So How *Does* the Mind Work?

STEVEN PINKER

Abstract: In my book *How the Mind Works*, I defended the theory that the human mind is a naturally selected system of organs of computation. Jerry Fodor claims that 'the mind doesn't work that way' (in a book with that title) because (1) Turing Machines cannot duplicate humans' ability to perform abduction (inference to the best explanation); (2) though a massively modular system could succeed at abduction, such a system is implausible on other grounds; and (3) evolution adds nothing to our understanding of the mind. In this review I show that these arguments are flawed. First, my claim that the mind is a computational system is different from the claim Fodor attacks (that the mind has the architecture of a Turing Machine); therefore the practical limitations of Turing Machines are irrelevant. Second, Fodor identifies abduction with the cumulative accomplishments of the scientific community over millennia. This is very different from the accomplishments of human common sense, so the supposed gap between human cognition and computational models may be illusory. Third, my claim about biological specialization, as seen in organ systems, is distinct from Fodor's own notion of encapsulated modules, so the limitations of the latter are irrelevant. Fourth, Fodor's arguments dismissing of the relevance of evolution to psychology are unsound.

In 2000 Jerry Fodor published a book called *The Mind Doesn't Work That Way* (hereafter: *TMDWTW*). The way that the mind doesn't work, according to Fodor, is the way that I said the mind does work in my book *How the Mind Works* (*HTMW*).[1] This essay is a response to Fodor, and one might think its title might be *Yes, It Does*! But for reasons that soon become clear, a more fitting title might be *No One Ever Said it Did*.

Fodor calls the theory in *How the Mind Works* the New Synthesis. It combines the key idea of the cognitive revolution of the 1950s and 1960s—that the mind is a computational system—with the key idea of the new evolutionary biology of the 1960s and 1970s—that signs of design in the natural world are products of the natural selection of replicating entities, namely genes. This synthesis, sometimes known as evolutionary psychology, often incorporates a third idea, namely that the mind is not a single entity but is composed of a number of faculties specialized for solving different adaptive problems. In sum, the mind is a system

Supported by NIH grant HD 18381. I thank Clark Barrett, Arthur Charlesworth, Helena Cronin, Dan Dennett, Rebecca Goldstein, and John Tooby for invaluable comments.

Address for Correspondence: Department of Psychology, William James Hall 970, Harvard University, Cambridge MA 02138.
Email: pinker@wjh.harvard.edu

[1] Fodor discusses *HTMW* together with a second book, Henry Plotkin's *Evolution in Mind* (Plotkin, 1997), which is similar in approach. But Fodor focuses on *HTMW*, as will I.

Mind & Language, Vol. 20 No. 1 February 2005, pp. 1–24.

of organs of computation that enabled our ancestors to survive and reproduce in the physical and social worlds in which our species spent most of its evolutionary history.

Readers who are familiar with Fodor's contributions to cognitive science but who have not read *TMDWTW* might be puzzled to learn that Fodor begs to differ so categorically. The first major theme of *HTMW* is computation, and Fodor, more than anyone, has defended what he calls the computational theory of mind: that thinking is a form of computation. The second major theme is specialization, and Fodor's most influential book is called *The Modularity of Mind*, a defense of the idea that the mind is composed of distinct faculties rather than a single general-purpose learning device or intelligent algorithm. The third theme is evolution, the source of innate biological structure, and Fodor, like many evolutionary psychologists, is willing to posit far more innate structure than is commonly accepted in contemporary philosophy and psychology. So it is surprising that Fodor insists that *HTMW* is wrong, wrong, wrong. Fodor and I must disagree on how the concepts of computation, faculty psychology (specialization), and innate biological organization should be applied to explaining the mind. This essay will be organized accordingly.

The Concept of Computation in *How the Mind Works*

According to *HTMW* (pp. 24–27; chap. 2), mental life consists of information-processing or computation. Beliefs are a kind of information, thinking a kind of computation, and emotions, motives, and desires are a kind of feedback mechanism in which an agent senses the difference between a current state and goal state and executes operations designed to reduce the difference. 'Computation' in this context does not refer to what a commercially available digital computer does but to a more generic notion of mechanical rationality, a concept that Fodor himself has done much to elucidate (Fodor, 1968; 1975; 1981; 1994).

In this conception, a computational system is one in which knowledge and goals are represented as patterns in bits of matter ('representations'). The system is designed in such a way that one representation causes another to come into existence; *and* these changes mirror the laws of some normatively valid system like logic, statistics, or laws of cause and effect in the world. The design of the system thus ensures that if the old representations were accurate, the new ones are accurate as well. Deriving new accurate beliefs from old ones in pursuit of a goal is not a bad definition of 'intelligence', so a principal advantage of the computational theory of mind (CTM) is that it explains how a hunk of matter (a brain or a computer) can be intelligent.

CTM has other selling points. It bridges the world of mind and matter, dissolving the ancient paradox of how seemingly ethereal entities like reasons, intentions, meanings, and beliefs can interact with the physical world.

It motivates the science of cognitive psychology, in which experimenters characterize the mind's information structures and processes (arrays for images, tree structures for sentences, networks for long-term memory, and so on). Since computational systems can have complex conditions, loops, branches, and filters which result in subtle, situationally appropriate behavior, the CTM allows the mind to be characterized as a kind of biological mechanism without calling to mind the knee-jerk reflexes and coarse drives and imperatives that have made people recoil from the very idea. Finally, mental life—internal representations and processes—appears to be more lawful and universal than overt behavior, which can vary with circumstances. This is behind Chomsky's idea that there is a single Universal Grammar that applies to all the world's languages despite their differences in overt words and constructions. Much of *HTMW* extends this idea to other areas of human psychology, such as the emotions, social and sexual relations, and humor.

Fodor, as I have acknowledged, deserves credit for capturing the sense of 'computation' in which it can sensibly be said that the mind is a kind of computer. That sense—in which a system's state transitions map onto logical relationships, or, as Fodor often puts it, the components of the system have both causal and semantic properties—says nothing about binary digits, program counters, register operations, stored programs, or any of the other particulars of the machines that process our email or compute our taxes and which are improbable characterizations of a human brain. The beauty of Fodor's original formulation is that it embraces a variety of systems that we might call 'computational', including ones that perform parallel computation, analogue computation (as in slide rules and adding machines), and fuzzy computation (in which graded physical variables represent the degree to which something is true, or the probability that something is true, and the physical transitions are designed to mirror operations in probability theory or fuzzy logic rather than in classical logic). Any adequate characterization of the concept of 'computation' should embrace these possibilities. After all, the term *digital computer* is not redundant, and the terms *analogue computer* and *parallel computer* are not self-contradictory.

At the same time, the computational theory of mind is by no means empty or necessary. It can be distinguished from the traditional belief that intelligence comes from an immaterial substance, the soul. It differs from the claim that intelligence is made possible only by specific biochemical properties of neural tissue. It differs from the assertion that mental life can be understood only in terms of first-person present-tense subjective experience. And it differs from the claim that intelligence can be understood only by considering what mental states refer to in the world, or by examining the incarnate person embedded in a physical and social context. Fodor emphasizes the idea that the representations in a computational system are *syntactic*: they are composed of parts in some arrangement, and the causal mechanisms of the system are sensitive to the identity and arrangement of those parts rather than to what they refer to in the world.

