Has Mentalese Earned Its Keep?
On Jerry Fodor’s LOT 2

Jesse Prinz
City University of New York
jesse@subcortex.com


On my first day in college, I confessed to my academic advisor, Raziel Abelson, an interest in Noam Chomsky’s theory that syntax is innate. ‘You think that’s crazy?’ he chortled, ‘well, there’s a philosopher at City who thinks all concepts are innate’. That afternoon, if memory serves, I bought a copy of The Language of Thought, and thus began my love/hate relationship with the writings of Jerry Fodor. The Language of Thought was a groundbreaking book, which marked the end of the Wittgensteinian stranglehold on the philosophy of mind, and played a funeral dirge for behaviourism. It also departed in exciting ways from prevailing forms of mentalist materialism, by shifting focus from sensations to thoughts; the identity theorists had said behaviourists were right about thinking, and the functionalists had implicitly followed suit focusing on states such as hunger and pain. The idea that thoughts could be built up of symbols was an inevitable consequence of the computer revolution, but Fodor gave it a philosophical defence that reset the agenda for philosophy of mind in profound ways. LOT 2 appears about 33 1/2 years after the original hit the shelves, and in those decades, Fodor has been rebuffing dissenters so frequently that the grooves in his long-playing arguments have began to show the signs of wear. LOT 2 is a progress report, and Fodor seems at times weary of stressing the same old points against critics who did not get it the first time. There is little new here, but the book is a welcome encapsulation of the various ways in which Fodor’s thinking has changed over the years and the ways in which the original story remains. Fodor still thinks the language of thought hypothesis is the only game in town, and, with his characteristic panache, he slings arrows at enemies old and new. His new formulations are often more
concise than in previous works (for better and worse), and he adds a few wrinkles to his time-tested tale. In this critical commentary, I will discuss both the content of the book and comment on what he leaves out.

LOT 2 is one-stop-shopping for those who want Fodor’s take on philosophy, and that makes it more cohesive and more encompassing than his other recent books. For the uninitiated, it is a useful introduction (though Fodor often presupposes an audience or readers who have tuned in to earlier episodes). For dedicated readers, there is also much of interest, including occasional revisions, reformulations, and augmentations of earlier arguments. In re-asserting his case for LOT, Fodor says surprisingly little about the many commentators and critics who have discussed his earlier work. The message is clear enough: these opponents have not come up with any objections worth considering in detail, and a mere re-statement of the core arguments will be enough to show why. That is an unfortunate posture to adopt, because more attention to critics might have pushed the book into another epicycle instead of retracing the same old orbit.

I will illustrate by walking through each chapter. Along the way, I will try to show that Fodor has not given us adequate reason to reject theories that compete with his version of the language of thought hypothesis. To illustrate, I will focus on prototype theory, since it has been his most frequent foil, but my own view is that many concepts comprise the full range of representations used to categorize, including prototypes, exemplars, folk theories, and so on.

In his introductory chapter, Fodor identifies pragmatists as his primary opponents, and he defines pragmatism as the view that ‘the distinctive function of the mind is guiding action’ (p. 13). This cannot be right, Fodor complains, because thinking about what exists and what would happen under counterfactual conditions must be prior to planning an action. This might sound like a promising argument, but its success depends on the specific assumptions of the imagined opponent. For example, the argument has no impact against the pragmatist who says that the mind has the function of guiding action in a teleological sense, for it is consistent with that view that minds must represent states of affairs in order to achieve the end for which they are selected. Fodor’s argument is also consistent with the view that conceptual knowledge includes knowledge of how the items in the world around us might be used or transformed. For example, my representation of a coffee cup might include a stored record of cup-directed behaviours. Indeed, without further elaboration, Fodor’s
argument does not even work against current defenders of enactive perception who say that we perceive features of objects by registering sensorimotor contingencies. If such representations can be primitive (they may be in simple organisms), then planning could involve selecting between the many possibilities of action that are delivered directly, as it were, to the senses. I do not endorse such a view myself. My point is simply that Fodor has not delivered a master-argument against prevailing theories that might count as pragmatist by his definition. More alarmingly, he does not keep to that definition. Fodor often uses ‘pragmatism’ as an umbrella term for a wide range of views, including any that violate the strictures of his conceptual atomism, such as functional role semantics. If his anti-pragmatist argument is supposed to work against those views, it is not clear how. Nor is it clear how it undermines the agendas of Sellars, Brandom, Quine, Wittgenstein, Churchland, and other authors singled out as pragmatists in this chapter. Nothing much hangs on this, but I do think it is unfortunate that Fodor tries to find a single label for his opponents, since it is hard to know when label applies (Am I a pragmatist? Are you?), and dubious that its diverse adherents can be felled by the same general objection.

