Clark Glymour

Invasion of the Mind Snatchers

Ten years ago I hoped, even expected, that the computational revolution would revive a dying enterprise, philosophy of science, and make it more intelligent and rigorous and insightful and interesting. This book and the recent work that seems to have provoked it have convinced me that my expectations were completely wrong. To judge by this sample, wherever the discipline of philosophy of science has been touched by cognitive science the result has been a zombie — philosophy of science killed dead and brought back to ghoulish, mindless, pseudolife.

"That grumpy guy, at it again," you say. Well, pilgrim, look at what we have from intelligent people who have in the past done good philosophical work:

Paul Churchland likes connectionist computational models of how the brain computes. Three essays in Churchland's recent book, *A Neurocomputational Perspective*[1] ("On the Nature of Theories: A Neurocomputational Perspective," "On the Nature of Explanation: A PDP Approach," and "Learning and Conceptual Change"), seem typical of his recent connectionist crush. They contain accounts of some well-known nonstochastic connectionist systems and accounts of some weight adjustment algorithms. In the first of these three essays Churchland describes some of the reasons why these connectionist models cannot be correct descriptions of how the brain functions if network nodes are identified with cell bodies and network edges with synaptic connections. He even calls for a study of more biologically realistic networks and learning procedures. Unfortunately, he does not propose any, investigate any, or prove anything about any. If he did, I would have no objection to the enterprise. The essays contain no new formal or computational results of any kind. Instead they solve all of the recently popular problems of philosophy of science. What solutions!

What is a theory? Says Churchland: Confused, old-fashioned people think a theory is a body of claims that can be expressed in a language, claims that (normally) are either true or false, and they even foolishly

think the business of science is to find the claims that are true, general, and interesting. Equally benighted people think that humans have attitudes toward propositions and even represent propositions locally in the brain. But connectionism reveals that a theory is no such thing: it is an assignment of weights within a neural network (p. 177). In particular, someone's total theory is an assignment of weights within the neural network of that person's brain. *What is a concept?* The same thing (p. 178). Churchland later changes the story and suggests that a concept is a *partition* of weight space — that is, a region of possible values of weights that for all and only input vectors in a specified class will produce a specific output vector — or perhaps an input/output function (p. 234).

One awaits the time when Churchland's opinion will triumph and theorists will routinely present their heads for examination. Here is a view fit for Lavoisier and Ichabod Crane. A few obvious questions spring to mind: If someone's entire theory is just the "weights" of all of his or her synaptic connections, and if no propositions are represented locally, what is a *part* of an entire theory — electrodynamics, say, or thermodynamics, or the theory of evolution? If theories are as claimed, what is a theory that someone considers, reasons with hypothetically, but does not believe? If that is what a theory is, how do people share theories? If a theory is a pattern of weights, then what is *testing* a theory? If that is what a theory is, then what are people doing when they claim to be arguing about theories — are they arguing about the weights in someone's head? And so on. An obvious assessment of Churchland's *nouveau* philosophy of theories suggests itself: the whole business is an elementary confusion between a theory, or a system of claims, and the physical/computational states someone is in when she or he has a system of beliefs. Astonishingly, you will not find a single one of these questions addressed in Churchland's three essays. Not one. Honest.

Some philosophers hold that evidence and observation are *theory-laden*, whatever that means. I thought it meant that there is no level of description of observations such that one and the same objective circumstance (however individuated, whether by the psychologist, the philosopher, or God) will be reported, at that level of description, in the same way by an observer regardless of his or her prior beliefs. *Is observation theory-laden?* Says Churchland: the connectionist picture shows it is indeed, because there are an astronomical number of different input/output functions — hence "concepts" — that a large network can have (p. 190).

You might wonder, as I do, what the connection is between the issue of theory ladenness and Churchland's remark about the multiplicity of concepts that a network can represent. The two seem so, well, *disconnected.*

Churchland's idea seems to be that an *observation* is a particular pattern of values for a specific set of *output* nodes. If someone has sufficiently different weights, she or he will make different observations from the same input pattern. But then it would seem, contrary to Churchland's story, that the content of an observation is indeed locally represented by the values of a specific set of nodes. And, that aside, in would seem that in connectionist systems there is a level of description of observation that is not theory-laden, namely the values of the "input" nodes. But of course Churchland cannot really talk intelligibly about theory ladenness at all, because that issue is about *descriptions*, and in his view (not mine) networks do not describe anything because they do not say or claim anything.

Churchland's three essays contain more like this: for example, a parallel account of "inference to the best explanation." If, however, you ask where methodology fits in — e.g., why and in what sense inference to the best explanation can form part of a reliable method — you will find no answer. You will not even find the question recognized. If you look for a contribution to methodology you will find none; and if you look for any new results about the power and limits of the computational picture for which Churchland is an enthusiast, there too you will find nothing new. This in an area rocking and seething with new ideas and real problems.

