# 37

# Jerry Fodor (1935– )

GEORGES REY

Jerry Fodor is widely regarded as the most significant philosopher of mind in recent times. With Noam Chomsky at MIT in the 1960s he mounted a decisive attack on the behaviorism that then dominated psychology and most philosophy of mind, and has tried to present in its place a naturalistic and realist account of mental processes that renders them amenable to serious scientific study.

Indeed, he is one of the few philosophers who has combined philosophical and empirical psychological research, publishing work in both domains, developing at least two theories that have become highly influential in both of them: the computational/representational theory of thought processes (see the section "CRTT: Computation") and the modularity theory of perception ("Modularity and the Limits of CRTT"). These theories are, however, best appreciated against the backdrop of a number of other themes in Fodor's work, which provide the best overview of his work, as follows: (1) Intentional Realism; (2) Nomic Explanation; (3) The Problems of Mind; (4) CRTT: Computation; (5) CRTT: Representation; (6) Solipsism and Narrow Content; (7) Nativism; (8) Modularity and the Limits of CRTT.

## Intentional realism

Fodor's primary concern is to defend the familiar "belief/desire," or "propositional attitude" psychology with which the folk routinely explain each other's behavior; for example, someone's heading south is explained by their wanting water and thinking there's some there. What we might call "scientific intentional realism" is simply a way of taking folk psychology seriously as the beginning of a serious scientific psychology.

Not that the folk are always right about the mind. Indeed, many of the claims and even the specific terms they employ: "learning," "memory," perhaps even "belief" and "desire," may well turn out to be theoretically inadequate. Fodor only presumes that, whatever the particular kinds of phenomena invoked by an ultimate psychology, they will display certain crucial properties:

1 as some or other species of *propositional attitude*, they will be *intentional* (being "about" things) and *semantically valuable* (capable of being *true or false*);
2 as phenomena involved in rational thought, they need to be *logically structured*;

3 as ultimately explanatory of action, they need to be *causally efficacious*; and
4 given the fundamental role of physics in our understanding of the world, they had better be *composed of or identical to* physical phenomena.

Fodor's postulation of states displaying these properties may seem rather truistic until one notes the quite substantial efforts of a good number of philosophers and psychologists in the twentieth century to argue otherwise: there are not only the usual Cartesian dualists, who resist the materiality of the mind (and hence (4)), but more recently there have been the logical positivists, behaviorists, Wittgenstein, Quine, Gibson, Davidson, Dennett, the Churchlands, many connectionists, all of whom have tried in one way or another either to deny the causal reality of the mind altogether, or to relegate mentalistic ways of describing the world to some sort of "second grade status," in some way less objective than physics (denying (3)). Much of Fodor's work consists in defending intentional realism against these attacks, not only as they arise in philosophy, but particularly in relation to psychology, where what is at stake are entire research programs committed to behaviorism, connectionism, neurophysiology – or, as he would recommend instead, to intentional realism (1968a, 1998b: chs 1, 8–10).

Fodor is specifically concerned with the kinds of challenges to a materialist mentalism that were raised by Descartes's concern with rationality, and Brentano's with intentionality, and so is consequently most concerned to be a realist both about attitude *states*, such as belief or desire, and about their *contents* (the belief that *snow is white* or *God is dead*). Indeed, he is adamant that the latter are decidedly not to be explained away as matters of "interpretation" or mere "similarity relations" (1998a: 30–4). He is also as realist as anyone about "qualia" and consciousness, but has relatively little to say about them. He is convinced that empirical psychology at least since Freud has given us reason to suppose that rational and intentional phenomena needn't be conscious (1968a, 1998b), and that he can therefore address the formidable difficulties of the former without worrying about what he regards as the currently impenetrable problems posed by the latter (1994: 121, and 1998b; but see 1972 and 1998b: 73 for some stray substantive remarks).

## Explanation as nomic subsumption

Fodor also takes it for granted that explanation in general is subsumption under laws, and that the realm of the mental is no exception (1994: 3, 1998a: 7). Except possibly in ultimate physics, he assumes these are *ceteris paribus* laws, or laws that are not "strict" and "exceptionless," but hold in abstraction from various interferences or "completers" that a fuller theory of the world might include (1991c). Much of his view here is of a piece with the kind of idealization that Chomsky noted is typical of any science (see CHOMSKY). But whereas Chomsky is largely concerned with only a specific set of idealizations – those capturing linguistic "competence" in abstraction from its "performance" – Fodor is concerned with what he regards as the necessary variety of them that are enlisted in the explanation of psychological *processes*.

Towards addressing this concern, Fodor (1968a) presented one of the first lengthy defenses of *functionalism*, according to which psychological states are

individuated by their causal relations. Since different physical phenomena can satisfy these relations, functionalism naturally gives rise to cross-classificatory *layers* of explanation: one level of causal relations may be "multiply realized," or variously "implemented," by different mechanisms at a lower level (1968a, 1998b: ch. 2). Specifically, the *intentional level* of a cognitive psychology may be implemented at a lower level by various computational/syntactic processes (§4), which in turn may be implemented by different physical mechanisms – brains in the case of people, transistors in the case of machines. (Note, though, that Fodor is nevertheless skeptical one can provide any *definitions* of mental states, functionalist or otherwise; see §5.1).

