DANIEL C. DENNETT

The Logical Geography of Computational Approaches: A View from the East Pole

Westward the course of empire takes its way.
—BERKELEY

With many different people claiming to be explaining the mind in "computational" terms, and almost as many denying that this is possible, empirical research and ideological combat are currently proceeding on many fronts, and it is not easy to get one's bearings. But some themes are emerging from the cacophony, and they tempt me to try to sketch the logical geography of some of the best-known views, with an eye to diminishing the disagreements and misrepresentations that sometimes attend them.

There are still dualists and other mystics in the world who assert (and hope and pray, apparently) that the mind will forever elude science, but they are off the map for me. A goal that unites all participants in the conflict area I will explore is the explanation of the aboutness or intentionality of mental events in terms of systems or organizations of what in the end must be brain processes. That is, I take it as agreed by all parties to the discussion that what we want, in the end, is a materialistic theory of the mind as the brain. Our departure point is the mind, meaning roughly the set of phenomena characterized in the everyday terms of "folk psychology" as *thinking about* this and that, *having beliefs about* this and that, *perceiving*

this and that, and so forth. Our destination is the brain, meaning roughly the set of cerebral phenomena characterized in the *non*intentional, *non*symbolic, *non*-information-theoretic terms of neuroanatomy and neurophysiology. Or we can switch destination with departure and construe the task as building from what is known of the plumbing and electrochemistry of the brain toward a theory that can explain—or explain away—the phenomena celebrated in folk psychology. There has been a surfeit of debate on the strategic question of which direction of travel is superior, top-down or bottom-up, but that is now largely behind us and well understood: obviously both directions can work in principle, both have peculiar pitfalls and opportunities, and no one with an ounce of sense would advocate ignoring as a matter of principle the lessons to be learned from the people moving from the opposite end.

A much more interesting clash concerns what to look for in the way of interstitial theory. It is here that manifestos about "computation" vie with each other, and it is this issue I will attempt to clarify. Consider the extreme positions in their purest forms.

First, there is what I shall call *High Church Computationalism,* which maintains that intervening between folk psychology and brain science will be at least one level of theory quite "close" to the high level of folk psychology that is both "cognitive" and computational." The defining dogmas of High Church Computationalism (HCC) are a trinity:

(1) *Thinking is information processing*. That is, the terms of folk psychology are to be spruced up by the theorist and recast more rigorously: "thinking" will be analyzed into an amalgam of processes ("inference" and "problem solving" and "search" and so forth); "seeing" and "hearing" will be analyzed in terms of "perceptual analysis" which itself will involve inference, hypothesis-testing strategies, and the like.

(2) *Information processing is computation (which is symbol manipulation).* The information-processing systems and operations will themselves be analyzed in terms of processes of "computation," and since, as Fodor says, "no computation without representation," a medium of representation is posited, consisting of *symbols* belonging to a *system* which has a *syntax* (formation rules) and *formal rules of symbol manipulation* for deriving new symbolic complexes from old.

(3) the semantics of these symbols connects thinking to the external world. For instance, some brain-thingamabob (brain state, brain event, complex property of brain tissue) will be the symbol for MIT, and some other brain-thingamabob will be the symbol for budget. Then we will be able to determine that another, composite brain-thingamabob refers to the MIT budget, since the symbolic structures composable within the representational medium have interpretations that are a systematic function of the semantic interpretations of their elements. In other words, there is a language of thought, and many of the terms of this language (many of the symbols manipulated during computation) can be said to *refer* to things in the world such as Chicago, whales, and the day after tomorrow.

At the other extreme from the High Church Computationalists are those who flatly deny all of its creed: there is no formal, rule-governed, computational level of description intervening between folk psychology and brain science. Thinking is something going on in the brain all right, but is not computation at all; thinking is something holistic and emergent—and organic and fuzzy and warm and cuddly and mysterious. I shall call this extreme version Zen Holism.[1]

In between these extremes are all manner of intermediate compromise positions, most of them still rather dimly envisaged at this inchoate stage of inquiry. It would be handy to have a geographical metaphor for organizing and describing this theory-space, and happily one is at hand, thanks to Fodor.

In a heated discussion at MIT about rival theories of language comprehension, Fodor characterized the views of a well-known theoretician as "West Coast"—a weighty indictment in the corridors of MIT. When reminded that this maligned theoretician resided in Pennsylvania, Fodor was undaunted. He was equally ready, it turned out, to brand people at Brandeis or Sussex as West Coast. He explained that just as when you are at the North Pole, moving away from the Pole in any direction is moving south, so moving away from MIT in any direction is moving West. MIT is the East Pole, and from a vantage point at the East Pole, the inhabitants of Chicago, Pennsylvania, Sussex, and even Brandeis University in Waltham are all distinctly Western in their appearance and manners. Boston has long considered itself the Hub of the Universe; what Fodor has seen is that in cognitive science the true center of the universe is across the Charles, but not so far upriver as the wild and woolly ranchland of Harvard Square. (To a proper East Pole native, my outpost farther afield at Tufts is probably imagined in terms of crashing surf and ukuleles.)