The Concept of Specialization in *How the Mind Works*

HTMW does not try to account for all of human behavior using a few general-purpose principles such as a large brain, culture, language, socialization, learning, complexity, self-organization, or neural-network dynamics. Rather, the mind is said to embrace subsystems dedicated to particular kinds of reasoning or goals (pp. 27–31). Our intelligence, for example, consists of faculties dedicated to reasoning about space, number, probability, logic, physical objects, living things, artifacts, and minds. Our affective repertoire comprises emotions pertaining to the physical world, such as fear and disgust, and emotions pertaining to the social and moral worlds, such as trust, sympathy, gratitude, guilt, anger, and humor. Our social relationships are organized by distinct psychologies applied to our children, parents, siblings, other kin, mates, sex partners, friends, enemies, rivals, trading partners, and strangers. We are also equipped with communicative interfaces, most prominently language, gesture, vocal calls, and facial expressions.

The intended analogy is to the body, which is composed of systems divided into organs assembled from tissues built out of cells. Our 'organs of computation', therefore, are not like discrete chips laid out on a board with a few solder tracks connecting them. Just as some kinds of tissue, like the epithelium, are used (with modifications) in many organs, and some organs, like the blood and the skin, interact with the rest of the body across an extensive interface, some kinds of specialized thoughts and emotions may serve as constituents that are combined into different assemblies. The concept of an artifact, for example—an object fashioned by an intelligent agent to bring about a goal—combines the concept of an object from intuitive physics with the concept of a goal from intuitive psychology. The psychology of sibling relations embraces the emotion of affection (also directed toward mates and friends), an extra feeling of solidarity triggered by perceived kinship, and a version of disgust pinned to the thought of having sexual relations with the sibling.

This kind of faculty psychology has numerous advantages. It is consistent with models of cognitive faculties such as language, spatial cognition, and audition that require specialized machinery (nouns and verbs, allocentric and egocentric frames of reference, and pitch and timbre, respectively). It is supported by the existence of neurological and genetic disorders that target these faculties unevenly, such as a difficulty in recognizing faces (and facelike shapes) but not other objects, or a difficulty in reasoning about minds but not about objects or pictures. Finally, a faculty psychology is necessary to account for many of the complex but systematic patterns in human thought and emotion. The fact that we love our siblings but don't want to have sex with them, and may want to have sex with attractive strangers without necessarily loving them, is inexplicable by a theory of social psychology that doesn't distinguish among kinds of human relationships but appeals only to global drives like 'positive affect'.

The Appeal to Evolution in *How the Mind Works*

Evolution is the third key idea in *HTMW* (pp. 21–24; chap. 3). The organs of computation that make up the human mind are not tailored to solve arbitrary computational problems but only those that increased the reproductive chances of our ancestors living as foragers in pre-state societies.

One advantage of invoking evolution is that it provides psychology with explanatory adequacy. It helps account for *why* we have the specializations we do: why children learn spoken language instinctively but written language only with instruction and effort, why the system for recalling memories satisfies many of the specifications of an optimal information-retrieval system, why our preferred sexual partners are nonsiblings who show signs of health and fertility. More generally, it explains why the human psyche has specific features that could not be predicted from the mere proposition that the brain engages in computation.

Evolutionary psychology also helps to explain many instances of error, irrationality, and illusion—why we gamble, eat junk food, fall for visual illusions, obsess over celebrities, and fear snakes and heights more than hair dryers near bathtubs or driving without a seatbelt. The nature of the explanation is that there can be a mismatch between the ancestral environment to which our minds are evolutionarily adapted and the current environment in which we find ourselves.

The most general attraction of a synthesis between cognitive science and evolutionary psychology is that it continues the process of the unification of putatively incommensurable metaphysical realms that has been the major thrust of science for four centuries (Tooby and Cosmides, 1992; Wilson, 1998). Newton united the sublunary and superlunary spheres, Lyell united the formative past and static present, Wöhler united living tissue and nonliving chemistry, and Darwin, Mendel, and Watson and Crick united seeming teleological design in organisms with ordinary processes of forward causation. In the same way, the idea that the human mind is an evolved computer aims to bridge the last major chasm in human knowledge, that between matter and mind, biology and culture, nature and society, the natural sciences and the humanities. This consilience promises not only a more parsimonious metaphysics but greater depth and explanatory power for the disciplines that study the mind and its products. Hypotheses about psychological function cannot be conjured up by whim but must be compatible with evolutionary biology and in some cases may be deduced from it.

I turn now to how each of these themes is treated in Fodor's critique of *HTMW*.

The Concept of Computation in *The Mind Doesn't Work that Way*

In *TMDWTW*, Fodor argues that he never meant that *all* of the mind could be explained as a kind of computation. On the contrary, there is a key thing that a human mind can do but which a computational system cannot do. I will discuss

this allegedly special human feat soon, but the debate cannot proceed if *HTMW* and *TMDWTW* don't mean the same thing by the word 'computation'.

And they don't. In *TMDWTW*, Fodor departs from the generic characterization of computation in his previous writings and assumes a far more specific and far less psychologically plausible version. He now defines the Computational Theory of Mind as 'whether the architecture of (human) cognition is interestingly like the architecture of Turing's kind of computer' (p. 105, note 3). Similarly, he evaluates the idea that 'cognitive architecture is Classical Turing architecture; that is, that the mind is interestingly like a Turing machine' (p. 30).[2]

A Turing Machine is a design for a hypothetical computer that Alan Turing found convenient to use in his proof that partial recursive functions could be computed by formally specified mechanical systems. It consists of a control unit and an infinite tape divided into squares which can be imprinted with any of a fixed number of symbols. The tape serves as the machine's input, output, and working memory; the control unit can 'look at' one square at a time. The control unit can be in a finite number of states, and is governed by a finite transition network which senses the machine's state and the visible symbol on the tape, and in response can change state, print or erase a symbol, and move the tape one square to the left or right. A Turing machine can compute any partial recursive function, any grammar composed of rewrite rules, and, it is commonly thought, anything that can be computed by any other physically realizable machine that works on discrete symbols and that arrives at an answer in a finite number of steps.

No one has ever built a Turing Machine (other than for pedagogical purposes) because it is maddeningly difficult to program and stupefyingly inefficient to run. It was invented only as a convenient mathematical construction, not as a prototype for a workable computer, and certainly not as a model of the functioning of the human mind. No one has ever taken seriously the idea that 'cognitive architecture is Classical Turing architecture', so the central premise of *TMDWTW*—that a Turing Machine is unsuited to solve a certain kind of problem that the human mind easily solves—is not relevant to anything. It is certainly not relevant to *HTMW*, which took pains to differentiate Turing Machines and current digital computers from the generic notion of computation (pp. 26–27, 64–69).