Especially in Fodor's work, this functionalist conception is responsible for a considerable "autonomy" of cognitive psychology from details of its implementation, analogous to the way a computer program can be specified in abstraction from the electronic details of the computers that run it. Because they are genuine laws, involving, for example, "projectible" predicates, the kinds they mention are not reducible to mere finite disjunctions of the kinds at the lower levels (1998b: ch. 2), although Fodor presumes that they "supervene" on them.

## The demand for mind

If we live in a purely physical universe, however, it might be wondered what serious explanatory role mental phenomena have to play. Why doesn't physics alone suffice? The short answer is, of course, that purely physical processes can come to exhibit all manner of special structures and organizations – molecules, crystals, cells, living organisms, and sometimes minds – that it is the business of special "macro"-sciences to describe.

Fodor is particularly impressed by the sensitivity of human beings to indefinitely many non-local, non-physical properties: not only, as Chomsky has emphasized, to highly *abstract* grammatical properties, like being a morpheme or a noun phrase, but also to arbitrary *non-physical* or *non-local* properties, such as *being a crumpled shirt*, *a grieving widow*, or *a collapsing star* (1986). These sensitivities are particularly impressive given that they seem to be (1) *productive* and (2) *systematic*. (1) People seem capable of discriminating stimuli of indefinite logical complexity, such as *being a crumpled shirt that was worn by the thief who stole the cat that chased the rat . . .* (1975a: ch. 1); and (2) anyone capable of discriminating one logical form is capable of discriminating logical permutations of it; for example, one can discriminate *John's loving Mary* if and only if one can discriminate *Mary's loving John* (1987b: 147ff.).

A good deal of Fodor's work has been devoted to showing that no non-mentalistic account can explain these phenomena. Thus he has argued at length that purely physicalist, behaviorist, Gibsonian, syntactic, and eliminative connectionist accounts of behavior are either vacuous or empirically inadequate (1968a, 1981b, 1987b: 161–3, 1988a, and 1991a). It is difficult to see how any physical mechanism could be sensitive to such an extraordinary range of arbitrary properties of the world without exploiting internal processes of logical combination, inference, and hypothesis confirmation that essentially involve phenomena satisfying the four demands listed on pp. 451–2.

## CRTT: Computation

Fodor's main proposal for meeting those four demands is the *computational/representational theory of thought* (CRTT). Indeed, much of his work can be regarded as an effort to incorporate into psychology Alan Turing's crucial work on mechanical computation, according to which certain rational processes could be realized mechanically: for example, each of the rules of logic can be shown to involve mechanical operations on the formally specified sentences of a formal language. Fodor regards this as promising for psychology, since, he argues, people at least sometimes engage in the sort of *truth*-preserving inferential processes captured by logic (1994: 9). This already marks an important break with traditional psychology, which tended to rely on mere *associations* among ideas (Hume) or stimuli (Skinner). These seem incapable of capturing the relation between, for example, the premises and the conclusion of a valid argument; mere associations, like that between "salt" and "pepper," are neither necessary nor sufficient for understanding valid arguments.

This concern with logic and truth also commits psychology to vehicles capable of the requisite representational richness. Traditional empiricist psychology (as in Locke and Hume) tended to be not only associationist, but also to regard mental representations (or "ideas") as *images*. But although images may have a role to play in thought (1975a: 174–94), it is doubtful that they are remotely adequate for the expression of it in general. What image could express the conditionals, quantifiers, negations, and modals in such a thought as *If everyone drinks then no one should drive*? Merely a picture of a lot of non-driving drinkers won't quite do. The only vehicles that seem remotely capable of expressing such thoughts are the logico-syntactic vehicles of a language, natural or artificial, with precisely the resources of operators (quantifiers, connectives) and referential devices (predicates, variables, names) that we ordinarily use to express those thoughts. That is, there must be some sort of language in which a thinker thinks, a "language of thought" (an "LOT" or "mentalese").

Talk of "sentences" in the brain mustn't be taken on the model of sentences as they are inscribed on pages of books. Sentences are highly abstract objects that can be entokened in an endless variety of ways: as wave forms (in speech), as sequences of dots and dashes (Morse code), as sequences of electrically charged particles (on recording tape). It is presumably in something like the latter form that sentences would be entokened in the head. Indeed, CRTT is best viewed as simply the claim that the brain has logically structured, causally efficacious states, a thesis that, whatever its merits, isn't *patently* absurd. (Note also that this is not a thesis that is supposed to be *introspectibly* plausible: CRTT does not entail that people's mental lives should appear "introspectively" to involve sentences, much less sentences of a natural language.)

