Because the constituent agents of emergent AI offer almost tangible objects-to-think-with, it prepares the way for the idea of mind as machine to become an acceptable part of everyday thinking.

The diffusion of popular versions of connectionist ideas about mind has been greatly facilitated by the fact that small neural net programs are easily run on widely available desktop computers. The two-volume *Parallel Distributed Processing,* what might be thought of as the Bible of connectionism's rebirth, was published in 1986. It inspired a flurry of programming activity, and not just among AI researchers and professional programmers. The PDP programs were simple enough for high school hackers and home computer aficionados to experiment with. For years, James McClelland and David Rumelhart, the editors of the PDP volume, had made their programs available to students in the universities where they taught—Carnegie Mellon, Stanford, and the San Diego campus of The University of California. But after the PDP volume was published, the demand for the programs was so great that Rumelhart and McClelland decided to put them on disks that would run on personal computers. The hardware requirements for running the programs were scarcely state-of-the-art. In the IBM version, you could be a connectionist with 256 kilobytes of memory, two floppy disk drives, a standard monochrome monitor, and version 2.0 of MS-DOS. It was like being told that you could be a cordon bleu chef using only a small Teflon frying pan and spatula.

In writing about the spread of ideas about microbes and the bacterial theory of disease in late nineteenth-century France, the sociologist of science Bruno Latour has argued that what spread the word was not the message put out by Louis Pasteur's writings, but the social deployment of an army of hygienists, state employees who visited every French farm.[46] They were the foot soldiers of Pasteur's revolution. PDP programs on floppy disks functioned similarly as carriers of emergent theories of mind.

The PDP disks and an accompanying workbook were published in 1988 with an "exhortation and a disclaimer." The disclaimer was that Rumelhart and McClelland could not "be sure that the programs are perfectly bug free."[47] They encouraged people to work around difficulties, to fix things where they could, and to send in their comments and suggestions. The developers of a cutting-edge scientific field were asking for help from their lay audience. This openness to criticism and collaboration was appealing. The exhortation was to

> take what we offer here, not as a set of fixed tasks to be undertaken, but as raw material for your own explorations. . . . The flexibility that has been built into these programs is intended to make exploration as easy as possible, and we provide source code so that users can change the programs and adapt them to their own needs and problems as they see fit.[48]

In other words, PDP was presented in a way that spoke directly to the learning style of the tinkerer—try it, play with it, change it—*and* to those who wanted to go below the surface. PDP combined the magic of emergence with the possibility of getting your hands dirty at the level of the source code. It was a powerful combination: hard and soft, bricolage and algorithm. It transcended old dichotomies: You could have your bricolage and feel like a real scientist too.

Emergent AI's message about complexity and emergence seems to be something that many people want to hear. The nondeterminism of emergent systems has a special resonance in our time of widespread disaffection with instrumental reason. Seymour Papert speculates that at least in part, emergent AI has been brought back to life by the resonance of its intellectual values with those of the wider culture, which has experienced 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."[49]

So emergent AI manages to be seductive in many ways. It presents itself as escaping the narrow determinism of information processing. Its images are appealing because they refer to the biology of the brain. Like fuzzy logic and chaos theory, two other ideas that have captured the popular and professional imagination during the last decade, emergent AI ac-

knowledges our disappointments with the cold, sharp edges of formal logic. It is consonant with a widespread criticism of traditional Western philosophy, which, as Heidegger once put it, had focused on fact in the world while passing over the world as such.[50] Emergent AI falls into line with postmodern thought and a general turn to "softer" epistemologies that emphasize contextual methodologies. And finally, its constituent agents offer a theory for the felt experience of multiple inner voices. Although our culture has traditionally presented consistency and coherence as natural, feelings of fragmentation abound, now more than ever. Indeed, it has been argued that these feelings of fragmentation characterize postmodern life.[51] Theories that speak to the experience of a divided self have particular power.

Like all theories that call into question an autonomous, unitary ego, emergent AI lives in a natural state of tension. Among the reasons decentered theories of mind are powerful is that they offer us a language of the self that reflects our sense of fragmentation. On the other hand, they are also under pressure from our everyday sense of ourself as unified. No matter what our theoretical commitments to a notion of a decentered self, when we say, "I do, I say, I want," we are using a "voice" that implies unity and centeredness. This tension between theory and the assumptions of everyday language has been played out in the history of the popular appropriation of psychoanalysis for nearly a century, most starkly in repeated confrontations with ideas that present the ego as the central executive of the mind. And now it is being played out in the history of the popular appropriation of artificial intelligence.