Since MIT is the Vatican of High Church Computationalism, and since the best-known spokesmen of Zen Holism hold forth from various podia in the Bay area, I propose to organize the chaos with an idealized map: positions about computational models of mental phenomena can be usefully located in a logical space with the East Pole at the center and the West coast as its horizon. (This is not, of course, just what Fodor had in mind when he discovered the East Pole. I am adapting his vision to my own purposes.) In between the extremes there are many positions that

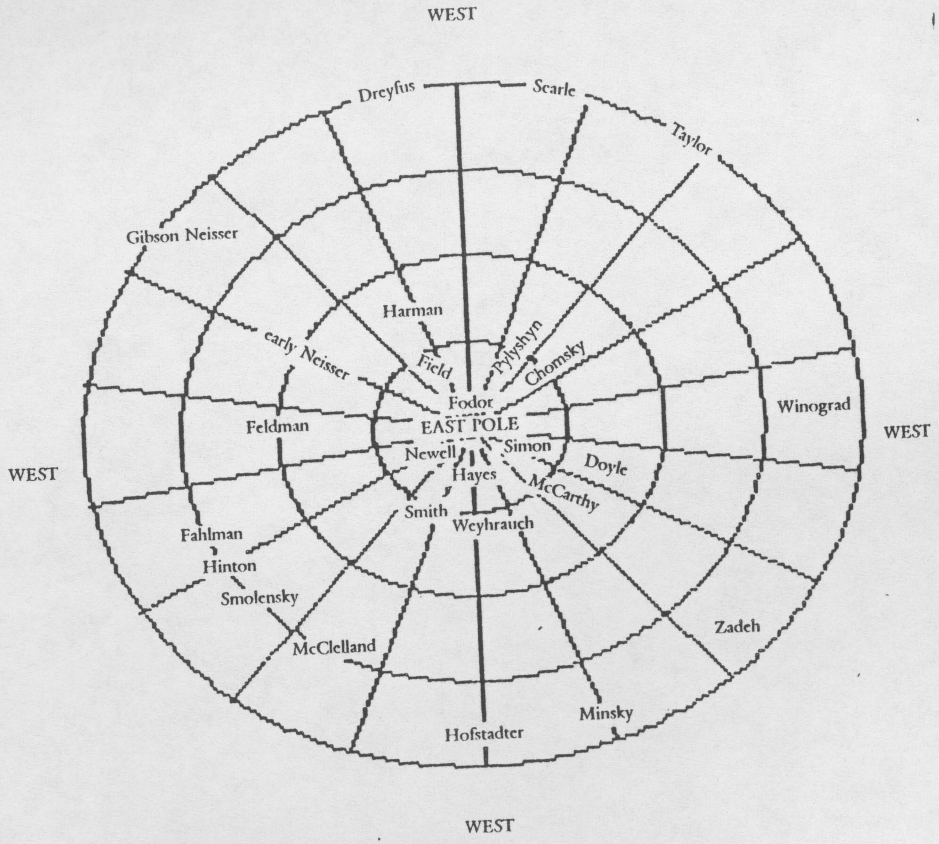

Fig. 1. A view from the East Pole.

disagree sharply over many matters (they are, as one says, "diametrically opposed"), but can nevertheless all be seen to be more or less Western, depending on which denials or modifications of High Church Computationalism they defend. As in any attempt at cartography, this is just one of many possible projections, claiming no essential rightness but inviting your consideration as a useful organizer.[2]

These warring doctrines, High Church Computationalism and its many heresies, are not themselves theories; they are ideologies.[3] They are ideologies about what the true theory of the mind will or must be like, when we eventually divine it. Various attempts to create genuine theories—various research programs—*seem* to be committed to various ideologies arrayed in our space, but as we shall see, the bond between research program

and ideology is rather loose. In particular, the fact that great progress is (or is not) being made on a research program might tell us next to nothing about the ultimate soundness of its inspiring ideology.

And vice versa: refutation of an ideology sometimes bodes not very much at all for the research done under its banner. Ideologies are important and even unavoidable; they affect how we imagine the issues, and how we express the questions. A false ideology will typically tempt us to frame the wrong questions and thereby waste time and effort on low-grade research and avoidable artifactual puzzles. But sometimes one can make progress even while asking awkward and misguided questions, and sometimes (quite often, in fact, I think) researchers half-consciously know better than actually to ask the questions their ideology holds to be the right questions. Instead, they ask questions they can see how to answer and hardly notice that these inquiries are rather remotely related to the official questions of their school of thought.

Not surprisingly, it is philosophers who have been the most active formulators and guardians of the ideologies. Jerry Fodor, in *The Language of Thought* (1975) and *RePresentations* (1981), has for some time been the theologian in residence at the East Pole [his more recent heresy in *The Modularity of Mind* (1983) will be discussed in due course]. Hubert Dreyfus and John Searle at Berkeley are the gurus of West Coast Zen Holism. Hartry Field (1978) is another East Polar apologist, and so is Gilbert Harman (1973), though he is not entirely Orthodox.

Joining Dreyfus and Searle on the West Coast is Charles Taylor.[4] Other philosophers range in between: Stephen Stich, Robert Cummins, John Haugeland, and Margaret Boden, to name a few.[5] (For my part, I have always considered myself bi-coastal.[6]) Philosophers are not the only major participants in this ideological conflict however. Allen Newell,[7] Noam Chomsky,[8] and Zenon Pylyshyn[9] have staunchly represented the East, while Terry Winograd,[10] Lotfi Zadeh, and Douglas Hofstadter are non-philosophers in the West who have contributed more than passingly to the formulation of doctrine.[11] And in psychology we have for instance the Apostate Ulric Neisser, whose book, *Cognitive Psychology* (1963), was a founding document of High Church Computationalism, but who, under the influence of J. J. Gibson, renounced the faith in his book, *Cognition and Reality* (1975), and helped to turn Ithaca into something of a West Coast colony.[12]

Real-world geography is obviously only an intermittent clue to logical geography. The San Diego school of Norman, Rumelhart, and McClelland is appropriately Western in its attitude, but now that McClelland is joining the rather Middlewestern group of Hinton and Fahlman at Carnegie–Mellon, Pittsburgh (given the Dreyfusian and hence Coastal sym-

pathies of John Haugeland) would begin to look like another Western colony were it not for the counterbalancing Eastern voices of Newell and Simon (and Jon Doyle). Jerry Feldman at Rochester is a founding member of the distinctly Western school of New Connectionists (of which more later), but his colleague Patrick Hayes (who once told me he was quite sure the brain conducts its business in the predicate calculus) is about as East Pole as you can get. John McCarthy at Stanford is also an East Pole missionary, of course, as are Richard Weyhrauch and Brian Smith, in spite of their real-world locations.