An extremely simple version of CRTT could be true of an intelligent system in the following way: there are sensory modules (e.g. visual and auditory systems, see "Modularity and the Limits of CRTT," below) that transduce ambient energy forms into electrical signals that in turn produce structured sentences as input to a central cognitive system (perception). This central system selects certain sentences from a pre-established (perhaps innate) set, tests their deductive consequences against this input

for a "best fit," and produces as output those sentences that pass that test above threshold (belief). These sentences in turn may be the input to a decision-making system in which, on the basis of that input, innate preferences, and utility functions, a course of action is determined, that is, a basic act-description is selected (intention) that causes a basic act satisfying that description to be performed (action).

A common objection to such an account is that it requires "homunculi" in the brain, possessing precisely the same intelligence that is supposed to be explained, in order to "read" these sentences. However, what Turing's theory of computation has taught us is that complex computational processes, such as operations upon symbols in a language, can be broken down into simpler operations that are eventually so simple that (it's obvious that) a machine can execute them (1975a: 73) without at that point exploiting any intelligence at all.

Besides the straightforward argument from the expressive power of a language, Fodor advances a number of other reasons for CRTT:

1 It is presupposed by standard theories of perception, hypothesis confirmation, and decision-making, all of which involve the agent representing the world and assigning utilities to the consequences of various courses of action (1975a: ch. 2).
2 It is able to explain the *truth*-preserving transitions in thought of which rational creatures are at least *sometimes* capable: for example, the ability to deduce "Rats die" from "Cats live" and "If cats live then rats die" (1994: 9).
3 It is easy to conceive of computational architectures exploiting an LOT that would explain the productivity and systematicity of our minds mentioned earlier, for example, a machine that was sensitive to syntactic structures could in standard recursive ways produce indefinitely complex representations (productivity), and could access one representation if and only if it could access a logico-syntactic permutation of it (systematicity) (1987b).
4 It offers a perspicuous account of the intensionality of thought: Oedipus' thinking he'll marry Jocasta is distinct from his thinking he'll marry his mother since the vehicles of the two thoughts (the LOT equivalents of "Jocasta" and "my mom") are different.

It is at least these four arguments that Fodor deploys both against traditional associationism in psychology, as well as against what he regards as its contemporary manifestation, radical connectionism (1988a, 1991a). (Against connectionist proposals that would merely propose a novel *implementation* of CRTT, Fodor has no objection.)

Despite his commitment to a CRTT, Fodor has doubts about its eventual scope, to which we'll return below (in "Modularity and the Limits of CRTT").

## CRTT: Representation

So far, we've discussed Fodor's views about mental *processes*, as computations over logico-syntactic representations. Fodor would be the first to recognize that this is at best only half an account of mentality: computations may well preserve truth better than associations do, but where do these representations acquire any semantic properties like truth in the first place?

## *Inferential role theories*

In early work (1963), Fodor was drawn to what is, broadly speaking, an "inferential role semantics" (IRS). This is a family of views according to which the meaning of an expression has to do with its inferential relations to other expressions, as in the case of "bachelor" entailing "unmarried." These relations might involve definitions (as in traditional "analyses" and "meaning postulates" in philosophy), "semantic decomposition" in linguistics, "procedural semantics" in artificial intelligence, or "prototypes" and "whole theories" in psychology.

By the late 1970s Fodor became convinced that the standard arguments for IRS suffered from serious empirical and philosophical difficulties: proposed linguistic decompositions were seriously inadequate; there was an embarrassing paucity of psychological evidence for anything like definitions (1975b), which the history of analytic philosophy had shown were notoriously difficult to provide in the first place (1970); and Quine had cast serious doubt on whether there could ever be any theoretically satisfactory way of distinguishing constitutive inferential relations (the "analytic") from merely common beliefs (the "synthetic") (1998a) (see QUINE). Indeed, although Fodor has no patience with Quine's behaviorism, he wholeheartedly endorses his rejection of definitions as having any serious explanatory status in any science whatsoever (the intuitive appearance of an "analytic/synthetic" distinction, Fodor argues, is due either to the "centrality" of a claim to one's thought, or to its involving "one-criterion" concepts (1998a: 80–6)).

Under the influence of Quine, and especially of his dictum, "The unit of meaning is the whole of science," many IRS theorists have themselves tended to forgo the analytic/synthetic distinction and regard *all* of an expression's inferential relations as constitutive. This "meaning holism" has the disturbing consequence that it would be virtually impossible for two people ever to mean exactly the same thing, indeed, for even one person to mean the same thing over any change of belief – rendering memory impossible! Fodor thinks that this renders any serious psychological generalizations impossible as well, and so is at pains to block the many arguments for it (1992a), and for any IRS, which, he thinks, inevitably invites it.

In Fodor's view, the original sin endemic to IRS theories consists in conflating *semantics* (or a theory of the content of concepts) with *epistemology* (or a theory of how we apply concepts). This conflation not only burdens semantics with the notorious problems of a verifiability theory of meaning that lurks in most of the above IRS proposals, but also presents substantial problems for accounting for the aforementioned productivity and systematicity of thought. Fodor argues that these latter phenomena require a compositional semantics (i.e. one in which the meaning of a complex expression is a function of the meaning of its parts), and epistemological capacities are not in general compositional: one could know a lot about pets and a lot about fish without knowing much at all about pet fish, for example, what one typically looks like (1990b, 1998b: chs 4–5). Unconfounding epistemology and semantics, Fodor instead forgoes any "molecular" account of meaning that depends upon relations among symbols, and instead embraces an "atomistic" theory that requires only that a symbol stand in a specific relation to the external world.

