

Initially, I was confident that formal artificial intelligence systems could achieve strong AI. Modern computers, following only a very limited set of formal bit-pushing rules, have memory, can perform mathematical and logical procedures almost instantaneously, and can run programs that seem very flexible. My reasoning seemed straightforward: intelligence arises from the action of the brain, which is composed of neurons, each of which is formally controlled by the formal laws of physics and chemistry. Since we observe intelligence arising from one formal system, then, it is conceivable that it could arise from a formal AI system. In fact, if one were to model all the formal interactions of an entire brain's worth of neurons in a computer program, this simulated brain would be just as capable of intelligence as an actual one.

The AI community's collective experience, though, has shown that formal approaches to AI, what's called good old-fashioned AI (GOF AI), show very little promise of achieving humanlike intelligence. This is not simply an observation that strong AI is not a realistic outcome to expect from GOF AI research in the near future. Instead, it seems that there is a systematic shortcoming in the GOF AI approach that will prevent it from ever achieving strong AI, no matter how much computing power becomes available or how extensively research in this field develops. Recent investigation, therefore, has largely turned away from GOF AI, pursuing other approaches instead. Parallel distributed processing (PDP) techniques, in particular, have been successful in overcoming problems with relative ease that proved to be some of the most challenging for AI.

Formal AI systems may be experts at logic and calculation, but when attempting independent real-world tasks, such as navigation, visual information processing, and face and speech recognition, they not only require massive computing power and not insignificant processing time, the quality of their performance is inconsistent at best. PDP systems, on the other hand, handle these activities with consistent success using minimal computing resources. The PDP

technique uses what are called neural networks, interconnected networks of “neurodes,” simplified representations of the neurons found in animals. Neural network artificial intelligence systems (NNAIs) can be “trained” through a fairly simple process akin to classic psychological conditioning: after presenting the NNAI with sample stimuli, the researcher reinforces desired reactions and discourages undesired ones. Rather than a system of rewards and punishments, reinforcement for neural networks involves back-propagation, a subtle but systematic adjustment of mathematical parameters that dictate how a given neurode responds to its inputs.

Neural networks are no more physically manifested than GOFAI computer programs; the structure of the entire network and all the neurodal interactions occurring within it are virtually represented within computer software. Because of this, I expect John Searle would still protest that neural networks are intrinsically unable to generate strong AI, as they are still technically subject to the Chinese Room argument. However, though the behavior of neural networks is technically formal, in that each individual element follows explicit rules, GOFAI and PDP differ fundamentally in the level at which formal manipulation is intended to occur. GOFAI is based on the concept of symbol manipulation; for a GOFAI system, concepts are represented explicitly as symbols, which are manipulated according to the formal rules of logic and mathematics. Though the mechanics of this technique may involve further formal manipulation of electronic bits, the level at which meaning is assigned and dealt with is that of concepts and symbols. In contrast, the PDP approach manipulates only signals formally, leaving the handling of concepts to emergent behavior. NNAIs are essentially the realization of my conjectured “simulated software brain;” however, the level at which formal control is exercised distinguishes PDP from GOFAI.

One of the significant goals of AI research, besides gaining deeper insight into the functioning and nature of the human mind, is correcting the shortcomings of human intelligence. The ultimate application for an ideal AI is essentially to create robotic slaves: machines that can be

as independently intelligent as humans, but over which humans still exert ultimate control. For example, this sort of control is illustrated by Isaac Asimov's Three Laws of Robotics, from his 1942 short story "Runaround":

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Whether doing housework (à la *The Jetsons*) or planetary exploration (the Mars rover, for example), humans desire an AI that is as intelligent as a human at some tasks, but not so intelligent that it can refuse to do them. Furthermore, our dream AI would be more intelligent than us in ways we can control, perhaps with more complex memory than we have, faster thinking, automatic and permanent learning, and a natural integration of logic. All of these (except the revocation of free will, of course) are desirable characteristics that human beings lack, but that would enhance our ability to solve problems and interact with the world.