All that should ruffle a few feathers. I doubt that the points of agreement between, say, Searle and Hofstadter, or Winograd and Feldman, loom very large in their minds, nor do I suppose Chomsky and Fodor are all that comfortable being lumped with McCarthy and Simon and vice versa. And rightly so, for the defining issues that place these people in the same "latitudes" are less integral to their widely differing research programs and methodological principles than they have acknowledged. So my cartography is indeed in some regards superficial. The dependence of some researchers on the dogmas of High Church Computationalism can be made to loom large at first glance, but I hope to show that it dissolves under scrutiny.

Over the years there has been an intermittent but intense focus on Fodor's and Chomsky's grounds for taking High Church Computationalism (HCC) seriously in their research programs in linguistics, psycholinguistics, and cognitive psychology, and I consider the issues raised in that focus to be sufficiently well known and analyzed that I need not review them here. Instead I will draw attention to how some other research programs in cognitive science *apparently* broaden the support for—or reveal the breadth of the dependence on—HCC.

Consider John McCarthy's quest for a *formal representation* of the knowledge an agent requires in order to *prove* that various courses of action are best (or just acceptable) under various circumstances.[13] This is, as McCarthy says, an epistemological question, and the formality constraint McCarthy imposes on the answer *apparently* has the same rationale as Fodor's formality constraint:[14] the brain, as a mechanism, can respond only to the formal (not semantical) properties of its states.

Or consider Newell and Simon's quest for the *rules* of problem-solving strategies adopted by self-conscious, deliberate human problem solvers, and their characterization of the phenomenon of problem solving as a transition between one (formally explicit) representation of the problem and another (formally explicit) representation of the solution.[15] Doesn't their empirical exploration of problem solving cast in these terms presuppose a commitment to the hypothesis that problem solving in human beings is a computational process taking explicit, formal symbol structures into other explicit, formal symbol structures? They often say as much, and when they do, they express what is in any case the common understanding about the whole point of their research program.

In fact, from one vantage point all AI seems unproblematically committed to HCC. If your models are written in LISP and are actually designed to run on computers, how can you take yourself seriously as a theorist or modeler of mental processes without taking yourself to be presupposing the thesis—or at least playing the hunch—that mental processes are analogous to the constituent computer processes of your model at least to the extent of being formal, computational processes of symbol manipulation?

Many cognitivist theorists have been content to avow just such an ideology. After all, what's wrong with it? This is the question that has thrown down the gauntlet to the ideological foes of HCC, who have been so concerned to demonstrate the shortcomings of HCC doctrines *as ideology* that they have seldom cared to ask whether the research programs of the proponents are as deeply committed to the doctrines as the proponents have maintained. (Mightn't the manifestos be more a matter of fancy advertising jingles than enabling assumptions? Your true-blue ideologue doesn't care; all that matters is under what conditions the advertising is *defensible*.)

Thus a recurrent and powerful line of criticism of High Church Computationalism points out that such computational models as have actually been proposed by workers in AI or cognitive psychology are ludicrously underdetermined by the available data, even when they are quite plausible, as they often are.[16] This criticism usefully reveals how far from being demonstrated the central claims of High Church Computationalism are, but otherwise it strikes a glancing blow, since if there happen to be deep methodological reasons for hoping for a winning computational model, the prospect that early exploratory models will be drastically undetermined by the data should be viewed as a tolerable (and entirely anticipated) phase of the research program.

And Fodor has provided us with a candidate for a deep methodological reason for throwing our lot with HCC: it is the only remotely explicit positive idea anyone has.[17] Fodor's challenge ("What else?") has very effectively embarrassed the opposition for years, since Zen Holism itself is not a positive alternative but only a brute denial. Saying that thinking is holistic and emergent only announces the flavor of your skepticism and gestures in the direction of an alternative.

In the absence of plausible, explicit alternatives and faced with the drastic underdetermination of any HCC theory, ideologues and critics have been lured into a startlingly premature debate on what *would be* good evi-

dence for one brand or another of computational theory. I'm all for thought experiments, but in this instance it seems to me that things have gotten out of hand. While scarcely measurable progress is being made on garnering real evidence for any particular computational theory or model,[18] tremendous effort has been expended on reaching and defending verdicts on imagined future evidential circumstances. (I have played this game as exuberantly as others—I do not exempt myself from these second thoughts.)

But we can ask the challenging question again: what is the alternative? That is, what else ought the troubled skeptic do, faced with an implausible and underdetermined ideology that defends itself with the challenge: what is the alternative? Dreyfus and the other West Coast philosophers have taken the extreme course: they have attempted to find a priori arguments showing that HCC *couldn't possibly* be true—without notable success. They have formulated their arguments but won few converts with them, and the verdict of many onlookers is that the debate conducted in those terms is a standoff at best.

If the a priori gambit has been overdone, there is a more modest Western tactic that has seldom been adopted by philosophers but has been quite influential in some AI circles: trying to explain not why HCC is impossible, but why, even if it is both possible (for all one can tell) and the only articulated possibility to date, it is so unlikely.

High Church Computationalism does seem to me (and to many others) to be highly implausible, for reasons that are hard to express but that hover around the charge that a computational, symbol-manipulating brain seems profoundly unbiological.[19] This unelaborated suspicion should not be trusted, for one's intuitions about what is biological and what is not are (for most of us, surely) an undisciplined crew. What could seem more unbiological (from one intuitive vantage point) than the clockworky mechanisms of DNA replication, for instance? So if this is to be more than just another way of saying Nay to NCC, we need to say something more explicit about why we think an HCC-style brain would not be Nature's Way.