Chapter two continues the crusade against ‘pragmatism’ but moves into terrain that stands a better chance of thwarting those pragmatists who are not wedded to the link between concepts and behaviour, namely those who think many lexical concepts are not atomic. That includes just about everyone other than Fodor and a few of his students. The majority view is that lexical concepts are structured entities, containing a collection of concepts (or representational components) as parts. Some researchers think of these parts mereologically, while for others, concepts are individuated by inferential roles, where tokening a concept disposes or entitles a person to infer that other concepts apply. Rather than ‘pragmatism’ this might be called ‘molecularism’ about concepts. Fodor’s invective takes the form of a dilemma, which is familiar to his readers. Concepts cannot be constituted by defining features, since there are no good definitions, and concepts cannot be constituted by non-defining features, such as prototypes, because prototypes are not compositional, and concepts must be compositional. Therefore, concepts are not constituted by features at all (unless they are explicitly phrasal), and molecularism is wrong. Few critics are willing to defend definitions these days, so much hangs on Fodor’s compositionality argument, which is one reason he has repeated it so often over the years. Fodor admits that
prototypes exist (indeed, as we will see, they do important work in his theory), so denying that concepts are prototypes rests heavily on the assumption that they do not compose. Is he right about that?

Many critics have taken up this argument, and I will mention here only one reply (see Prinz 2002). The compositionality argument says that concepts must compose because otherwise we would not be able to generate an unbounded number of thoughts from finite resources. (There is also an argument involving ‘systematicity’, but the remarks that follow can be easily adapted.) It can be shown empirically that prototypes often fail to compose: the prototypical features of ‘carnivorous plants’ include ‘eats insects’, which is neither prototypical for plants nor carnivores. But it does not follow that prototypes cannot figure into our explanation of how people generate novel thoughts from finite means. The fact that prototypes do not always compose is readily explained. Since we have seen carnivorous plants, we can acquire a prototype directly from those encounters without relying on compositional mechanisms. But suppose we encounter a truly unfamiliar concept, such as ‘carnivorous chair’—the subject of a new horror film. To mentally represent this English phrase, we might well compositionally combine features of the corresponding prototypes. Carnivorous chairs have mouths, stalk their victims, have four legs, back support, and can be sat on. In a similar manner, we can generate a prototype for any arbitrary compound. When we do this, we may modify the resulting prototypes using the power of reasoning. For example, we might conceive of a ‘tinfoil chair’ as incapable of being sat on, because we reason that tinfoil easily bends under human weight. But such modifications are wholly consistent with the crucial point: prototypes can be combined compositionally, and that is sufficient for explaining how they could be used to represent arbitrary combinations. It does not follow that compounds generated in this way will be very good for identifying category instances. For all I know, real carnivorous chairs eat through their pores rather than through their teeth. But compositionally generated prototype compounds are still very useful. If you send me to the store to buy a purple sofa, searching for an object shaped like a typical sofa and shaded in typical purple will surely help. In cases where I have reason to think the compound will not be like its components, I can simply amend my prototype (as with the tinfoil chair). Furthermore, compound prototypes will surely be better for identifying category instances than a pair of structureless atoms. Prototypes provide an imperfect method for predicting the features of novel
compounds, but Fodor offers no method at all, and there is no doubt that predicting features is something we readily do. One might think that explaining our productive capacity to anticipate the features of novel categories is something that cries out for explanation, but Fodor leaves this capacity unexplained.