## *Information theories*

Fodor takes as his point of departure the "information" theoretic semantics developed by Fred Dretske, which treats semantic meaning as a species of "natural" meaning (whereby dark clouds "mean" rain, or smoke fire). This idea often appears in psychological discussions under the guise of *discrimination abilities*: for example, something is a "shape receptor" if and only if it reliably discriminates shapes. The idea is naturally spelt out in terms of certain *counterfactual* dispositional properties to co-vary with specific phenomena in the world.

So stated, information semantics is open to several immediate objections.

1 pan-semanticism: something needs to be said about what's special about semantic or psychological meaning, since everything is causally related (and so "carries information") about *something*;
2 transitivity: "information" is transitive, but meaning isn't: if "smoke" co-varies with smoke, and smoke, itself, with fire, then "smoke" co-varies with fire; but "smoke" doesn't mean fire (1990b: 93);
3 robustness: *most* tokenings of sentences are produced in the *absence* of the conditions that they nevertheless mean. "That's a horse" can be uttered on a dark night in the presence of a cow, or just idly in the presence of anything. Fodor (1987b) calls these latter usages "wild," and the property whereby tokens of symbols mean things that aren't on occasion their actual cause, "robustness."
4 In accounting for robustness, a semantic theory needs to say what distinguishes the "wild" from the meaning-constitutive causes, a problem made vivid by the "disjunction" problem: what makes it true that some symbol "F" means HORSE and not HORSE OR COW ON A DARK NIGHT, or HORSE OR COW ON A DARK NIGHT OR $W^2$ OR $W^3$ OR . . . (where each $W^i$ is one of the purportedly "wild" causes) (1987b).

A fifth problem could be raised regarding the contents of logical and mathematical symbols, which do not obviously enter into causal relations with any worldly phenomenon. Fodor sets aside this problem for the nonce, although suspecting that they are the only symbols for which an IRS is plausible.

## *Teleological views*

A natural suggestion regarding the meaning-constitutive conditions is that they are in some sense "optimal" conditions that obtain when nothing (e.g. poor vision, limited spatiotemporal access) is "interfering" with belief fixation, and it is functioning as it was "designed to." Fodor (1987b) calls such theories "teleological" and he himself proposed a version of one in the widely circulated paper called "Psychosemantics" (1990a) (to be distinguished from the book (1987b), of the same title in which he *rejects* any such theory!). The attraction of such a theory lies in its capturing the idea that two individuals *meaning* the same thing by some symbol consists in their agreeing about what it would apply to, *were they to agree about everything else*. Their disagreements are to be explained as due to their differing epistemic positions and reasoning capacities.

Although Fodor nowhere suggests such theories are false, he does think they are subject to a number of difficulties, the chief one consisting of the circularity that seems

unavoidable in specifying the optimal conditions: it would appear that those conditions cannot be specified without employing the very intentional idiom the theory is supposed to explain (1987b: 104–6).

In order to avoid this and other problems Fodor (1987b, 1990b) went on to propose his "asymmetric dependency" theory. Although it makes no explicit appeal to ideal epistemic conditions, much of its motivation can be appreciated by thinking of the ideal co-variational theory in the background.

## *Asymmetric dependency*

According to the ideal co-variational theory, tokens of an expression may be "wild," that is, produced by a property it doesn't express. Now, one way to understand the asymmetric dependency theory is first to notice that, plausibly, *all such wild cases depend upon the ideal case, but not vice versa*: the wild tokenings depend upon the ideal ones, but the ideal ones don't depend upon the wild ones. Getting things wrong depends upon getting things right in a way that getting things right doesn't depend upon getting things wrong. Thus, the property HORSE causes "cow" because some horses, for example, those at the far end of the meadow, look like cows and, under ideal conditions, COW causes "cow."

So formulated, of course, the account still mentions ideal conditions, and these Fodor has conceded cannot be specified without circularity. His further interesting suggestion is that mention of the ideal conditions here is entirely inessential: *the structure of asymmetric causal dependency alone, abstracted from any specific conditions or causal chains, will do all the required work* (1990b: 99, 1998a: 156ff).

To simplify the discussion, we can define a predicate, "*x* is locked onto *y*," to capture this asymmetric causal structure:

A symbol "S" is locked onto property F just in case:

1 there's a (*ceteris paribus*) law that F causes tokenings of "S";
2 tokenings of "S" are robust: i.e. are sometimes caused by a property G other than F;
3 when Gs (other than Fs) cause tokenings of "S," then their doing so asymmetrically depends on (1) i.e. on the law that F causes "S"s,

where X's causing Ys "asymmetrically depends" on a law, L, if and only if X's causing Y wouldn't hold but for L's holding, but not vice versa: L could hold without X's causing Y. Thus, smoking's causing cancer, depending upon many laws, asymmetrically depends upon Newton's, since Newton's doesn't depend upon smoking's causing cancer. Fodor's proposal about content then is:

(M) if "S" is locked onto F, then "S" expresses F.