Douglas Hofstadter has recently found a way of expressing this misgiving that strikes me as being on the right track.[20] HCC systems, designed as they are "through a 100% top-down approach" (p.284), are *too efficient* in their utilization of machinery. As we work our way down through the nested black boxes, "functions calling subfunctions calling subfunctions," decomposing larger homunculi into committees of smaller, dumber homunculi, we provide for no waste motion, no nonfunctional or dysfunctional clutter, no featherbedding homunculi or supernumeraries. But that is not Nature's Way; designing systems or organizations with that sort of

efficiency requires genuine foresight, a *detailed* anticipation of the problem spaces to be encountered, the tasks the system will be called upon to perform. Another way of saying it is that such systems, by being designed *all the way down*, have too much intelligence implicated in their design at the lower levels.

Nature's Way of providing flexibility and good design involves a different kind of efficiency, the sort of efficiency that can emerge opportunistically out of prodigious amounts of "wasteful" and locally uninterpretable activity—activity that isn't from the outset "for" anything, but is enlisted to play some very modest role (or many roles on many different occasions) in some highly distributed process.

This is a theme to counterbalance the themes of HCC that have so dominated imagination, but is it just a theme? Until very recently, Fodor's challenge stood unanswered: no one had any explicit proposals for how such bottom-up systems could do any recognizable cognitive work. The only suggestions forthcoming from the philosophers (and neuroscientists as well) were metaphorical and mysterious.[21]

But now from out of the West something better is coming. Explicit proposals, and even working, testable models are emerging from a variety of workers clustered around the so-called New Connectionism. (I am still only beginning my attempt to map out the relations between these kindred spirits and no doubt will leave out, misinclude, and misrepresent some people in the course of providing my introductory list, but since this paper cannot wait for a year, I will have to present my half-formed reflections.)

The most compelling *first* impression of the New Connectionists (and the point of their name) is that they are looking closely at neural architecture and trying to model much closer to the brain than the mind. That is, if East Pole AI programs appear to be attempts to *model the mind*, New Connectionist AI programs appear to be attempts to *model the brain*. And some of the purer or more extreme approaches feature explicit commentary on the parallels between neurons or neuron assemblies and the functional units of their models.[22] But it is a mistake, I think, to read the movement as "neurophysiology carried on by other means."[23] Nor is the distinctive difference simply or mainly a matter of being much more bottom-up than top-down.[24] For whereas specifically brainish-looking bits and pieces and assemblies do often appear in these new models, what is more important is that at a more abstract level the systems and elements—whether or not they resemble any known brainware—are of recognizable biological types.

The most obvious and familiar abstract feature shared by most of these models is a high degree of parallel processing, either simulated or based on actual parallel hardware.[25] Although the point has been brought home to

everybody by now that the brain is a massively parallel processor and that this is important in understanding how the mind's work is done by the brain, there is less interest among the New Connectionists in the question of just what kind of parallel processor the brain is than in what the powers of massively parallel processors in general are. Hence some of the parallel-processing models are almost willfully "unrealistic" as models of brain organization. For instance, one of the guiding analogies of Hofstadter's Jumbo architecture is the constructing of molecules by enzymes floating freely within the cytoplasm of a cell—but of course Hofstadter doesn't think the cognitive tasks the Jumbo architecture is designed to perform (the example exploited in the exposition and testing of the architecture is solving anagrams) are performed within the cell bodies of people's brain cells![26]

Another widely and diversely used New Connectionist idea derives from statistical mechanics: "simulated annealing."[27] Computational analogues of alternatively "warming" and "cooling" structures to get them to settle into the best combinations have proven to be powerful new methods in several different domains.[28]

Although there is a lot of diversity and disagreement among the people in my Western cluster around the New Connectionists, a few characteristics—family resemblances—are worth noting. In these models, typically there is:

(1) "distributed" memory and processing, in which units play multiple, drastically equivocal roles, and in which disambiguation occurs only "globally." In short, some of these models are what you might call computational holograms. For instance, Pentti Kanerva's distributed recognition memory[29] has a strictly limited capacity for high-quality memory, but when it is overloaded, the effect is not to create a simple overflow in which no new information can be input. Rather, the input of too much information leads to the partial degradation of information previously stored; the superimposition of the excess information smudges or obscures the information already in memory.[30]

(2) no central control but rather a partially anarchic system of rather competitive elements. (See, e.g., the discussion in Feldman and Ballard of "winner take all" or WTA networks. Many of these ideas can be seen to be new versions of much older ideas in AI—e.g., Selfridge's Pandemonium, and of course perceptrons.)

(3) no complex message-passing between modules or subsystems. [For instance, no discursive messages "about the outside world." "The fundamental premise of connectionism is that individual neurons *do not transmit large amounts of symbolic information*. Instead they compute by being *appropriately connected* to large numbers of similar units" (Feldman and Ballard 1982, p.208).]

(4) a reliance on statistical properties of ensembles to achieve effects.

(5) the relatively mindless and inefficient making and unmaking of many partial pathways or solutions, until the system settles down after a while not on the (predesignated or predesignatable) "right" solution, but only with whatever "solution" or "solutions" "feel right" to the system. This combines the idea of simulated annealing (or a close kin of it) with the idea that in nature not all "problems" have "solutions" and there is a difference between a process stopping and a process being turned off.

The models being explored are still computational, but the level at which the modeling is computational is much closer to neuroscience than to psychology, What is computed is not (for instance) an implication of some predicate-calculus proposition *about Chicago*, or a formal description *of a grammatical transformation*, but (for instance) the new value of some threshold-like parameter of some element *which all by itself has no univocal external-world semantic role*. At such a low level of description, the semantics of the symbolic medium of computation refers only (at most) to events, processes, states, addresses within the brain—within the computational system itself. In short, on this view the only formal, *computational* "language of thought" is *rather* like a machine language for a computer, and you can't say "it's raining in Chicago" in machine language; all you can express are imperatives about what to do to what contents of what address and the like.