In the last twenty years there has been considerable study in psychology of patterns of human irrationality. The studies include limitations on the reliability of human judgments of logical relations, judgments of probability, judgments of causality, and limitations in the reliability of experts of various kinds. The real source of this work lies in the pioneering studies done by Paul Meehl — Ronald Giere's colleague — in the 1950s on the reliability of clinicians. Herbert Simon's analysis of corporate behavior is in the same spirit. According to Simon, corporations do not act like ideally rational agents; they do not maximize profits or expected profits. Instead they try to make enough profit to keep up with the competition, please the directors, obtain bonuses for the management, etc. In Simon's phrase, they *satisfice.*

Giere's *Explaining Science*[2] tries to apply the idea of satisficing to scientific practice. Individual scientists, he claims, do not maximize their expected utility; they satisfice. The trouble with Giere's starting point is that while there is formal structure implicit in the idea of maximizing expected utility, there is scarcely any in the idea of satisficing; it functions as a phrase partly marking the idea that an agent may be unable to do what is ideally best (when cognitive limitations are ignored) and partly marking the idea that an agent may be unwilling to give over

the cognitive resources required to determine what is normatively best. The interesting questions are the general forms these irrationalities take, and what can be done to overcome them. The first question is explored by those who follow Meehl through the work of Kahneman, Tversky, Dawes, and others, the second by a wide range of contemporary work in artificial intelligence. So there is plenty of serious work to do along the general line Giere takes. The trouble is he does not do any of it.

The oddity of Giere's book is not in the content but in the speech act. Indiana University, where he worked for many years, has a cyclotron facility, and Giere took advantage of its location to gather evidence of the obvious. One chapter of the book argues that in designing experiments with the cyclotron, physicists act and reason as if particle theory were true. Another long chapter examines the fortunes of the phenomenological Dirac equation in accounts of scattering and, in particular, the work of a particular physicist, Bunny Clark. The philosophical claims are, first, that scientists prefer being right to being wrong, and second, that the choices they make about what to work on are influenced by their "cognitive resources," e.g., by the theories they have experience with, the mathematical methods they know how to use, their ability to write computer programs. No kidding.

This is not quite the same as setting out to gather evidence that $2+2=4$, but it is close. Shorn of the philosophical trappings Giere's book is sustained by human interest, just as, say, *The Soul of a New Machine* is sustained by human interest. The interweaving of science and human interest is very well done, and so far as I can tell Giere gets the relevant science right. I have nothing against intellectual journalism. I find it more than curious that it should pass as philosophical research.

Paul Thagard's ECHO program is of a genre. I think the genre is dog and pony show. Thagard's computer program, ECHO, is supposed to measure "explanatory coherence." The idea is that propositions cohere with each other and with the data because of their explanatory and analogical relations. Scientists, it is claimed, prefer (ought to prefer? — too subtle a distinction, that) coherent systems of propositions. The principles of coherence Thagard gives are these:

1. Coherence is symmetric and so is incoherence.
2. If P1 ... Pm explain Q, then Pi and Q cohere and Pi, Pj cohere for i, j from 1 to m, and the degree of coherence is inversely proportional to m.
3. If P1 explains Q1 and P2 explains Q2 and P1 is analogous to P2 and Q1 is analogous to Q2, then P1 and P2 cohere and Q1 and Q2 cohere; if the circumstances are otherwise the same but P1 is disanalogous to P2, then P1 and P2 incohere.
4. Observational propositions have a degree of acceptability of their own.
5. If P contradicts Q, then P and Q incohere.
6. The acceptability of a proposition P in a system S depends on its coherence with the propositions in S, and if many results of relevant experimental observations are unexplained, then the acceptability of a proposition P that explains only a few of them is reduced.
7. The coherence of a system of propositions is a function of the pairwise coherence of its members.

This is more substance than in the first two examples, but as a piece of philosophical analysis one would object: (1) all the hard questions — what is explanation? what is a better or a worse explanation? what is analogy? how is consistency to be maintained by a computationally bounded system? — have been begged; (2) the account of the dependence of acceptability on coherence is vague; (3) there is no argument for the account as any kind of norm; (4) there is no empirical evidence that the account is correct as a description of anything. The program is supposed to address some of these questions.