Information processing AI challenged the idea of the centered self when it equated human minds with rule-driven machine processes, but it offered an opening to centralized views of the mind. If the mind was conceived as a hierarchical program, it was relatively easy to imagine a program on top of the hierarchy that could be analogous to an executive ego. Emergent theories of AI are more radically decentralizing in their intent. Yet, like Freudian theory, they too are challenged by those who would recast them into centralized forms.

In the mid-1980s, MIT students influenced by Marvin Minsky's society theory said they were content to see their minds in a radically decentralized fashion. They spoke of their minds as "a lot of little processors," or as one put it, "In my mind, nobody is home—just a lot of little bodies."[52] However, even among those most committed to Minsky's views, some were tempted to put centralized authority back into his system, to make one of the little processors more equal than the others.[53] One young woman told me that one of the agents in her society of mind had the ability to recognize patterns. She said that the pattern-recognition agent was able to build on this skill, "grow in its ability to manipulate data,"

and "develop the ability to supervise others." Like ego psychologists with their conflict-free zone, she had reintroduced an executive agent into her idea of decentered mind.

Minsky's society model leaves a fair amount of room for such recentralizing strategies. Connectionist models leave rather less. Their neuron-like structures are poorly equipped to develop minds of their own, although I have interviewed several people who have imagined that *associations* of neural pathways might take on this executive role. Today, the recentralization of emergent discourse in AI is most apparent in how computer agents such as those designed to sort electronic mail or scour the Internet for news are being discussed in popular culture. As we have seen, the intelligence of such learning agents emerges from the functioning of a distributed and evolving system, but there is a tendency to anthropomorphize a single agent on whose intelligence the users of the program will come to depend. It is this superagent that is often analogized to a butler or personal assistant. The appropriation of decentered views of mind is a complex process; decentering theories are made acceptable by recasting them in more centralized forms, yet even as this takes place, some of the decentered message gets through all the same.

TROJAN HORSES

The Freudian experience taught us that resistance to a theory is part of its cultural impact. Resistance to psychoanalysis, with its emphasis on the unconscious, led to an emphasis on the rational aspect of human nature, to an emphasis on people as logical beings. In the 1970s and 1980s, resistance to a computational model of people led to an insistence that what was essential in humans—love, empathy, sensuality—could never be captured in language, rules, and formalism. In this way, information processing reinforced a split between the psychology of feeling and the psychology of thought. There was a dissociation of affect. The cognitive was reduced to logical process and the affective was reduced to the visceral. But the relationship between thought and feeling is more complex than that. There is passion in the mathematician's theorem and reason behind the most primitive fantasy. The unconscious has its own, structured language that can be deciphered and analyzed. Logic has an affective side, and affect has a logic.

Perhaps the models of human mind that grow from emergent AI might come to support a more integrated view. The interest of psychoanalysts in these models suggests some hope that they might, but there is reason to fear that they will not. In fact, the way emergent AI attempts to include feelings in its models provides some basis for pessimism. Take, for exam-

ple, Marvin Minsky's way of explaining the Oedipus complex in *The Society of Mind*. Minsky sets the stage for the child to develop a strong preference for one parent in cognitive terms. "If a developing identity is based upon that of another person, it must become confusing to be attached to two dissimilar adult 'models.' "[54] For Minsky, Oedipus is simply an adaptive mechanism that facilitates the construction of an unconfused agent by removing "one [of the models] from the scene."[55]

Here, Minsky enters a domain where pure information processing seldom dared to tread, the domain of personality, identity, and subjectivity. The tenor of Minsky's emergent society theory is romantic and impressionistic. But when he actually applies his model, he turns the Oedipal moment (in the psychoanalytic world, thought of in terms of jealousy, sexuality, and murderous emotion) into an engineering fix for a purely cognitive problem. It is more economical and less confusing to have one role model than two so the cognitive unconscious acts to reduce dissonance. Minsky transforms the psychoanalytic consideration of primitive feelings into a discussion of a kind of thinking. Information processing left affect dissociated; emergent AI may try to integrate it but leave it diminished.