It's hard to credit that Fodor takes seriously the idea that the human memory is like a tape divided into squares, even for the domains (like language parsing) in which he believes that CTM is true. Could Fodor have something more abstract in mind, notwithstanding his explicit references to computational

---

[2] These are not the only quotations in *TMDWTW* in which Fodor equates computation with Turing Machines. For example, he suggests in another place that 'we will *all* have to give up on the Turing story as a general account of how the mind works, and hence, a fortiori, that we will have to give up on the generality of the [synthesis of computation and evolution in *HTMW*]' (pp. 46–47).

architecture? He does bring up the weaker idea that 'the mind is Turing-equivalent' (p. 105, note 3; see also p. 33), and that 'minds are "input-output equivalent" to Turing machines' (p. 30). But no one would actually defend this version of the Computational Theory of Mind either. The class of functions computable by Turing machines includes every computer program you would ever have reason to think of (calculating the digits of *pi*, organizing a company's payroll), and countless ones that one would never have reason to think of. In the domain that Fodor considers most amenable to computational analyses, language, it is axiomatic among linguists that the set of possible human languages is vastly smaller than the set of languages generable by Turing machines. If it weren't, characterizing Universal Grammar would be trivial, and languages would be unlearnable (Pinker, 1979). So Turing-Machine equivalence is as red a herring as Turing-Machine architecture.

In one place Fodor adumbrates what he means by properties of computational architecture that are 'interestingly like' Turing Machines. He attributes to Turing a version of CTM in which 'mental processes are operations defined on syntactically structured mental representations that are much like sentences' (p. 4). This characterization of CTM is also puzzling. For one thing, Turing machines *aren't*, by design, sensitive to the structure of representations: they can 'see' only a single symbol at a time, and at best can be programmed to *emulate* systems that are sensitive to structure. Nor did Turing himself say anything about structured or sentence-like representations.[3] One could, in principle, program a Turing machine to emulate a structure-sensitive architecture, but then would program a Turing machine to emulate a connectionist architecture as well (with the analogue values approximated to some arbitrary degree of precision). As for real computers, they avail themselves of many representational formats, most of which are not particularly like sentences (relational databases, image files, list structures, and so on). And with the possible exception of Chomsky's 'Logical Form' and other representations of the semantically relevant information in syntax, computational models of the human mind rarely posit 'mental representations that are much like sentences'. Consider, for example, visible-surface arrays (a.k.a. 2½–D sketches), semantic networks, mental models, phrase-structure rules, and analogue imagery representations.

At times Fodor invokes a still weaker ('minimal') form of CTM, namely that 'the role of a mental representation in cognitive processes supervenes on some syntactic fact or other' (p. 29), that is, that mental representations affect cognitive processing by virtue of the identity and arrangement of the symbols composing them. He refers to a system of this sort as having 'classical'

[3] Despite Fodor's frequent invocation of Turing and Quine in *TMDWTW*, he does not actually cite anything by them. I will assume that the arguments Fodor has in mind come from Turing's 'Computing Machinery and Intelligence' (Turing, 1950) and Quine's 'Two Dogmas of Empiricism' (Quine, 1960).

computational architecture (e.g., p. 31), contrasts it with connectionist and associationist alternatives (which are sensitive only to collections of features and lack syntactic organization), and grants it the computational power of Turing machines (p. 30). In a rather abstruse and disorganized discussion, (pp. 28–33), Fodor seems to contrast minimal-CTM with his strong, Turing-machine-architecture-CTM in the following way. Turing machines can process only local information, such as the information inside a proposition, and hence are unable to respond to global properties of the total set of propositions, such as whether they are collectively parsimonious, or whether they are mutually relevant or consistent. A minimal-CTM computer, in contrast, can process an arbitrarily large set of propositions at once, including propositions that help determine whether the remaining propositions satisfy some global property. Fodor warns the reader that in order to do this, a minimal-CTM computer would have to swallow implausibly large databases in one bite, perhaps the entirety of the system's knowledge.

In all these characterizations, Fodor describes a computational system at a level close to what hackers call 'bare metal': the elementary information-processing steps built directly into the hardware. This leads to his repeated emphasis on how myopic and inflexible computational systems are, an emphasis that, we shall see, he compares unfavorably to the human mind. Nowhere does Fodor acknowledge that real computers overlay the bare metal with many layers of software that endow them with more global reach and more flexible powers, and that it is this 'virtual machine', the one visible to programmers and users, that specifies the system's powers in practice. An obvious example is an internet search engine, which repeatedly examines pages on the World Wide Web and constructs a database capturing which words are found in which documents and which documents are linked to which other documents. By processing this database, instead of the entire Web directly, the search engine can respond to global properties, such as which page on the Web is likely to be most relevant to a search query. A person who learned about the nature of computation from *TMDWTW* would have no idea that computers might be capable of such a feat.

Just as curiously, Fodor says nothing about the computational architectures that have been proposed as actual models of the *mind*. There is no mention in *TMDWTW* of production systems, semantic networks, knowledge representation languages, unification systems, dynamic binding networks, massively parallel architectures, and hybrid connectionist-symbolic systems. All of these are 'computational' in Fodor's original, generic sense (that is, they contain symbols that have both semantic and causal properties), and all of them are syntactic (rather than connectionist or associationist) in that at least some of their operation depends on the internal relations among the elements in their representations. Yet they do not work like Turing Machines or the variants that Fodor presents as the essence of the computational theory of mind, and Fodor does not refer to them in his discussion. As we shall see, this is a key omission.

Fodor on the Limits of Computational Psychology

Fodor believes he has identified a feat that human minds can do but that Turing machines and their kin cannot do.[4] He calls this feat 'abduction', 'globality', the 'frame problem', and 'inference to the best explanation'.

Frustratingly, Fodor never gives a clear definition of what he means by abduction, nor does he work through an example that lays bare exactly how a computational system (Turing machine or other) fails to do something that humans easily do. He often seems to use 'abduction' and its relatives to embrace any really hard problem about cognition, as if the real title of the book was *We Don't Understand Everything About the Mind Yet*. But Fodor's general idea is that when people solve a problem they have an uncanny ability to bring to bear on it just the information that is most relevant to it. Moreover, people can absorb the implications of some new fact or conclusion, and can be sensitive to the overall parsimony and consistency of a belief system, without exhaustively searching the contents of memory and testing the implications of a fact against everything they know.

Fodor asserts that abduction is beyond the abilities of a classical computational system, because such a system can only apply rules to circumscribed strings according to local conditions of matching and mismatching symbols. This may suffice to parse a sentence using a set of grammatical rules, or to derive a conclusion from a set of premises using modus ponens. But it does not allow the system to revise a belief that is an indirect implication of an entire set of beliefs. In those cases there is no simple 'if-then' rule that cranks such implications out.

Fodor offers an example from common-sense reasoning. A reasoning system implemented on a classical computer would have depart from human reasoning, he claims, in the following scenario:

> The thought that there will be no wind tomorrow significantly complicates your arrangements if you had intended to sail to Chicago, but not if your plan was to fly, drive, or walk there. But, of course, the syntax of the mental representation that expresses the thought *no wind tomorrow* is the same whichever plan you add it to. The long and short of it is: The complexity of a thought is not intrinsic: it depends on the context (p. 26).