How then do we ever get anything happening in such a system that is properly *about Chicago*? On these views there must indeed be higher levels of description at which we can attribute external-semantical properties to brain-thingamabobs (this brain-thingamabob refers to Chicago, and that one refers to MIT), but at such a level the interactions and relationships between elements will not be computational but (and here we lapse back into metaphor and handwaving) statistical, emergent, holistic. The "virtual machine" that is recognizably psychological in its activity will not be a *machine* in the sense that its behavior is not formally specifiable (using the psychological-level vocabulary) as the computation of some high-level algorithm. Thus in this vision the low, computational level is importantly *un*like a normal machine language in that there is no supposition of a direct translation or implementation relation between the high-level phenomena that do have an external-world semantics and the phenomena at the low level. If there were, the usual methodological precept of computer science would be in order: ignore the hardware since the idiosyncracies of its particular style of implementation *add nothing* to the phenomenon,

provided the phenomenon is rigorously described at the higher level. (Implementation details do add constraints of time and space, of course, which are critical to the assessment of particular models, but these details are not normally supposed to affect *what information processing is executed*, which is just what makes this Western proposal a break with tradition.)

My favorite metaphor for this proposal is meteorology. (What would you expect from the author of *Brainstorms*? But the analogy is developed in detail in Hofstadter's *Gödel, Escher, Bach,* pp.302–309.) Think of meteorology and its relation to physics. Clouds go scudding by, rain falls, snowflakes pile up in drifts, rainbows emerge; this is the language of *folk meteorology*. Modern day folk meteorologists—that is, all of us—know perfectly well that *somehow or other* all those individual clouds and rainbows and snowflakes and gusts of wind are just the emergent saliencies (saliencies relative to *our* perceptual apparatus) of vast distributions of physical energy, water droplets, and the like.

There is a gap between folk meteorology and physics but not a very large and mysterious one. Moving back and forth between the two domains takes us on familiar paths, traversed many times a day on the TV news. It is important to note that the meteorologist's instruments are barometers, hygrometers, and thermometers, not cloudometers, rainbometers, and snowflakometers. The regularities of which the science of meteorology is composed concern pressure, temperature, and relative humidity, not the folk-meteorological categories.

There is not, today, any field of computational cloudology. Is this because meteorology is in its infancy, or is such an imagined science as out of place as astrology? Note that there are patterns, regularities, large scale effects, and, in particular, reactive effects between items in folk-meteorological categories and other things. For instance, many plants and animals are designed to discriminate folk-meteorological categories for one purpose or another. We can grant all this without having to suppose that there is a formal system governing those patterns and regularities, or the reactions to them. Similarly—and this is the moral of the meteorological metaphor—it does not follow from the fact that the folk-psychological level of explanation is the "right" level for many purposes that there must be a computational theory at or near that level. The alternative to HCC is that it is the clouds and rainbows in the brain that have intentionality—that refer to Chicago and grandmother—but that the rigorous computational theory that must account for the passage and transformation of these clouds and rainbows will be at a lower level, where the only semantics is internal and somewhat strained as semantics (in the same way the "semantics" of machine language is a far cry from the semantics of a natural language).

But how are we to move beyond the metaphors and develop these new low-level hunches into explicit theory at the "higher," or more "central," cognitive levels? The bits of theory that are getting explicit in the New Connectionist movement are relatively close to the "hardware" level of description, and the cognitive work they so far can do is often characterized as either relatively peripheral or relatively subordinate. For instance, pattern recognition appears (to many theorists) to be a relatively early or peripheral component in perception, and memory appears (to many theorists) to be a rather subordinate ("merely clerical" one might say) component in the higher intellectual processes of planning or problem solving. To the ideologues of the West, however, these appearances have misled. All thinking, no matter how intellectual or central or (even) rule-governed, will turn out to make essential use of fundamentally *perceptual* operations such as versatile pattern recognition; it is no accident that we often say "I see" when we come to understand. And, according to the Western view, the apportionment of responsibility and power between memory and intelligent processing will be unlike the underlying (and ineluctably influential) division of labor in von Neumann machines, in which the memory is inert, and cold storage and all the action happens in the central processing unit; a proper memory will do a great deal of the intelligent work itself.

So far as I know, no one has yet come up with a way of sorting out these competing hunches in a medium of expression that is uniform, clear, and widely understood (even if not formal). What we need is a level of description that is to these bits of theory *roughly* as software talk is to hardware talk in conventional computer science. That is, it should abstract from as many low-level processing details as possible while remaining in the spirit of the new architectures.

The problem is that we do not yet have many clear ideas about what the functions of such systems must be—what they must be able to do. This setting of the problem has been forcefully developed by David Marr in his methodological reflections on his work on vision.[31] He distinguishes three levels of analysis. The highest level, which he rather misleadingly calls computational, is in fact not at all concerned with computational processes, but strictly (and more abstractly) with the question of what function the system in question is serving—or, more formally, with what function *in the mathematical sense* it must (somehow or other) "compute." Recalling Chomsky's earlier version of the same division of labor, we can say that Marr's computational level is supposed to yield a formal and rigorous specification of a system's *competence*—"given an element in the set of x's it yields an element in the set of y's according to the following rules"—while remaining silent or neutral about implementation or *perfor-*

mance. Marr's second level down is the *algorithmic* level, which does specify the computational processes but remains neutral (as neutral as possible) about the *hardware*, which is described at the bottom level.

Marr's claim is that until we ge a clear and precise understanding of the activity of a system at the highest, "computational" level, we cannot properly address detailed questions at the lower levels, or interpret such data as we may already have about processes implementing those lower levels. Moreover, he insists, if you have a seriously mistaken view about what the computational-level description of your system is (as all earlier theorists of vision did, in his view), your attempts to theorize at lower levels will be confounded by spurious artifactual problems. [It is interesting to note that this is also the claim of J. J. Gibson, who viewed all cognitivistic, information-processing models of vision as hopelessly entangled in unnecessarily complex Rube Goldberg mechanisms posited because the theorists had failed to see that a fundamental reparsing of the inputs and outputs was required. Once we get the right way of characterizing what vision receives from the light, he thought, and what it must yield ("affordances"), the theory of vision would be a snap.]