Thus, a predicate "C" expresses COW if (a) it were a law that the property COW causes "C" tokenings, and (b) other causal relations between properties (e.g. HORSE, MILK, etc.) and "C" tokenings asymmetrically depend upon this law.

Note that (M) supplies only a *sufficient*, not a necessary physicalistic condition for predicate expression. Fodor believes this is all that he is required to do, given his merely "supervenient" physicalism mentioned on p. 453. Fodor argues that, if there are no

counterexamples to (M), then he has done all that he needs to do to show that, contrary to dualism, certain physical arrangements are sufficient for intentionality.

Note also that Fodor avails himself of the convenient largesse (some might regard it as a profligacy) of properties in the world. For him, as for many philosophers, there's virtually a property for every primitive predicate, *whether or not the property happens to be instantiated*. Thus, there are properties of being a unicorn and being phlogiston, despite the lack of any instantiations of them in the actual world. And so the concepts UNICORN and PHLOGISTON are distinguishable in this way by the respective lockings.

Fodor (1987b, 1990b, 1991b) defends (M) with considerable ingenuity. Whether or not this atomistic account of meaning can succeed, it is crucial to understanding Fodor's later work where it is simply taken for granted. (Fodor has written almost nothing further on (M) since 1991.)

## Solipsism and narrow content

Despite the attractions of an "externalist" theory like (M), it is hard to resist the idea that there is *something* semantic purely "in the head." There are two standard ways of pressing this point: so-called "Frege" cases, and "Twin" cases.

### *Frege cases*

There seem to be plenty of expressions with the same worldly reference that nevertheless have patently different meanings. Frege's example was "the morning star" and "the evening star," but these are distinguishable in different possible worlds, and so involve different properties. But what about predicates that are necessarily co-extensive, not only in this but all possible worlds, like "equilateral" versus "equiangular" triangle? Here Fodor avails himself of the resources of the LOT: "equi-angu-lar" and "equi-lateral" are syntactically complex, and so thoughts involving them can be distinguished thereby (1998a: 15–21, 163–5). In his terminology, they are different concepts with the same content.

Can all cases be handled in these ways? Are all differences in thought either differences in the denoted properties or structural differences in the way the properties are represented? Proper names present one kind of problem; terms for kinds ("lawyer," "attorney") another; necessarily co-instantiated terms, such as Quine's notorious "rabbit/undetached-rabbit-parts" example, still another. Fodor claims that, so long as coreferential names and simple kind terms are treated as tokens of different types *internally* by any agent, *inter*agent comparisons are merely pragmatically different (1994: 109–12), a view that marks a change from his earlier view and which he shares with many "direct reference" theorists such as David Kaplan. He also provides a detailed response to the Quinian challenge, exploiting a distinction in the logical role the different co-instantiated thoughts play (1994: 55–79).

### *Twin cases*

Twin cases are the converse of Frege cases: instead of two expressions with the same reference but different senses, here we have expressions with the same sense but different references. Hilary Putnam invited us to imagine there was a faraway planet,

"Twin Earth," exactly like Earth in every way (including history) except for having in place of $H_2O$ a superficially similar, but atomically different chemical XYZ (see PUTNAM). Oscar on Earth thinks about water, i.e. $H_2O$, where his twin on Twin Earth doesn't think about water at all, but about twater, i.e. XYZ. The question now is whether psychology should care about distinguishing Oscar from his twin: after all, aren't their internal mental lives indistinguishable?

Fodor's views about twin cases have changed over the years. In a much discussed paper (1980a) he argued that psychology couldn't wait upon a full theory of all an agent's environment, telling us which was water and which twater. And so it had better adopt what Putnam called a policy of "methodological solipsism," theorizing only about what goes on inside an agent's head. He takes this to comport well with what he calls a "formality condition" that follows from CRTT: mental states have their efficacy as a consequence of the formal character of their tokens.

However, he also recognizes that intentional properties of mental representations are essential to their role in psychological explanation, and so, if psychology is solipsistic, there must be some solipsistic, or "narrow" kind of content that supervenes on the internal states of a thinker.

For a while, Fodor settled on the following conception (originally proposed by Stephen White, developing the seminal work of Kaplan): *the narrow content of a Mentalese expression is a function (in the set theoretic sense) that maps a person's life context onto a broad content*. For example, the narrow content of Oscar and Twin Oscar's "water" is the function that maps Oscar's context onto $H_2O$ and his twin's context onto XYZ. When Oscar utters "Water is wet," he thereby expresses the content "$H_2O$ is wet," while when Twin Oscar utters it he expresses the content "XYZ is wet." Two symbols have the same narrow content just in case they serve to compute the same such function: it is this that is shared by Oscar and his twin. Whether this function can actually be specified by psychology is, however, not altogether clear: how does one continue to specify it beyond Earth and the fanciful case of Twin Earth?