The example is quite unconvincing. Even the stupidest reasoning system would be programmed to test for wind conditions before sailing, and branch to an appropriate course of action depending on the outcome, but not run that test before driving or walking. Perhaps realizing this, Fodor spends most of his time on

[4] Actually it is unclear whether Fodor is making the strong mathematical claim that he has identified a function that cannot be computed by a Turing machine at all, or that he has merely identified a function that Turing Machines cannot do with humanlike speed and efficiency. The latter may be the upshot of his discussion of the 'minimal' computational theory of mind.

examples from the history of science. For example, he ventures, a classical computer cannot understand that in Newtonian mechanics heavier objects don't necessarily fall faster, or that in modern chemistry metals need not be solid. Fodor mentions the names W.V.O. Quine and Pierre Duhem without explanation; presumably it is an allusion to their arguments that the entire set of one's beliefs form an interconnected whole and that the only criterion for justifying a *particular* belief is its effect on the coherence and simplicity of the entire *system* of beliefs. One's definition of art, for example, might depend on assumptions about the universality of art across human cultures, which may depend on the antiquity of artistic creation in human prehistory, which may depend on radiocarbon dating of cave paintings. So a revision in physicists' understanding of the process of radioactive decay could alter our definition of art, despite the fact that any explicit set of canons for physics would say nothing about art or vice versa.

Fodor's argument, then, is that there is an unbridgeable chasm between the feats of human abduction and the powers of computational systems. It is a crisis for cognitive science, not just for *HTMW*: 'I'm inclined to think that Chicken Little got it right. Abduction really is a terrible problem for cognitive science, one that is unlikely to be solved by any kind of theory we have heard of so far' (p. 41). Inverting the paper-ending cliché, Fodor suggests that in this area, *less* research needs to be done. 'Do nothing about abduction', he advises; 'Wait until someone has a good idea' (p. 52). Until that day, cognitive scientists should concentrate on parts of the mind in which global interactions with knowledge are minimal, such as vision, speech, and syntactic parsing.

Problems with Fodor's Critique of Computational Psychology

But Fodor's supposed chasm can be narrowed from both sides. Let's begin with the human mind's powers of abduction. Fodor's reliance on examples from the history of science to illustrate the inimitable feats of cognition has an obvious problem: the two work in very different ways. A given scientific inference is accomplished by a community of thousands of scientists, who work over centuries, use sophisticated mathematical and technological tools, pool their results in journals and conferences, and filter out false conclusions through empirical hypothesis-testing and peer criticism. And their accomplishments are appreciated in hindsight through histories written by the victors with the false starts edited out (the phlogiston, the N-rays, the Lamarckism, the cold fusion). A common-sense inference, in contrast, is accomplished by a single brain working in seconds, doing its best with what it has, and scrutinized by cognitive scientists in real time, errors and all. Granted that several millennia of Western science have given us non-obvious truths involving circuitous connections among ideas; why should theories of a single human mind be held to the same standard? Would a typical person, working alone and unprompted, abduce that in modern chemistry, solidity is not a necessary property

of metals, or that planetary regression complicates geocentric theory? This is far from obvious.

Fodor briefly concedes that this is a problem for his argument. As he notes, one could argue that 'the apparent nonlocality of quotidian cognitive processes is somehow an illusion . . . *Scientific* inference may really sometimes be abductive; but then, science is social, whereas quotidian cognition, of the kind psychologists care about, is carried out in single heads. Psychology isn't, after all, philosophy of science writ small' (p. 52). Quite so. And here is how he replies to this objection: 'It strikes me as wildly implausible that the structure of human cognition changed radically a few hundred years ago' (p. 53). This is not an excerpt or a summary of Fodor's reply; it is the totality of his reply. And it is a startling non sequitur. The problem for Fodor's argument is not the difference between how the mind works today and how the mind worked a few hundred years ago. It's the difference between how a single human mind works and how the entire edifice of Western science works.

The gap between minds and computational models can be narrowed from the other side as well. Fodor argues that cognitive science is utterly clueless about how to address the problems he lumps together as abduction. 'The substantive problem is to understand, even to a first approximation, *what sort* of architecture cognitive science ought to switch to insofar as the goal is to accommodate abduction. As far as I know, however, nobody has the slightest idea' (p. 47). As a particularly damning indictment, he writes, 'The frame problem doesn't [even] make it into the index of Pinker's . . . book' (p. 42). As a matter of fact, the frame problem *is* in the index of *HTMW* (p. 639, fourth entry). And contra Fodor, cognitive scientists do have the slightest idea of what sort of cognitive architecture, at least to a first approximation, might explain abductive inference.

Recall that Fodor's touchstone for the abduction problem is Quine's analysis of the interconnectedness of knowledge. Quine (1960) wrote:

> The totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic, is a man-made fabric which impinges on experience only along the edges. Or, to change the figure, total science is like a field of force whose boundary conditions are experience (p. 42).

Quine's metaphors of a fabric and a force field, with constraints along their edges that propagate across their surface, are reminiscent of the kind of computational system sometimes known as soap-film or soap-bubble models and which cognitive scientists call constraint satisfaction networks (Attneave, 1982; Marr and Poggio, 1976; Rumelhart *et al.*, 1986; Waltz, 1975). The key idea is that a global property (in the soap film example, minimal area of a curved surface) can emerge without a master plan through numerous local interactions among the constituents (in this case, surface tension among neighboring molecules). In *HTMW* (pp. 103–109; see also pp. 233–236, 242–255) I presented an example (see figure 1, originally from

(Feldman and Ballard, 1982) of soap-film computation in a constraint network that settles on a global 3-D shape defined by locally ambiguous 2-D contours, in this case those of a Necker cube. The units represent possible interpretations of local features. The interpretations that are mutually consistent in a single 3-D object excite each other (arrows) while the ones that are inconsistent inhibit each other (dots).

More generally, a constraint satisfaction network consists of a large number of representations of possible states of affairs (the nodes) and a dense array of information pathways interconnecting them (the connections or pathways). Each node typically has a scalar value that represents the likelihood that the proposition it represents is true, and the pathways encode relationships of consistency among the propositions. Nodes representing mutually consistent propositions (where the truth of one would increase one's confidence level for the other) are connected by pathways that cause a high confidence value for the first node to increment the confidence level of the second node; nodes representing inconsistent pathways (where the truth of one leads one to doubt the other) are connected by pathways that have the opposite effect. Computation in such networks consists of setting initial values for some of the nodes, letting constraints propagate through the network in parallel, and allowing it to settle into a stable set of new values, which represents a new state of knowledge. The values for the pathways (which can be set by various combinations of innate tuning, experience with the correlations among the truth values, or deduction) are designed so that the system as a whole tends to move in the direction of global criteria such as consistency and simplicity.