Now Marr claims to have gotten the computational level right for vision, and his claim is not obviously too optimistic. But vision, like any peripheral system, is apparently much more tractable at Marr's computational level than are the central systems of thought, planning, problem solving, and the like that figure so centrally in AI explorations. Fodor argues in *The Modularity of Mind* (1983) that while there has been dramatic progress on the peripheral perceptual "modules" that "present the world to thought," "there is no serious psychology of central cognitive processes."

> We have, to put it bluntly, no computational formalisms that show us how to do this, and we have no idea how such formalisms might be developed. . . . If someone—a Dreyfus, for example—were to ask us why we should even suppose that the digital computer is a plausible mechanism for the simulation of global cognitive processes, the answering silence would be deafening. (p. 129)

But what is this? One would have thought that never the twain would meet, but here is Fodor, Archbishop of the East Pole, agreeing with Dreyfus, Guru of the West Coast, that High Church Computationalism has made no progress on "central cognitive processes." If Fodor is right in his pessimism—and I think for the most part he is—what might a reasonable theoretician do?

My proposal: go right on doing the sort of AI that has traditionally been associated with High Church Computationalism but abandon the computationalist ideology altogether and reinterpret the programs of these AI practitioners as *thought experiments*, not *models*.

Here is what I mean. If Marr is right to insist that progress must first be made on the problems at the computational level, then the first task confronting us if we want a theory of "central cognitive processes" is just to say what those processes are supposed to be able to accomplish. What is the nature of those central faculties? Forget for the moment *how* they do what they do. Just what is it that they (are supposed to) do? What is the competence the theorist should try to explain? As Fodor insists, no one has a clear, crisp explicit account of this. But several researchers are trying. Allen Newell, for instance, calls this level of description the Knowledge Level. It is, in effect, Marr's computational level as applied to the central arena. McCarthy draws a similar level distinction.[32] What these and many other theorists in AI have been doing is not proposing HCC models of human cognition or testing theories with empirical experiments, but casting about, in a *thought*-experimental way, for constraints and relationships that might inform the description (at Marr's "computational" level) of the mysterious "central cognitive processes." And at least this much progress has been made: we have enlarged and refined our vision of what powers human minds actually have. And we now know quite a few ways *not* to try to capture the basic competence—let alone the implementation—of the central cognitive systems. The process of elimination looms large in AI research; virtually every model seriously considered has been eliminated as far too simple for one reason or another. But that is progress. Until the models are seriously considered and eliminated they lurk as serious possibilities to tempt the theorist.

Thus McCarthy's formality constraint is not a commitment to High Church Computationalism (or need not be). It might be nothing more than the demand for enough rigor and precision to set the problem for the next level down, Marr's algorithmic level, except that this would probably not be a good term for the highest level at which the processes (as contrasted with the products) of the New Connectionist sort were described.

And Newell and Simon's search for "rules" of "thinking" need not commit them or their admirers to the HCC doctrine that thinking is rule-*governed* computation. The rules they discover (supposing they succeed) may instead be interpreted as regularities in patterns in the emergent phenomena—the cognitive "clouds" and "rainbows"—but not "mere" regularities. The well-known distinction (in philosophy) between rule-following behavior and rule-described behavior is often illustrated by pointing out that the planets do not compute their orbits, even though *we* can, following rules that describe their motions. The "rules" of planetary

motion are law-like regularities, not "followed" rules. This is true, but it ignores a variety of regularity intermediate between the regularities of planets (or ordinary cloud formations) and the regularities of rule-following (that is, rule-*consulting*) systems. These are the regularities that are preserved under selective pressure: the regularities dictated by principles of good design and hence homed in on by self-designing systems. That is, a "rule of thought" may be much more than a mere regularity; it may be a *wise* rule, a rule one would design a system by if one were a system designer, and hence a rule one would expect self-designing systems to "discover" in the course of settling into their patterns of activity. Such rules no more need to be explicitly represented than do the principles of acrodynamics honored in the design of birds' wings.

For example, Marr discovered that the visual system operates with a tacit assumption that moving shapes are articulated in rigid linkages and that sharp light-intensity boundaries indicate physical edges. These assumptions are not "coded" in the visual system; the visual system is designed to work well only in environments where the assumptions are (by and large) true. Such rules and principles should be very precisely formulated at the computational level—not so they can then be "coded" at the algorithmic level but so that the (algorithmic) processes can be designed to honor them (but maybe only with a high degree of regularity).

Are there "rules" of (good) problem solving that must be (and are) tacit in the regularities that emerge in the information processing of mature thinkers? One might discover them by attempting to *codify* such rules in a rule-following system whose behavior exhibited those regularities because those were the regularities it was "told" to follow. Such systems can be put through their paces to test the adequacy of the rules under consideration.[33] It is this testing that has led to the (often informative) elimination of so many tempting models in AI.

In sum, there is no reason I can see for AI or cognitive science to take on the rather unlikely burden of defending HCC. It seems to me that all the valuable AI research that has been done can be viewed as attempts to sketch competences. (Marr, of course, went far beyond that.) As such it is best viewed as consisting of (preliminary) thought experiments, not as more "mature" genuinely experimental science. But its thought experiments are subject to a modicum of control. One can *test* such a sketch of a competence by test driving an unbiologically produced Rube Goldberg device with that competence (the actual "computational" AI program) to see how it would perform.