Fodor does not address this reply to his compositionality argument, and he condemns his one published response to a footnote. In that paper (Connelly et al. 2007), Fodor and his collaborators provide experimental evidence against the claim that prototypes compose for arbitrary unfamiliar cases. In so doing, he implicitly admits that the debate hangs on largely untested empirical issues, rather than his in-principle argument. The in-principle argument can establish that we must be capable of combining concepts compositionally; but that demand is easy to meet. The empirical issue concerns the conditions under which we actually do compose, and here Fodor thinks prototype theorists get it wrong. But his study suffers from a doubly doomed design: (1) it tries to establish a conclusion about prototypes using a method of measuring prototype structure that departs inexplicably from the vast empirical literature on the subject, and (2) it fails to take into account existing models of prototype combination that build in systematic adjustments to feature weights under combination; for example, on Hampton’s (1991) model, prototype features pool together compositionally, but weights are systematically lowered for any feature that is not present in both prototypes being combined (see Prinz forthcoming). Perhaps that is why the study does not make it into the main body of the book. But, without this study, we find Fodor repeating an old argument without replying to those who have, by my lights, exposed a fatal flaw. That worn-out argument packs more bark than bite.

The debate about molecularism is of crucial importance because the most notorious problems with Fodor’s view could be avoided if he abandoned his atomism. Consider, for example, chapter three, in which he takes on Frege cases. Fodor is a referentialist: he says a concept’s content is exhausted by what it refers to. This makes it hard to explain cases of informative identities. If content is exhausted by reference, then anyone who knows that Jack the Ripper killed Mary Nichols should know that Walter Sickert killed Mary Nichols (assuming that Sickert was the Ripper). The molecularist has an explanation of why this identity is informative: we think of ‘The Ripper’ by means of a vicious murderer prototype and we think of ‘Sickert’ by means of an eccentric painter prototype. Fodor admits that
we have prototypes, but he denies that these are concepts, so he
wants to distinguish RIPPER and SICKERT without appeal to the
meaningful features with which they are associated.

Fodor’s solution? Co-referring concepts can have distinct realiza-
tions, where realizations are individuated below the level of semantic
roles. But what are these realizations? They cannot be neural, because
neural implementations are likely to be highly variable both across and
within individuals; consider the fact that two connectionist nets will
produce distinct weightings when trained on the same learning cor-
pora, and the weightings change in a single network with each episode
of learning. Fodor seems to suppose that we can type neurons into
equivalence classes that are neither physical nor semantic, but then we
need an actual story of how these classes get typed. Avoiding semantic
features also seems unhelpful. It is a bit like explaining substitution
failures by appeal to differences in fonts. From a psychological point
of view, the difference between SICKERT and RIPPER owes to diver-
gent features attributed to their referents, including the simple belief
that they have different names. As has been pointed out in the
literature, non-semantic theories of Mentalese typing seem especially
hopeless when it comes to inter-personal Frege cases (Aydede 1998,
Schneider 2005). It is grossly implausible that everyone who has
thought about Jack the Ripper has neurons that can be type-identified
by anything other than their psychological or semantic similarities.
What these people share is a set of beliefs about heinous crimes
committed by someone called ‘Jack the Ripper’. Leaving that out of
the story fails to explain why people with shared concepts act in
similar ways; for example, many people read books about Jack the
Ripper, but not about Sickert as such.