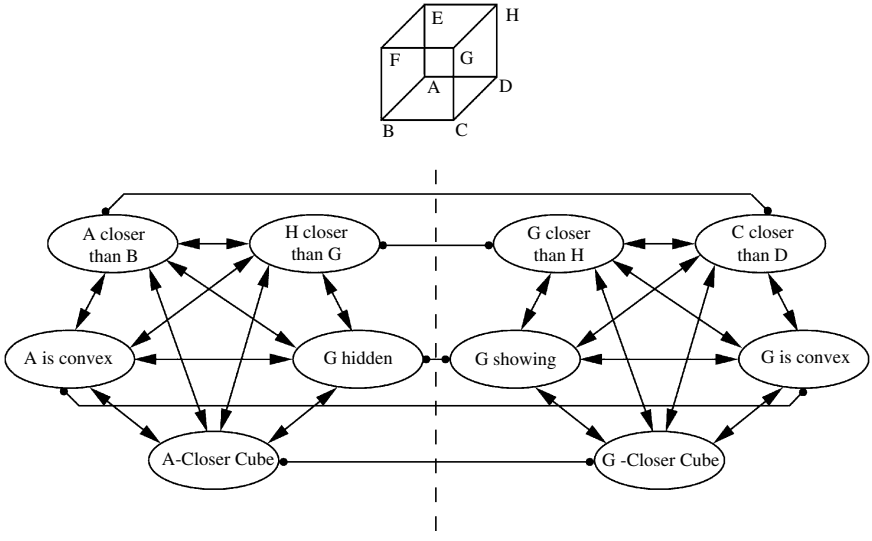

Figure 1 *Soap-film computation in a Necker cube (adapted from Feldman and Ballard, 1982)*

As I explain in *HTMW* (pp. 103–109), constraint satisfaction networks have a number of properties that help explain common-sense reasoning in the same manner as they help explain perception. One is content-addressability. We know that human memory is not searched exhaustively, nor by fixed physical addresses or file names (as in conventional computers), nor by a huge index compiled off-line (as in search engines). Rather, a concept in one set of beliefs can activate its counterpart in another set of beliefs in a single computational step. This feature falls out of the parallel, densely interconnected architecture of constraint networks. A second and related advantage is pattern completion: when some subset of a collection of mutually relevant beliefs is activated, the remaining ones are filled in automatically. Constraint satisfaction networks, at least in principle, allow far-flung information to be applied to a current problem based on overlapping content. For example, the liquid state of mercury can call to mind the liquid state of water; the associated knowledge that liquidity is not a constant property of water but one that is dependent on its temperature can then be applied back to mercury, allowing the person to abduce that the liquidity is not a necessary property of metals but a temperature-dependent one.

Constraint networks, then, are designed to do what Fodor, with his focus on Turing Machines, claims cannot be done: maintain a system of beliefs that satisfies some global property (such as consistency or simplicity) through strictly local computations. Though constraint networks are difficult to implement in standard programming languages and computer architectures (to say nothing of Turing machines), they are reminiscent of the parallel, densely interconnected, and graded signal-processing architecture of the human brain, and there is every reason to think that this is no coincidence.

The best-known constraint satisfaction networks are those developed in the connectionist school, in which the nodes tend to represent some simple feature, the propagated constraints are scalar activation levels, and the adjustments consist of summation or some other simple aggregation (e.g. Rumelhart *et al.*, 1986). Fodor detests connectionism, which he calls 'simply hopeless', because the connectionist models most popular among cognitive psychologists are ill-equipped to represent the logical structure of propositions (Fodor and Pylyshyn, 1988). I think Fodor is right about that class of models (Pinker, 1999; Pinker and Prince, 1988), but long before the ascendance of connectionism there were *symbolic* constraint satisfaction networks (e.g., Waltz, 1975), and there are now hybrids that explicitly combine the structure-sensitivity of symbol-processing architectures with the content-addressability and pattern-completion abilities of connectionist architectures (Hummel and Biederman, 1992; Hummel and Holyoak, 1997; Marcus, 2001; Shastri, 1999). Unless such hybrids are impossible in principle, which Fodor has not shown, his arguments about the limitations of Turing Machines and other serial architectures are irrelevant.

It would be a mug's game to try to defend current models of cognition against criticisms about their numerous shortcomings. Perhaps all these ideas will turn out to fail in the long run. But it makes little sense to argue about whether to be an

optimist or a pessimist when gazing into a crystal ball. The point is that Fodor claims to have identified a *principled* reason why computational approaches will not succeed at modeling cognition, and that argument is undermined by his failure to consider architectures (such as constraint-satisfaction networks) in which the computations are not necessarily serial, discrete, or sensitive only to local properties of a representation.[5]

Constraint satisfaction networks deal with representations of graded confidence levels and computations involving probabilistic inference. As such they fall under the broad rubric of heuristic reasoning, which tends toward likely solutions rather than guaranteeing correct ones. Fodor's only mention of cognitive scientists' attempt to deal with abduction is an evaluation of heuristic approaches. 'Perhaps', Fodor writes, 'real cognition in real heads achieves an appearance of abductive success by local approximations to global processes; and perhaps the problem of calculating these approximations is solved heuristically, case by case' (p. 42). But he quickly dismisses this objection by asserting, not very convincingly, that 'there is every reason to think that' the 'the inferences that are required to figure out *which* local heuristic to employ are [often] themselves abductive' (p. 42).

> If it's hard to model the impact of global considerations in *solving* a problem, it's generally equally hard to model the impact of global considerations in *deciding how* to solve a problem. . . . Suppose I'm unclear whether, on balance, in the current state of the market, it would be reasonable for me to invest in potato futures. Then I am likely to be equally unclear *how to decide* whether, on balance, in the current state of the market, it would be reasonable for me to invest in potato futures . . . I'm told that Jones advises buying potatoes; so, for practical purposes, my question whether it is wise for me to buy potatoes is reduced to the question whether it is wise for me to do as Jones advises. But what weight I ought to assign to Jones's advice itself depends a lot on what the context is. If, for example, it's *Dow* Jones, it may matter a lot that the context is financial. Deciding whether to take Jones's advice depends, in all sorts of ways, on what my prior beliefs about Jones are, just as deciding whether to buy potatoes depends, in all sorts of ways, on what my prior beliefs about the market are. There is nothing to indicate that the determinants of reliable cognitive processing become decreasingly global . . . as one goes up this hierarchy of decision making (pp. 42–43).

The key word here is *reliable*. Fodor's argument here assumes that people invariably make reliable decisions under uncertainty, such as whether to invest in potato futures. But real people's decisions are not so reliable, and they would appear to be not-so-reliable in just the way one would expect if they applied heuristics based on

[5] See (Barrett, forthcoming), for discussion of another computational architecture for cognition that escapes the limitations of Turing machines and related designs.

a few readily available cues. As the recent dot-com bubble has shown, real people tend to base investment decisions on, among other things, what they hear that everyone else is doing, what their brother-in-law advises, what a cold-calling stranger with a confident tone of voice tells them, and what the slick brochures from large investing firms recommend. People, in other words, use heuristics. It is in the very nature of common-sense inference (and, for that matter, Fodor's favored examples of visual perception and sentence processing) that available cues give statistically useful but frequently fallible information. Fodor provides no evidence that human beings have a power of reliable abductive inference that is better than heuristic, and by extension, that is beyond the reach of theories of cognition in the computational framework.