This leaves us with an almost embarrassingly ecumenical conclusion. Everyone is right about something. Dreyfus and the Zen Holists are right that we need not commit ourselves to the defining dogmas of High

Church Computationalism, but the people engaged in devising computational models of cognitive processes are right that their methodology is probably the best way to make headway on the mysteries of the mind. Everybody agrees that something or other in the brain must be capable of having the semantic property of referring to Chicago and that it is the task of *some* sort of computational theory to explain and ground this power. Residual disagreements are either based on unmotivated allegiances to bits of outworn creed or are substantive disagreements on just which brand of interstitial computational theory is apt to be most promising. There is only one way to settle these latter disagreements: roll up our sleeves and devise and test the theories.[34]

Notes

This paper was prepared for the Conference on Philosophy and Cognitive Science at MIT, May 17–20, 1984, sponsored by the Sloan Foundation. Written under a deadline for the purpose of providing a glimpse of the state of the art in mid-1984, it will no doubt have a short shelf life. So read it now, or if now is later than 1986, read it as a quaint reflection on how some people thought back in 1984.

1. Richard Dawkins speaks of those who are "holistier than thou" in *The Extended Phenotype* (1982, p.113).

2. For another attempt at a systematic spatial ordering of views on a closely related issue, see John Haugeland's paper, "The Intentionality All-Stars," (still unpublished, alas) which identifies various positions on intentionality with baseball positions, and in the process creates some striking and illuminating juxtapositions. For instance, Fodor, Kant, and Husserl are all on first; Wittgenstein, Quine, and Ryle are shortstops; Searle is out in right field with Derrida. (Nagel is at the plate, of course, wondering what it's like to be at bat.)

3. Allen Newell would call them intellectual issues. See Newell 1983.

4. See, e.g., Taylor 1983.

5. Stich 1983; Cummins 1983; Haugeland 1978, 1981; Boden 1984.

6. See, e.g., Dennett 1982, 1983a, 1984a (my review of Fodor's *Modularity of Mind*).

7. Newell 1980.

8. See, e.g., Chomsky 1980a, 1980b.

9. See, e.g., Pylyshyn 1978, 1980, 1984.

10. In the someday-forthcoming 8th edition of *The Philosophical Lexicon*, "winograd" (n. sometimes pronounced "wino-grad") is defined as the degree of intoxication occasioned by moving to the West Coast.

11. Hofstadter, on reading an earlier draft of this paper, suggested that Zadeh's fuzzy set theory is actually better seen as an attempt, entirely within the Eastern

Orthodoxy, to achieve West Coast ends with East Pole means. I am inclined to agree; this is one of the fine points of interpretation in need of further work.

12. A forthcoming collaborative effort by psychologist Steven Kosslyn and philosopher Gary Hatfield, "Representation without Symbols" in *Journal of Social Research*, expresses Western views similar to those developed by Boden and by me in "Styles of Mental Representation."

13. See, e.g., McCarthy 1980.

14. Fodor 1980.

15. See, e.g., Newell and Simon 1972, 1976.

16. See, for instance, Edward Stabler 1983, pp. 391–422, and especially the commentaries. See also Chomsky 1980b, and the commentaries.

17. Newell and Simon rely on the same challenge in "Computer Science as Empirical Enquiry: Symbols and Search." See (in Haugeland reprinting) p.50: "The principal body of evidence for the symbol-system hypothesis that we have not considered is negative evidence: the absence of specific competing hypotheses as to how intelligent activity might be accomplished—whether by man or by machine."

18. Simon has objected (in conversation) that his work with Newell is replete with solid empirical evidence in favor of their "production system" models of human problem solving. But even if I were to grant that *something very like* Newell and Simon's productions have been shown empirically to be involved in—even the basis of—human problem solving, there would still be no empirical evidence as yet showing that the *computational implementation* of production systems in computers realistically models any level of the neural implementation of the production-like processes in human thinking. For a related argument see Stabler (1983) and my commentary on Stabler (1983b, p.406–407).

19. See Dennett 1984b.

20. Hofstadter 1983b. Hofstadter calls High Church Computationalism the Boolean Dream.

21. In "Cognitive Wheels" (1984b) I call this dodge the declaration that "the brain is wonder tissue."

22. See Feldman and Ballard 1982.

23. Clark Glymour, in "Android Epistemology," (forthcoming) declares that AI—and here he is surely referring to East Pole AI—is actually "logical positivism carried on by other means."

24. See, e.g., McClelland, unpubl. manuscript.

25. Hillis 1981; Fahlman, Hinton, and Sejnowski 1983.

26. Hofstadter 1983a.

27. Kirkpatrick, Gelatt, and Vecchi 1983.

28. Smolensky 1983.

29. Kanerva 1983.

30. John Haugeland, in "The Nature and Plausibility of Cognitivism," held out bravely for some sort of hologram-like alternative to a computationalist language of thought. Now there are some actual models to examine—and not just the (perhaps visionary but) metaphorical suggestions of Pribram and Arbib.

31. Marr 1982.

32. Newell 1982; McCarthy 1980.

33. I think it *may* be helpful to compare this interpretation of AI strategy with the simulations explored by evolutionary theorists such as John Maynard Smith and Richard Dawkins, who ask questions about whether certain behavioral "strategies" are evolutionarily stable by explicitly codifying the strategies in the behavior of imaginary organisms, and then pitting them against alternative (explicit, rule-governed) strategies embodied in rival imaginary organisms, to see which (pure, idealized) strategy would win in Nature under various conditions. See Dawkins (1976, 1982) for good introductory discussions. See also George Axelrod (1984) on prisoners' dilemma competitions between simulations for a similarly motivated research effort.

34. Douglas Hofstadter has had an even greater role than usual in shaping my thinking on these issues, so if I am all wrong about this, he *is* responsible, and will have to share the blame. But he is not responsible for my failures to understand or do justice to the various efforts I discuss here. Others who have commented on earlier drafts of this paper, including Robert Cummins, Jerry Feldman, John Haugeland, Hilary Putnam, and Herbert Simon, are hereby thanked and absolved in the usual manner.