As this book is being written, Minsky is working on a book somewhat forbiddingly entitled *The Emotion Machine*. The term refers to both the brain-that-is and the computer-that-will-be. In a recent discussion of that project he acknowledged Freud as "marvelous" and credited Freud's first writings as marking the "birth of psychology." However, when Minsky took up the Freudian topic of pleasure, he gave it a special twist. "Pleasure," said Minsky, "is not an ennobling thing. It is a narrowing thing. You don't have pleasure. Pleasure has you. Pleasure has something to do with keeping you from thinking too many other things when short-term memories are being transferred." Beauty was given a similarly functional definition. It is what "keeps you from being able to find something wrong with something." So if people find a sunset beautiful, it may derive from "our not being a nocturnal animal and it was time to start finding a cave to hide in."[56] I have noted that when Fredric Jameson characterized postmodern thought, he wrote of its tendency to support a "waning of affect."[57] Minsky's theory does nothing if not this, albeit in a way Jameson probably didn't have in mind.

For the foreseeable future, emergent machine intelligence will exist in only the most limited form. But even now, it is providing a rich store of images and metaphors for the broader culture. The language of emergence mediates between technical AI culture and the general psychological culture in a way that the language of information processing did not. The language of neurons, holism, connections, associations, agents, and actors makes it easier for people to consider themselves as that kind of

machine. The similarities between the object language of emergent AI and the object language of psychoanalysis have made AI appear more open to the concerns of the psychoanalytic culture that for so long has been hostile to it.

Some ideas require a Trojan horse for their appropriation—a vehicle in which they can be smuggled into unfriendly terrain. When AI was perceived as synonymous with information processing, it was generally unacceptable to humanists. Now that theorists of emergent AI use the language of biology, neurology, and inner objects to describe their machines, AI begins to seem interesting. Through such intellectual detours, romantic machines may have the effect that critics feared from the "classical," rational ones. For John Searle, no matter what a computer could do, human thought was something else, a product of our specific biology, the product of a human brain. But when connectionist neuroscience begins to revise the boundaries between brains, machines, and minds, it is harder to argue for the specificity of the human mind.

In the 1980s, in response to the computer presence, some people made a split between a mechanical vision of intelligence and an almost mystical vision of emotion. Others emphasized that computers could think but could not achieve the kinds of knowledge that come from being-in-the-world. Today, machines that promise to learn-in-the-world challenge us to invent new hybrid self-images, built from the materials of animal, mind, and machine. In the 1990s, artificial intelligence seems to be suggesting not modernist mind as mechanism but postmodern mind as a new kind of machine, situated somehow between biology and artifact.

In the 1980s, connectionism met the challenge posed by the romantic reaction to information processing AI by agreeing with the statement that people do not operate by simple preprogrammed rules. Connectionism simply added, "And neither do intelligent machines." But our desire to be something other than just machines did not disappear as we began to accept emergent views of both human and computer minds. By the late 1980s, the boundary between people and computers had been displaced in two ways. First, it was displaced away from thought to emotion. Computers might think, but people could feel. This displacement, however, was obstructed by dramatic advances in psychopharmacology, which suggested that the processes that underlie human motions are fairly "mechanical," predictable, and controllable. So while it remained true that computers don't have emotions, there grew up an increased uncertainty about what it means to say that people have them. This development made a second boundary displacement all the more important. People's sense of difference from computers shifted away from the domain of intelligence to the domain of biological life. Computers were accepted as intelligent, but people were special because they were alive.

When AI offered a rational and rule-driven machine, it led to a romantic reaction. Current romantic reconceptualiziations of the machine may now be supporting a rationalist reaction: a too-easy acceptance of the idea that what is essential about mind can be captured in what makes us akin to mechanism. In the past decade, our culture has more readily accepted the idea that human and machine minds might be similar. And for the moment, the question of intelligence is no longer the issue around which the border between people and objects is being contested. Now, people are more likely to distinguish themselves from machines by invoking biology. Our bodies and our DNA are becoming our new lines in the sand. The heat of the battle is moving to the issue of life.