The Concept of 'Modularity' in *The Mind Doesn't Work That Way*

Constraint satisfaction architecture is one part of the solution to abduction in *HTMW*, but one still needs principles on how such networks are organized into mutually relevant collections of knowledge. This is one of the motivations for the second major theme in the book, specialization or domain-specificity. Rather than comprising a single set of rules that apply across the board to all propositions in memory, the human mind organizes its understanding of reality into several domains, such as physical objects, living things, other minds, and artifacts. Each is organized around core intuitions that guide reasoning in those domains. Physical objects occupy space, persist over time, and are subject to physical forces. Living things are self-propelled and self-organized owing to a hidden essence. Minds consist of nonmaterial beliefs and goals. Artifacts are objects fashioned by a mind to bring about a goal.

A reasoning system organized in this way can, in effect, prune the combinatorial tree of inferences that make up the frame problem. To understand how a rock falls and bounces, look at its size and shape. To understand what a plant does, test its juices and fibers. To figure out what other people are likely to do, ask about their opinions and wants. To decipher a gadget, try to figure out what its inventor had in mind. These principles inhibit the system from reasoning about rocks by asking them their opinions, reasoning about chairs by putting bits of a chair under a microscope, and other possible false leads that make abduction such a hard problem.

For the most part, Fodor assimilates the concept of specialization in *HTMW* to his own notion of *modularity*. In his 1983 classic *The Modularity of Mind* Fodor defended a conception of a mental module as an informationally encapsulated processor. Here is how he explains it in *TMDWTW*: 'A certain piece of information. . .or rule of inference is, in principle, relevant to [a] creature's success both in tasks drawn from domain *A* and in tasks drawn from domain *B*. But though the creature reliably uses the information to perform one kind of task, it seems unable to do so when it is required to perform tasks of the other kind' (p. 62). One

paradigm case is the fact that consciously knowing that a visual illusion *is* an illusion does not make the illusion go away; this suggests that some aspect of visual perception is modular. Another example is that a perfectly sensible reading of a sentence may be unavailable to a person if his parsing mechanism has trouble assigning a correct syntactic parse to the sentence; for example, *Ashley said that Sue will leave yesterday* seems contradictory, even though it has a grammatical and plausible interpretation in which Ashley said it yesterday. This suggests that sentence parsing is modular.[6]

Fodor acknowledges that a mind with some kind of modular design could, in principle, meet the abduction challenge. But he immediately attempts to demolish the suggestion that the human inference system in fact has such a design. Syntax and vision, he argues, might be modules, but domains of reasoning cannot be.

As with the concept of computation, the concept of modularity has a number of meanings, and it's important to understand what work the concept does in *HTMW*. *HTMW* took pains to distinguish modules in Fodor's strong sense of encapsulated processors with modules in a much weaker sense of domain-specific functional organization (e.g., pp. 30–31, pp. 314–315). The subsystems in *HTMW* are not categorically sealed off from information that could be relevant to them, and they were posited not just to serve as a mail router for incoming information but rather to dictate what *kinds* of inferences and goals should be triggered by that information. A module for sexual attraction, for example, doesn't just make people pay attention to siblinghood; it says specifically, 'don't find your sibling sexually attractive'.

To be fair, in *TMDWTW* Fodor does distinguish among his own encapsulated processors and my functionally specialized mechanisms (chap. 4). But after making this distinction, he blurs it all over again. He identifies the central thesis of *HTMW* as 'massive modularity', which 'means that there is a more or less encapsulated processor for each kind of problem it can solve' (p. 64). But as mentioned, in *HTMW* (pp. 30–31, 314–315) the specializations need *not* be strictly encapsulated (though they are biased to consider some kinds of information before others), and their number is by no means 'massive'. Given the complexity of human behavior, a theory that posits some two dozen emotions and reasoning faculties (distinguishing, for example, fear from sexual jealousy from the number sense) is far from profligate (especially compared to Fodor's tolerance for the possibility that people are born with 50,000 concepts).

Fodor's imputation of his own version of modularity to *HTMW* is intended, I believe, to be charitable. A massive number of encapsulated modules *could* meet the abduction challenge if it were true, he concedes, since it would confine an inference engine to the information relevant to solving a problem. But in fact, he goes on to argue, it is *not* true, so that option is not available. His objection to

[6] As it happens, few psycholinguists believe that sentence parsing is as modular as Fodor argued, though I believe he was probably not completely wrong about it either; see Pinker, 1994, chapter 7.

multiple reasoning systems is based on the same diagnosis of regress that he invoked when arguing against heuristics. Fodor claims that the problem of routing information to the appropriate reasoning system—for example, detecting that an event is an example of a social exchange in order to activate a cheater-detecting mechanism—requires nothing less than a full solution to the abduction problem, leaving us where we started: 'Nobody has *any idea* what kind of cerebration is required for figuring out which distal stimulations are social exchanges' (p. 76).

In fact, psychologists have had an idea for more than half a century, at least since Fritz Heider (Heider and Simmel, 1944) showed that people automatically interpret certain patterns of moving dots as agents that seek to help and hurt each other—just the conceptual elements that are relevant to social exchange. Fodor never mentions this phenomenon, though he does make a rather odd argument against the general idea that domain-specific reasoning systems might be triggered by psychophysical cues. Any attempt to ground cognition in sensation would be tantamount to British empiricism, Fodor writes, and evolutionary psychologists (and other advocates of domain-specificity) claim not to be empiricists. The argument is odd because it is a form of genre criticism: shoehorning people into an orthodoxy and criticizing them for not adhering to it.

Of course, spatiotemporal trajectories are not the only or even the primary way that people recognize cognitive domains such as social exchange. Let's grant that not all perceptual information is earmarked by some psychophysical cue that can be used to shunt it to the most relevant reasoning system. Let's say that the input to one system may come from the output of another system. Perhaps the social exchange system is fed by an intuitive psychology that infers people's goals from their behavior. Perhaps sexual emotions are fed, in part, by a system for inferring who is related to the self. Perhaps some of these input systems are domain-general with respect to reasoning systems, feeding a number of them with information about objects or bodies or actions. This may speak against the mind as a collection of encapsulated modules, each wired directly to the eyeballs. But it does not speak against the mind as a network of subsystems that feed each other in criss-crossing but intelligible ways—the organ system metaphor on which *HTMW* is based.[7]

The Dismissal of Evolution in *TMDWTW*

Fodor advances four arguments that evolution has nothing to add to our understanding of how the mind works.

1. Fitness and truth. Treating the mind as an organ whose ultimate function is to promote Darwinian fitness, Fodor claims, has no advantage over the biologically

[7] See (Barrett, in press) for similar arguments, using 'enzymes' rather than organ systems as the central metaphor.

untutored view that the mind is an organ whose function is to arrive at the truth. 'There is nothing in the "evolutionary", or the "biological", or the "scientific" worldview that shows, *or even suggests*, that the proper function of cognition is other than the fixation of true beliefs' (p. 68). To suggest otherwise, he claims, is 'neo-Darwinist anti-intellectualism'.