## *Only wide content after all*

More recently Fodor has moved away from any reliance on narrow content at all. He argues that both Frege cases and Twin cases don't have to be taken seriously by psychology: they are violations of the *ceteris paribus* conditions under which serious psychological laws are satisfied (1994: ch. 2). With respect to the Twin cases, he claims that it's important to remember what he earlier forgot, that, although they are *conceptually* possible, they are not *nomologically* so, and "empirical theories are responsible only to generalizations that hold in nomologically possible worlds" (1994: 29).

With respect to the Frege cases, his position is more complex. Citing the case of belief, he argues for what he calls the "Principle of Informational Equilibrium" (PIE),

> Agents are normally in *epistemic equilibrium* in respect of the facts on which they act. Having *all* the information – having all the information that God has – would not normally cause an agent to act otherwise than as he does. (1994: 42)

He claims that, since the success of our action is no accident and tends to depend upon the truth of our beliefs, "no belief/desire psychology can fail to accept PIE" (1994: 42).

Consequently, Frege cases, in which an action depends upon being *ignorant* of an identity statement are, from the point of view of psychology, "aberrations." They do occur; and, following the earlier discussion, they can be described by invoking syntactically different LOT expressions as the different "modes of presentation" Frege thought were needed. But, *pace* Frege and his many followers, no "third realm" of "senses," or narrow contents, is needed. Whether a similar argument can be made for attitudes other than belief is a question that is unfortunately not addressed.

## Nativism

In the same book in which Fodor presented the CRTT hypothesis he also defended a highly controversial thesis about *the innateness of all concepts*. That thesis originally stirred more controversy than did CRTT itself, much of which was addressed in publication (1981c); but it is independent of CRTT, depending upon further claims about the nature of concepts, definitions, and learning, issues which, moreover, are not likely to be settled by commonsense thought on the matter (1998a: 28).

In (1975a: ch. 2) Fodor argued that, since standard models of learning involved hypothesis confirmation – one learns that "cat" in English means *cat* by confirming hypotheses about English usage – these models are committed to the constituent concepts of these hypotheses being innate (1975: ch. 2): after all, one can't form hypotheses such as " 'red' means *red*," without using and therefore already possessing those very concepts!

Later he deepens the discussion by considering the traditional empiricist suggestion that one acquires concepts by constructing complex ideas out of sensory primitives that are variously associated in experience (1981b). The persistent failure over the centuries to set out these constructions, which is closely related to the failures of any IRS (see section "Inferential role theories"), suggests that they are probably not to be had. As Plato observed early on (and Chomsky more recently), most of our cognitive capacities seem to transcend our specific sensory histories.

However, Fodor (1998a: chs 6–7) has recently raised what he regards as a serious objection to his earlier views that all concepts are innate, what he calls the "DOORKNOB/'doorknob' " problem:

> Why is it so often experiences of doorknobs, and so rarely experience with whipped cream or giraffes, that leads one to lock onto *doorknobhood*? . . . assuming that primitive concepts are triggered, or that they're "caught," won't account for their content relation to their causes; apparently only induction will. But primitive concepts can't be induced; to suppose they are is circular. (1998a: 127–32)

That is, as he argued in 1975a: ch. 2, you can't perform inductions to concepts that you can't already represent.

His solution to this problem is "ontological": the properties that correspond to our primitive concepts are just the properties to which we generalize, *from phenomenally specified stereotypes*. For example, it is constitutive of *being a doorknob* that it is the property onto which people lock as a result of exposure to stereotypical doorknobs. Fodor relies here on what he regards as the psychological evidence that stereotypical instances

of primitive concepts can be specified *independently of those concepts*, e.g. by enumerating the shapes, colors, functions that typical instances share (1998a: 137–45).

Fodor thinks we now have an acceptable account of the non-arbitrary relation between the acquisition of many of our concepts and the experience of typical instances of them. Moreover, it has the interesting consequence that someone *not exposed to typical doorknobs* might well not have the concept "doorknob," since, without that experience, there might well be no such locking. He speculates that this is the case for most commonsense concepts. So, he concludes, "maybe there aren't any innate *ideas* after all" (1998a: 143). All there are are innate dispositions to lock onto properties when exposed to their stereotypical instances.

There are a number of problems raised by this view. It might appear to undermine the reality of the referents of our commonsense concepts, making them "dependent upon us." But this is an illusion. A doorknob may be identified by its tendency to have a certain effect on us; but it can exist even if, if we didn't exist, it didn't have that effect. But, in any case, minds are in fact a part of the world, and doorknobs do in fact have their effects upon them (1998a: 148–9).

A more serious problem is raised by what seem to be genuine "natural kind" concepts, whose reference patently does not involve any relation to what people might do: being genuine *water* is a matter of being $H_2O$, whether or not people would generalize to it on the basis of stereotypical samples. These concepts, Fodor claims, are latecomers in our cognitive development, dependent upon the social institution of science, the development of sophisticated theories and the consequent deference to experts (1998a: 150–62). His hope is that the peculiar semantic effects of these late developments can be entirely spelt out in terms of details of the counterfactuals involved in his asymmetric dependence theory (see "Teleological views").