Bibliography

Axelrod, G. 1984. *The evolution of cooperation*. New York: Basic Books.

Boden, M. 1984. What is computational psychology? *Proceedings of the Aristotelian Society* 58 (suppl.):17–53.

Chomsky, N. 1980a. *Rules and representations*. New York: Columbia Univ. Press.

———. 1980b. Rules and representations. *Behavioral and Brain Sciences* 3:1–61.

Cummins, R. 1983. *The nature of psychological explanation*. Cambridge: MIT Press, Bradford Books.

Dawkins, R. 1976. *The selfish gene*. Oxford: Oxford Univ. Press.

———. 1982. *The extended phenotype*. Oxford and San Francisco: Freeman.

Dennett, D. 1978. *Brainstorms: Philosophical essays on mind and psychology*. Cambridge: MIT Press, Bradford Books.

———. 1982. Beyond belief. In *Thought and object*. ed. Andrew Woodfield, pp. 1–95. Oxford: Clarendon Press.

———. 1983a. Styles of mental representation. *Proceedings of the Aristotelian Society* 83:213–226.

———. 1983b. When do representations explain? *Behavioral and Brain Sciences* 6:406–7.

———. 1984a. Carving the mind at its joints. *Contemporary Psychology* 29:285–286.

———. 1984b. Cognitive wheels: The frame problem of AI. In *Minds, machines and evolution*, ed. C. Hookway, pp. 129–151. Cambridge: Cambridge Univ. Press.

Fahlman, S. E., G. Hinton, and T. J. Sejnowski. 1983. Massively parallel architectures for AI: NETL, Thistle and Boltzmann machines. *Proceedings of the American Association of Artificial Intelligence* 83:109–113.

Feldman, J. and D. H. Ballard. 1982. Connectionist models and their properties. *Cognitive Science* 6:205–254.

Field, H. 1978. Mental representation. *Erkenntnis* 13:9–61.

Fodor, J. 1975. *The language of thought*. New York: Crowell.

_____. 1980. Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 3:63–110. [Also published in Fodor (1981, pp. 225–253).]

_____. 1981. *RePresentations: Philosophical essays on the foundations of cognitive science*. Cambridge: MIT Press, Bradford Books.

_____. 1983. *The modularity of mind*. Cambridge: MIT Press, Bradford Books.

Gibson, J.J. 1975. *Cognition and reality*. San Francisco: Freeman.

Glymour, C. Forthcoming. Android epistemology.

Harman, G. 1973. *Thought*. Princeton: Princeton Univ. Press.

Haugeland, J. 1978. The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* 1:215–260. [Also published in Haugeland (1981, pp. 243–281).]

_____, ed. 1981. *Mind design: Philosophy, psychology, artificial intelligence*. Cambridge: MIT Press, Bradford Books.

_____. The intentionality all-stars. Unpublished.

Hillis, D. 1981. *The connection machine (computer architecture for the new wave)*. AI Memo 646 MIT, September.

Hofstadter, D. 1979. *Gödel, Escher, Bach: An eternal golden braid*. New York: Basic Books.

_____. 1983a. The architecture of Jumbo. In *Proceedings of the Second International Machine Learning Workshop*, 161–170. Urbana, Ill.: Univ. of Illinois.

_____. 1983b. Artificial intelligence: Subcognition as computation. In *The study of information: Interdisciplinary messages*, ed. F. Machlup and U. Mansfield, pp. 263–285. New York: Wiley.

Kanerva, P. 1983. *Self-propogating search: A unified theory of memory*. Techn. rept. Center for the Study of Language and Information, Palo Alto.

Kirkpatrick, S., C. D. Gelatt, Jr., and M.P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 13 May 1983, 671–680.

Kosslyn, S. and G. Hatfield. Forthcoming. Representation without symbols. *Journal of Social Research*.

Marr, D. 1982. *Vision*. New York: Freeman.

McCarthy, J. 1980. Circumscription–a form of non-monotonic reasoning. Stanford AI Lab memo AIM-334, February. (Also pub. in *Artificial Intelligence* 13:27–39.)

McClelland, J. L. Models of perception and memory based on principles of natural organization. UC San Diego. Unpublished.

Neisser, U. 1963. *Cognitive psychology*. New York: Appleton–Century–Crofts.

_____. 1975. *Cognition and reality*. San Francisco: Freeman.

Newell, A. 1980. Physical symbol systems. *Cognitive Science* 4:135–183.

_____. 1982. The knowledge level. *Artificial Intelligence* 18:87-127.

_____. 1983. Intellectual issues in the history of artificial intelligence. In *The study of information: Interdisciplinary messages*, ed. F. Machlup and U. Mansfield, pp. 187–227. New York: Wiley.

Newell, A., and H. Simon. 1972. *Human problem solving*. Englewood Cliffs, N.J.: Prentice-Hall.

_____. 1976. Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery* 19:113–126. (Also published in Haugeland. 1981. 35-66.)

Pylyshyn, Z. 1978. Computational models and empirical constraints. *Behavioral and Brain Sciences* 1:93–127.

_____. 1980. Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences* 3:111–169.

_____. 1984. *Computation and cognition: Toward a foundation for cognitive science*. Cambridge: MIT Press, Bradford Books.

Smolensky, P. 1983. Harmony theory: A mathematical framework for learning and parallel computation. *Proceedings of the American Association of Artificial Intelligence* 83:114–132.

Stabler, E. 1983. How are grammars represented? *Behavioral and Brain Sciences* 6:391–422.

Stich, S. 1983. *From folk psychology to cognitive science*. Cambridge: MIT Press, Bradford Books.

Taylor, C. 1983. The significance of significance: The case of cognitive psychology. In *The need for interpretation*, ed. S. Mitchell and M. Rosen, pp. 141–169. London: Athlone. New York: Humanities.