Putting aside the scope error in the anti-intellectualism charge, Fodor's claim that 'truth is cognition's proprietary virtue' runs into an obvious empirical problem: many kinds of human beliefs are systematically false. Members of our species commonly believe, among other things, that objects are naturally at rest unless pushed, that a severed tetherball will fly off in a spiral trajectory, that a bright young activist is more likely to be a feminist bankteller than a bankteller, that they themselves are above average in every desirable trait, that they saw the Kennedy assassination on live television, that fortune and misfortune are caused by the intentions of bribable gods and spirits, and that powdered rhinoceros horn is an effective treatment for erectile dysfunction. The idea that our minds are designed for truth does not sit well with such facts.

And contrary to Fodor's claim that nothing in the evolutionary worldview '*even suggests*' that the function of cognition is something other than believing true things, here are five things that suggest exactly that.

First, computing the truth has costs in time and energy, so a system designed for useful approximations (one that 'satisfices' or exhibits bounded rationality) might outcompete a system designed for exact truth at any cost. There is little point, for example, in spending twenty minutes figuring out a shortcut that saves you ten minutes in travel time.

Second, outside the realm of mathematics and logic, there is no such thing as a universal true-belief-fixer. Inductive inference systems must make fallible assumptions about the world, such as that surfaces are mostly cohesive, human languages conform to a universal grammar, and people who grow up with you are your biological siblings. If the world for which the system was designed has changed, those beliefs may be systematically false. Visual illusions are a prime example. In other words, there is an important difference between a system designed to fixate likely beliefs in an ancestral world and a system designed to fixate true beliefs in this world.

Third, beliefs have a social as well as an inferential function: they reflect commitments of loyalty and solidarity to one's coalition. People are embraced or condemned according to their beliefs, so one function of the mind may be to hold beliefs that bring the belief-holder the greatest number of allies, protectors, or disciples, rather than beliefs that are most likely to be true. Religious and ideological beliefs are obvious examples.

Fourth, publicly expressed beliefs advertise the intellectual virtuosity of the belief-holder, creating an incentive to craft clever and extravagant beliefs rather than just true ones. This explains much of what goes on in academia.

Fifth, the best liar is the one who believes his own lies. This favors a measure of self-deception about beliefs that concern the self.

The idea that the mind is designed for truth is not completely wrong. We do have some reliable notions about the distribution of middle-sized objects around us and the quotidian beliefs and desires of our friends and relatives. But the statement that the mind is designed to 'find out truths' would seem to be a rather misleading summary of the past fifty years of research on human reasoning.

2. Consilience. Fodor is puzzled by the idea that psychology might benefit by being connected to evolutionary biology, an idea that he calls 'a little odd. The New Synthesis is, after all, prepared to allow that psychology and botany, for example, actually don't have much to say to one another; let those chips fall where they may' (pp. 80–81). Similarly, he argues, astrophysical theory has few implications for botany, quantum mechanics is irrelevant to demography, and lunar geography does not constrain cellular mitosis. Why should it be any different for 'your favorite theory about how the mind works and your favorite theory of how evolution works?' (p. 82).

Here is why it should be different. The subject matter of psychology is the functioning of the brain. The subject matter of botany is plants. The brain is not a plant. Now, the subject matter of evolutionary biology is living things. The brain *is* a living thing. Therefore, the relationship between psychology and evolution is not the same as the relationship between psychology and botany (or the relationship between lunar geography and cellular mitosis, and so on). If anything is 'a little odd', it is Fodor's failure to distinguish pairs of disciplines whose subject matters are in a superset-subset relation from pairs of disciplines whose subject matters are disjoint. Fodor repeats his non-sequitur when he writes, 'It simply isn't true that all the sciences are mutually relevant'. The issue, of course, is not whether *all* the sciences are mutually relevant but whether evolutionary biology and psychology (and other pairs of sciences with overlapping subject matters) are mutually relevant.

Indeed, Fodor extends his argument from 'not all' to 'most not': 'Quite the contrary', he writes, 'most sciences are quite strikingly mutually irrelevant . . . It's generally hard work to get theories in different sciences to bear on one another' (p. 83). This strikes me as a remarkable misreading of the current state of science. A glance at a university catalogue or funding agency provides literally dozens of examples in which pairs of sciences are mutually relevant: astrophysics, astrobiology, atmospheric chemistry, biochemistry, biogeography, biophysics, chemical biology, geophysics, geochemistry, molecular biology, molecular genetics, physical chemistry, and on and on. A growing plaint among scientific and academic policy-makers is that disciplinary divisions are fossils of Nineteenth century ways of organizing knowledge and an impediment to scientific progress.

3. Teleology. Fodor argues that invoking *function* in a psychological explanation is logically independent of invoking *natural selection*. The strong connection between function and selective history in evolutionary psychology, Fodor writes, is

> . . . an uncomfortable feature of the Darwinian account of teleology, one that makes it hard to believe that it could be the one that biological/psychological explanation requires. Imagine, just as a thought experiment, that Darwin was comprehensively wrong about the origin of species . . . Would it then follow that the function of the heart is not to pump the blood? Indeed, that the heart, like the appendix, has no function? (p. 85).

But far from being an 'uncomfortable' feature, the logical independence of biological functionality and natural selection is what gives Darwinism its empirical content. A common (and lazy) criticism of the theory of natural selection is that it is circular. According to the criticism, Darwinism means 'survival of the fittest' but 'the fittest' is defined as 'what survives'. Or, natural selection says only that whatever gets selected gets selected. By noting that biological functionality can be identified independently of any invocation of natural selection, Fodor, to his credit, shows why such arguments are fallacious. Natural selection is a falsifiable scientific explanation of how biological functionality arises, not a part of the concept of functionality itself.

On the other hand, from a scientist's perspective functionality without natural selection is unacceptably incomplete. Adaptive organs such as the eye or heart are staggeringly improbable arrangements of matter, and we need an explanation as to how they come to exist. Faced with this puzzle, the only alternatives to natural selection are deliberate engineering by a deity or extraterrestrial; some kind of mysterious teleological force that allows future benefit to affect present design; and simply not caring. The last appears to be Fodor's preference, but there is no reason that other scientists should be so incurious.

Natural selection, moreover, does more than solve the puzzle of how biological functionality arises. It can also feed back to revise and constrain our characterization of a function itself. For example, if the explanation of biological functionality in terms of natural selection is correct, we can rule out adaptations that work toward the greater good of the species, the harmony of the ecosystem, beauty for its own sake, benefits to entities other than the replicators that create the adaptations (such as horses which evolve saddles), functional complexity without reproductive benefit (e.g. an adaptation to compute the digits of pi), and anachronistic adaptations that benefit the organism in a kind of environment other than the one in which it evolved (e.g., an innate ability to read, or an innate concept of 'carburetor' or 'trombone').

Natural selection also has a positive function in scientific discovery, impelling psychologists to test new hypotheses about the possible functionality of aspects of psychology that previously seemed functionless. Numerous success stories are recounted in *HTMW*, such as the hypothesis that social emotions (sympathy, trust, guilt, anger, gratitude) are adaptations for policing reciprocity in non-zero sum games, and that an eye for beauty is an adaptation for detecting health and fertility in potential mates. Conversely, other psychological traits, such as music and religion, are recalcitrant to any rigorous analysis of adaptiveness in the

evolutionary biologist's sense; they are better explained as by-products of adaptations. None of this research would be possible if psychologists had satisfied themselves with a naïve notion of function instead of the one licensed by modern biology.