The trouble with the PDP approach, however, lies in the level at which formalism and human control enter neural networks. Although neural networks have proven far more successful than GOFAI at efficiently overcoming basic hurdles of real-world performance, their behaviors are indirectly trained, rather than directly programmed. Humans have no control over the level of representations in a neural network, if indeed it can be said to exist at all; there is no complete explanation available of *how* NNAs recognize voices or faces, we just know that they *do*. With a PDP-powered robot, then, there is no way to implement Asimov's three laws. We would not understand how the neural net gives rise to the robot's behaviors any more than we understand our own behavior in terms of activity on the neural scale; our knowledge of how neural networks do what they do is just as limited as is our knowledge of how the human brain does what it does. Furthermore, NNAs tend to emulate, in a simplified way, human intelligence, both its strengths—

flexibility, real-world interaction, quick processing—and its weaknesses—unnatural handling of complex logic, slow mathematical reasoning, difficult learning, complicated memory. Our control over the robot would be limited to the same conceptual levels, essentially, as it is over other humans: at the neurodal/neural level, and at the behavior level. We would have no access to that mysterious representational level that we strive to control, and would therefore not be able to write in any explicit controls of its thinking or behavior.

Robot slaves, essentially, are the dream—in fact, the word “robot” comes from the Czech “robota,” meaning “compulsory labor”—but pure PDP can’t give it to us. We can’t get intelligence with GOFAI, and we can’t get the sort of intelligence we want with NNAI. Is there any reason, then, to pursue either of them? Even though current approaches to AI do not appear to be able to provide us with the customizable buddies that we want, they are each very useful for other purposes and are worthwhile to pursue.

Firstly, what I call the “science fair effect” applies for both techniques: humanity’s natural drive to be inquisitive is not to be ignored; the pursuit of knowledge is often its own quite valid justification. Furthermore, each technique promises practical gains, even if intelligent robot servants are out of their league. GOFAI may develop very powerful expert systems, specialized robotics that operate in well-defined and finite environments, and intelligent agents that work in large virtual realms, like search engines. PDP, on the other hand, may yield independently navigating robots for rescue work, mining, space exploration, or other applications, or even simple artificial life forms designed to terraform extraterrestrial planets. NNAI could also be studied to gain insight into human neuropsychology; with this knowledge, it could also be possible to create cybernetic devices: external memory enhancement, augmented, artificial, or interactive vision, or even dream recorders.

Seeing the shortcomings of both of these intellectual traditions, some researchers have suggested that they could be combined to more thoroughly emulate human intelligence. Though work in this direction is just beginning, it may hold the secret to eventually creating the ideal AI. With real-world flexibility as well as powerful logical reasoning, not only would the advanced progeny of this hybrid technique be able to grapple with more complex problems, their designers would likely be able to control their processing at the conceptual level.

As Braitenberg and Copeland discussed, however, sufficiently advanced NNAs may be judged to possess free will, since the complex interactions of even rudimentary neural networks produce unpredictable, chaotic outputs and behaviors that humans cannot completely understand. Whether or not this assessment would be correct is certainly debatable, but if NNAs could truly be said to have free will, the situation would be significantly complicated with regard to GOFAI/PDP hybrid systems. Directly and immutably controlling the decision-making ability of such a system through its GOFAI element would restrict the free will of the PDP portion and the system as a whole, and would therefore be a severe moral injunction.

Science fiction again forecasts the possibilities of an unrestricted hybrid AI: there is the possibility that this approach could yield autonomous robots of higher intelligence than their human inventors. With free will preserved, the opportunistic drives of a collection of these robots could very well decide that humans are impediments to their survival. While a *Terminator*-style war with technology may seem far-fetched, it is something to consider. One way of conceiving of this situation is as a logical extension of biological evolution: survival of the fittest dictates that our offspring, even computerized ones, could overtake our own species if they were more successful in competition for resources than us or if they would win in direct conflict. This is a frightening scenario, but it is difficult on the small scale for any parent to allow its children to replace it. If we

were capable of creating machines that were superior to us, we would be the honored servants of natural history by allowing them to develop independently and coexist with or replace us.