## *Contra selectionism*

Fodor's interest in nativist hypotheses might lead one to think that he believes that our cognitive capacities are the result of natural selection. Nothing could be further from the truth. Although he doesn't doubt for a moment that the human mind/brain evolved, he sees no reason whatever to think that its considerable cognitive capacities were specifically *selected*. In part his opposition to selectionism is like his opposition to empiricist theories of learning: he sees no reason to think that these capacities reflect some sort of regularities in our histories. Many of them – such as our grammatical or mathematical abilities – seem to far exceed anything that either our upbringing or our evolutionary history could plausibly have supported.

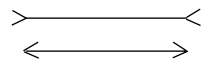
Additionally, he thinks that selectionist stories about the evolution of cognitive capacities, beside being flagrantly speculative, seriously underestimate the complexity of the relation between mind and brain: "make an ape's brain just a little bit bigger (or denser, or more folded . . .) and it's anyone's guess what happens to the creature's behavioral repertoire" (1998b: 209). It's as likely as not that some small change in our ancestors' brains made them *tremendously* smarter, like the modest change required to transform a finite state machine into a Turing machine. Doubtless this provided us with *some* selectional advantages. But that is no reason to suppose that anything like the majority of mental abilities we display – from acquisition of grammars to the grasp

of higher mathematics, physics, or folk psychology – were themselves individually selected.

## Modularity and the limits of CRTT

In discussing IRS theories in the section "Inferential role theories," we noted their tendency to become "holistic," a tendency Fodor sees as inimical to the interest of serious psychology. This tendency has also been prevalent in much contemporary "New Look" theories of perception: the work of Jerome Bruner in psychology, Thomas Kuhn in the history of science (see KUHN), and Nelson Goodman in philosophy (see GOODMAN), emphasizes how much people's theoretical expectations can color their perceptions, to the point that they and others insist that we ought to abandon "the myth of the given" (see also SELLARS). Fodor deplores this holistic tendency here as well, and in this case marshals interesting psychological evidence against it.

Fodor first of all calls attention to the surprisingly little noticed fact that the very perceptual illusions that New Look theorists often invoked to make their point actually tell against it: for *many of these illusions do not disappear when we know better*. No matter how sure we are that the Muller–Lyer illusion lines are equal, the upper still looks longer



than the lower. In a phrase Fodor takes from the psychologist, Zenon Pylyshyn, perception seems to be "cognitively impenetrable."

Fodor cites facts like these, and considerable data about language comprehension, to argue that there are a number of dedicated mental "modules" that are "informationally encapsulated" from the "central" system whereby we reason generally and fix our beliefs. These include the standard sensory systems, certain levels of language processing, and perhaps other dedicated systems such as face and musical perception. Among other things, such systems are, furthermore: *extremely rapid* (on the order of a quarter of a second, 1983: 61–4); *shallow* (their outputs are limited to "basic perceptual categories," such as chair or dog, 1983: 86–99); associated with a *characteristic development* (vision and language seem to develop in specific ways that are independent of other mental capacities, 1983: 100–1); and they are *domain specific* (confined to, for example, processing of light, faces, or grammar, 1983: 47–52). In this last respect, Fodor's conjecture overlaps, of course, with Chomsky's postulation of an innate grammatical competence, but, further, involves the modularity of not only the domain, but of the standard processing of information from that domain (1989, 1998b: ch. 4).

All of these properties contrast with properties of the central system, which is *voluntary* (you can choose what to think about), *slow* (you can potter for months), sometimes *deep* (you can think about non-perceptual categories), non-localized ("there is, to put it crudely, no known brain center for *modus ponens*," 1983: 98), and *domain independent* (you can think about almost anything).

Philosophically, what is perhaps most intriguing about the postulation of sensory modules is the way in which it provides a new basis for the controversial observation/theoretic distinction, although a basis that may only partially overlap the traditional introspective one (1984b, 1988b).

It might be thought that, given their encapsulation, these modules are not really *cognitive*, and so don't involve all the issues about computation and representation that are the focus of most of Fodor's work. But precisely the point of postulating modules as opposed to standard transducers is that they do seem to involve computation. Indeed, Fodor regards them as "compiled transducers. . . . 'compiled' to indicate that they have an internal computational structure, and 'transducer' . . . to indicate . . . information encapsulation" (1983: 41).

Indeed, in his recent work (2000), Fodor argues that modules may be the only appropriate domain for a CRTT. In being *un*encapsulated, the *central* belief fixation system exhibits a number of properties that present serious prima-facie difficulties for any standard computational treatment. Relying on what he regards as Quine's astute views about theory confirmation, Fodor claims that central systems are:

1 "*Quinian*," i.e. computed over the totality of a belief set, as when we settle on a theory that is, for example, simplest and most conservative overall;
2 *isotropic* (every belief is potentially relevant to the confirmation of every other, as when a pattern of light on a piece of paper confirms a theory about the age of the universe) (1983: 105ff).