4. Complexity. Fodor's final dismissal of evolution consists of a rejection of the argument that adaptive complexity requires an appeal to natural selection:

> . . . the complexity of our minds, or of our behavior, is simply irrelevant to the question of whether our cognitive architecture evolved under selection pressure (p. 87). . . . It's entirely possible that quite small neurological reorganizations could have effected wild psychological discontinuities between our minds and the ancestral ape's (p. 87–88).

The problem with this argument is that it confuses complexity with *adaptive* complexity, that is, improbable functionality. Fodor may be correct that as-yet-unknown changes in the developmental program for a primate brain could increase its complexity, for example, by giving it more neurons, a more intricate tangle of connections, or a more tortuous 3–D shape. But this is entirely different from increasing its *functionality* by making it better equipped to solve problems such as mate selection, coalition building, or toxin avoidance. The reason is that the proximate physical mechanisms that constitute our neurodevelopmental program—axon guidance signals, neural growth factors, cell adhesion molecules, and so on—cannot 'see' their effects on the functioning of the whole organism in its social and physical milieu. Natural selection *can* see those effects, and thereby can shape, over generations, just those developmental variations that enhance them.

Fodor, ironically, concedes a related point:

> . . . what is surely not conceivable is that relatively small, fortuitous changes in brain structure should produce massive increments in a creature's stockpile of true, contingent beliefs. . . . barring the rarest of accidents, it's simply not conceivable that a large database of logically independent, contingent beliefs that was formed fortuitously (e.g., in consequence of random alterations of brain structure) could turn out to be generally true. To get the feel of the thing, imagine cutting up the Manhattan telephone directory and then pairing all the numbers with all the names at random. How often do you suppose the number thus assigned to someone would be the number that he actually has? (pp. 93–94).

But Fodor's argument concerning beliefs that are contingently true in an environment applies in equal force to biological mechanisms that are contingently *fit* in an environment—that is, to mechanisms that attain some improbable state that enhances the organism's chances at reproduction. As Richard Dawkins (1986)

has put it, 'However many ways there may be of being alive, it is certain that there are vastly more ways of being dead, or rather not alive. You may throw cells together at random, over and over again for a billion years, and not once will you get a conglomeration that flies or swims or burrows or runs, or does *anything*, even badly, that could remotely be construed as working to keep itself alive' (p. 9).

Summary and Conclusion

In *HTMW*, I defended a theory of how the mind works that was built on the notions of computation, specialization, and evolution. Specifically, it holds that the mind is a naturally selected system of organs of computation. Fodor claims that 'the mind doesn't work that way' because (1) Turing Machines cannot do abduction, (2) a massively modular system *could* do abduction but cannot be true, and (3) evolution adds nothing to our understanding of the mind. In this paper, I have presented four reasons that Fodor's argument doesn't work.

First, the claim that the mind is a computational system is distinct from the claim that the mind has the architecture of a Turing Machine or some other serial, discrete, local processor. Therefore the practical limitations of Turing Machines are irrelevant.

Second, abduction—conceived as the cumulative accomplishments of the scientific community over millennia—is distinct from human common-sense reasoning. Therefore Fodor's gap between human cognition and computational models may be illusory.

Third, biological specialization, as seen in organ systems, is distinct from Fodorian encapsulated modules. Therefore the limitations of Fodorian modules are irrelevant.

Fourth, Fodor's arguments dismissing the relevance of evolution to psychology are unsound. Human cognition is not exclusively designed to arrive at true beliefs. Evolutionary biology is more relevant to psychology than botany is to astronomy. Biological function without natural selection is woefully incomplete. And adaptive complexity requires a non-fortuitous explanation, just as true beliefs do.

Some final thoughts. It should go without saying that we *don't* fully understand how the mind works. In particular, we don't have a complete theory of how the mind accomplishes feats of common sense and scientific inference. Scientific psychology is not over. On the other hand, Fodor has failed to show that there is some known, *in-principle* chasm between the facts of human cognition and the abilities of biologically plausible computational systems. Chicken Little is wrong, and more, not less, research needs to be done.

Harvard University
Cambridge, MA

References

Attneave, F. 1982: Pragnanz and soap bubble systems: A theoretical exploration. In J. Beck (ed.), *Organization and Representation in Perception*. Mahwah, NJ: Erlbaum.

Barrett, H.C. forthcoming: Enzymatic computation and cognitive modularity. *Mind & Language*.

Dawkins, R. 1986: *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design*. New York: Norton.

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

Fodor, J.A. 1968: *Psychological Explanation: An Introduction to the Philosophy of Psychology*. New York: Random House.

Fodor, J.A. 1975: *The Language of Thought*. New York: Crowell.

Fodor, J.A. 1981: *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Fodor, J.A. 1983: *The Modularity of Mind*. Cambridge, Mass.: MIT Press.

Fodor, J.A. 1994: *The Elm and the Expert: Mentalese and its Semantics*. Cambridge, Mass.: MIT Press.

Fodor, J.A. 2000: *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press.

Fodor, J.A. and Pylyshyn, Z. 1988: Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3–71.

Heider, F. and Simmel, M. 1944: An experimental study of apparent behavior. *American Journal of Psychology*, 57, 243–259.

Hummel, J.E. and Biederman, I. 1992: Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99, 480–517.

Hummel, J.E. and Holyoak, K.J. 1997: Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427–466.

Marcus, G.F. 2001: *The Algebraic Mind: Reflections on Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Marr, D. and Poggio, T. 1976: Cooperative computation of stereo disparity. *Science*, 194, 283–287.

Pinker, S. 1979: Formal models of language learning. *Cognition*, 7, 217–283.

Pinker, S. 1994: *The Language Instinct*. New York: HarperCollins.

Pinker, S. 1997: *How the Mind Works*. New York: Norton.

Pinker, S. 1999: *Words and Rules: The Ingredients of Language*. New York: HarperCollins.

Pinker, S. and Prince, A. 1988: On language and connectionism: Analysis of a Parallel Distributed Processing model of language acquisition. *Cognition*, 28, 73–193.

Plotkin, H. 1997: *Evolution in Mind*. London: Allen Lane.

Quine, W.V.O. 1960: Two dogmas of empiricism. In *From a Logical Point of View*. New York: HarperCollins.

Rumelhart, D.E., Smolensky, P., McClelland, J.L., and Hinton, G.E. 1986: Schemata and sequential thought processes in PDP models. In J.L. McClelland and

D.E. Rumelhart (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Psychological and Biological models*. (Vol. 2). Cambridge, MA: MIT Press.

Shastri, L. 1999: Advances in SHRUTI: A neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence*, 11, 79–108.

Tooby, J. and Cosmides, L. 1992: Psychological foundations of culture. In J. Barkow, L. Cosmides and J. Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press.

Turing, A.M. 1950: Computing machinery and intelligence. *Mind*, 59, 433–460.

Waltz, D. 1975: Understanding line drawings of scenes with shadows. In P.H. Winston (ed.), *The Psychology of Computer Vision*. New York: McGraw-Hill.

Wilson, E.O. 1998: *Consilience: The Unity of Knowledge*. New York: Knopf.