Thagard's program takes as *input* a list of "facts" of the following sort: "Q1 is observed to be the case"; "P1 ... Pn explain Q"; "P1 ... Pm explain Pn"; "Pj and Pk are inconsistent"; "Pk and Pm are analogous." The operator, Thagard or his assistant, specifies all of these facts; the system does nothing to determine what explains what, or what is analogous to what, or what is inconsistent with what. Thagard gives a number of examples of historical cases in which he gives his program "facts" about explanation and inconsistency and alternative hypotheses and the program finds the theory that was historically preferred. Copernicus wins out over Ptolemy, Darwin over creationism, etc. These examples are supposed to be the empirical evidence for the theory incorporated in the program.

What is wrong with this work? Here is a start:

If the aim is to describe how humans produce judgments about the best explanation given facts about what explains what, there are no arguments for the particular program. As a description, the ECHO picture assumes that theory comparison has two distinct modules, one for judging what explains what and another for judging "coherence" of separate propositions. There is neither empirical nor historical evidence for such

an assumption. The picture assumes that in judging what to believe, all explanations of observations are equally important; all that matters is their numbers. There is no evidence for that claim, either. There is no psychological case at all.

In their paper for this collection, Nowak and Thagard give one argument that is so bad it must be disingenuous. They consider an incredibly simple counting procedure proposed by Jerry Hobbs as an alternative to ECHO, and they claim that Hobbs's count does not give the intuitive answer in some cases. Therefore, Nowak and Thagard conclude, ECHO is "necessary." Grant the premises, and consider the form of the argument, which is worthy of "Star Trek":

If ECHO then P
If Hobbs procedure then ~P
Therefore, ECHO is necessary for P.

(Which reminds me of a story once told me by Hartry Field about an unnamed distinguished philosopher, said to be from Pittsburgh, who was informed in private by a junior colleague that one of the distinguished professor's forthcoming papers turned on a modal fallacy. "Tell me," the distinguished professor is said to have asked his young colleague, "is it a *well-known* modal fallacy?")

Most of what the ECHO program does is simple counting, and one could think of lots of other easy ways to count besides Hobbs's. Any number of alternatives suggest themselves that are much simpler than ECHO and could be implemented on a pocket calculator.[3]

The historical simulations are bogus experiments. What do they test? Thagard and his assistants get to choose the "facts" and count them, and they choose and count them so that ECHO gets the right answer. I assure you that if ECHO were given as a separate fact to be explained that each star does not show parallax, Ptolemaic theory would fare much better; anti-Darwinians could (and did) cite a myriad of facts that Darwin's theory could not explain, but their objections do not show up in Thagard's input. The experiments show nothing about any capacity of the ECHO program to get to the truth reliably, and they show nothing about the unique capacity of the program (as against some simple function whose implementation would scarcely cost military research or the McDonnell Foundation a dime) to reproduce conventional judgments in the examples Thagard uses.

If the aim of this book is to show how philosophy of science can be integrated with cognitive psychology and artificial intelligence, then I think its point is made. There is plenty of work in cognitive science that is more enthusiasm than substance; there is plenty of work in cognitive psychology that tells you no more than your grandmother could, and probably less; there is plenty of work in artificial intelligence that is dog and pony show. Giere, Churchland, and Thagard each present examples of work that could mingle indistinguishably with the worst of cognitive psychology and artificial intelligence. One hopes their examples will not become the standard for the union of cognitive science and philosophy of science. There are philosophers whose work could mingle with the best of cognitive science and computer science. Maybe we need a grant to discover what makes some philosophers immune to the invasion of the mind snatchers.

Notes

A fellowship from the John Simon Guggenheim Memorial Foundation afforded time to write this essay. I am grateful to Ronald Giere for his courtesy and consideration in publishing it.

1. Cambridge, Mass.: MIT Press, 1989.
2. Chicago: University of Chicago Press, 1989.
3. For example, keeping close to the spirit of Thagard's program, one could define:

E(P1P2) = the set of all tuples P1, P2 ... Pm, Q such that P1, P2 ... Pm explain observation Q;

N(P1P2) = the set of all tuples P1, P2 ... Pm, Q such that P1, P2 ... Pm explain not-Q, and Q is observed;

For $s = <P1, P2 ... <Pm, Q>$ in E(P1P2) or N(P1P2), L(s) = m

$$\sum_{s \in E(P1P2)} \frac{1}{L(s)} - \sum_{s \in N(P1P2)} \frac{1}{L(s)}$$

And take the "activation" or "acceptance" of a proposition to be the sum of its coherences with other propositions in the system. Clearly a lot of other simple functions are possible; for example, with simple recursion or iteration one could (as Thagard wants) give P1 and P2 some degree of coherence if they explain a hypothesis that explains a hypothesis ... that explains an observation; the degree of coherence contributed by such a relation could be weighted by the number of explanatory steps between P1, P2, and the observation. You could write the particular function illustrated above in a few minutes on a programmable pocket calculator.