Fodor (1999) argues that these features render belief fixation holistic and context-dependent, in a fashion that is not clearly amenable to the Turing computability invoked by CRTT. This latter depends upon exploiting a representation's local syntactic features: an argument's deductive validity can be checked by looking at its local spelling. However, abductive cogency seems to be ascertainable only by looking at a claim's relation to other, indefinitely remote representations, and its effect on the belief system as a whole. Fodor sees this as the problem underlying the so-called "frame problem" encountered in artificial intelligence (1987a), and is consequently pessimistic about the prospects of it being ultimately solved by CRTT. Although CRTT is necessary for an adequate theory of mind, it seems to be far from sufficient.

## Bibliography of works by Fodor

For a full bibliography of Fodor's work up until 1991, see B. Loewer and G. Rey (1991) *Meaning in Mind: Fodor and his Critics* (Oxford: Blackwell Publishers), which also contains critical essays by a number of prominent philosophers, and an introduction on portions of which the present entry relied.

1963 (with Katz, Jerrold): "The Structure of a Semantic Theory," *Language* 39, pp. 170–210.
1968a: *Psychological Explanation*, New York: Random House.
1968b: "The Appeal to Tacit Knowledge in Psychological Explanation," *Journal of Philosophy* 65, pp. 627–40.
1970: "Three Reasons for not Deriving 'Kill' from 'Cause to Die'," *Linguistic Inquiry* 1, pp. 429–38.
1972 (with Block, N.): "What Psychological States are Not," *Philosophical Review* 81, pp. 159–81.
1974 (with Bever, T. and Garrett, M.): *The Psychology of Language*, New York: McGraw Hill.
1975a: *The Language of Thought*, New York: Thomas Y. Crowell.
1975b (with Fodor, J. D. and Garrett, M.): "The Psychological Unreality of Semantic Representations," *Linguistic Inquiry* 6, pp. 515–31.

1978: "Tom Swift and his Procedural Grandmother," *Cognition* 6, pp. 229–47.
1980a: "Methodological Solipsism Considered as a Research Strategy in Cognitive Science," *Behavioral and Brain Sciences* 3, pp. 63–109 (with replies to commentators).
1980b (with Garrett, M., Walker, E. and Parkes, C.): "Against Definitions," *Cognition* 8, pp. 263–367.
1981a (with Pylyshyn, Z.): "How Direct is Visual Perception? Some Reflections on Gibson's 'Ecological Approach'," *Cognition* 9, pp. 139–96.
1981b: "The Present Status of the Innateness Controversy," in Fodor 1981c, pp. 257–316.
1981c: *Representations: Essays on the Foundations of Cognitive Science*, Cambridge, MA: MIT Press.
1983: *The Modularity of Mind: An Essay on Faculty Psychology*, Cambridge, MA: MIT Press.
1984a: "Semantics, Wisconsin Style," *Synthese* 59, pp. 231–50.
1984b: "Observation Reconsidered," *Philosophy of Science* 51, pp. 23–43.
1986: "Why Paramecia Don't Have Mental Representations," in *Midwest Studies in Philosophy*, vol. X, ed. P. French, T. Uehling, Jr., and H. Wettstein, University of Minnesota Press.
1987a: "Frames, Fridgeons, Sleeping Dogs and the Music of the Spheres," in Z. Pylyshyn (ed.) *The Robot's Dilemma; The Frame Problem in Artificial Intelligence*, Norwood, NJ: Ablex.
1987b: *Psychosemantics: The Problem of Meaning in Philosophy of Mind*, Cambridge, MA: MIT Press.
1988a (with Pylyshyn, Z.): "Connectionism and Cognitive Architecture," *Cognition* 28/1–2, pp. 3–71.
1988b: "A Reply to Churchland's 'Perceptual Plasticity and Theoretical Neutrality'," *Philosophy of Science* 55, pp. 188–98.
1989: "Why Should the Mind be Modular?," in A. George (ed.) *Reflections on Chomsky*, Oxford: Blackwell Publishers.
1990a: "Psychosemantics, or Where do Truth Conditions Come From," in W. Lycan (ed.) *Mind and Cognition*, Oxford: Blackwell Publishers.
1990b: *A Theory of Content and other Essays*, Cambridge, MA: MIT Press.
1991a (with McLaughlin, B.): "Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work," *Cognition* 35/2, pp. 185–204.
1991b: "Replies," in *Meaning in Mind: Fodor and his Critics*, ed. B. Loewer and G. Rey, Oxford: Blackwell Publishers, pp. 255–319.
1991c: "You Can Fool All of the People Some of the Time, Everything Else Being Equal: Hedged Laws in Psychological Explanations," *Mind* 100/1: pp. 19–34.
1992a (with LePore, E.): *Holism: A Shopper's Guide*, Oxford: Blackwell Publishers.
1992b: "Substitution Arguments and the Individuation of Beliefs," in G. Boolos (ed.) *Essays for Hilary Putnam*, Oxford: Blackwell Publishers.
1994: *The Elm and the Expert*, Cambridge, MA: MIT Press.
1998a: *Concepts: Where Cognitive Psychology Went Wrong*, Oxford: Oxford University Press.
1998b: *In Critical Condition*, Cambridge, MA: MIT Press.
2000: *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*, Cambridge, MA: MIT Press.