Fodor does not cite the literature on this problem, but he seems to
be on to it when he worries that the neural account cannot explain
how we share thoughts with Martians, who evidently lack our neuro-
physiology (p. 90). Fodor’s reply to the worry is puzzling. He says that
Mentalese types may be multiply realized at the level of physical
implementation. But that is not a solution to the problem; it is a
restatement of it, and it vitiates the proposal that typing can be
done by appeal to non-semantic properties. We need a way to type
Mentalese tokens that is not neurophysiological, and Fodor has not
offered one. Without that, his theory will fail. Fortunately, there is
an alternative. We can type Mentalese tokens by accompanying
prototypes, or better yet, drop Mentalese and say that concepts are
prototypes. In the appendix to chapter three, Fodor gives what can be
taken as another reason against this proposal (in case you are unmoved by the compositionality objection). He says that tokening concepts does not generally cause one to think of features associated with the concept, because that would be ‘derailing’ (p. 96). This claim is empirically false. Granted, it would be derailing to think of every associated feature, but it would not be derailing to think of prototypical features. Forty years of research on prototypes, semantic priming, and, more recently, neural activations confirm that this is exactly what occurs. For example, La Heij et al. (1990) show that the word ‘dog’ brings cats to mind, and Matsukawa et al. (2005) show that pictures of ants make it easier for people to identify pictures of butterflies. There are interesting empirical questions about how such associations arise without disrupting thought, but the claim that such associations do not arise illustrates the risk of doing armchair psychology.

Turn now to chapter four. Here Fodor rehearses his reasons for thinking that computation (as we understand it) cannot explain mentation. This theme in Fodor’s corpus is perhaps his most despairing (see e.g. Fodor 2002). The computational theory of mind is arguably the best hope we have for developing a science of the mental. If it cannot work, cognitive science is doomed. Fortunately, the argument is fishy from the get-go. After all, cognitive science is alive and well, and dozens of journals report new scientific findings about the workings of the mind each month. Human thought is surprisingly predictable. That does not mean anyone can predict my next thought, but, if you put a carrot in front of me and say ‘name it’, you can predict I will say ‘carrot’ and not ‘vegetable’ or ‘orange cone’ and for reasons that are fairly easy to characterize computationally. Likewise, we know variables that can reliably elicit emotions, promote helping behaviour, overwhelm attention resources, produce false memories, lead to parsing errors, engender illusions of control, and push evaluative judgements in either direction. Fodor doubts that these achievements will lead to an adequate computational theory of thinking, because of an insurmountable obstacle: the dreaded frame problem. Computation works locally, he reminds us, and there is no local solution to the way we search our vast knowledge stores to identify what bits of information are relevant to updating beliefs from moment to moment. But this assertion is just an observation about current science masquerading as a principled argument. Many researchers who grasp the challenge are optimistic about finding an empirically informed solution (e.g. Shanahan and Baars (2005)).
The mere fact that anything we know might be relevant at any given moment does not entail that we need non-local processes to search our knowledge stores. We need a search procedure and we need a way to rank the items that we find, and neither of these requirements looks insuperable. To make this point, one might follow Andy Clark (2002) who finds reason for optimism in work on Internet search engines. To search the Web is a monumental task, but good search engines use simple, purely syntactic, local, computational processes to deliver ranked lists of hits very quickly. The good ones give hits in rough order of relevance. Heavily travelled paths with many connections have an advantage, much as they would in the brain. Clark’s point is not that Google has solved the frame problem, but that the success of search engines should help us see that global, weighted searchers are not computationally intractable. Perhaps Fodor should cheer up and learn to love computational approaches to central cognition again.

Fodor swaps scepticism for credence in chapter five, where he takes up the most provocative thesis in his earlier book: radical concept nativism. Fodor updated his earlier discussions in Concepts: Where Cognitive Science Went Wrong (Fodor 1998), and here he offers an economical restatement of the view advanced in that book, but he also adds some things, making it one of the most interesting chapters of the book. In Concepts, Fodor had noticed that his argument against concept learning does not count as an argument for innateness because unlearned does not entail innate. He repeats that refrain here, but now recognizes that this is not really a departure from his original view, whether you call it nativism or not. The main point is that, contrary to the view of all sane and decent cognitive scientists, most lexical concepts are unlearned. Here he strengthens the argument to include phrasal concepts as well. Concept learning is impossible, according to Fodor. To learn the concept C we would need to formulate a hypothesis about what Cs are, but to do that, we must already have a concept of Cs, and this is true no matter how logically complex C happens to be.

I was never much bothered by Fodor’s nativism, despite the mocking dismay of my undergraduate advisor. After all, as an atomist, Fodor does not think we have innate knowledge of any categories. Rather, he thinks the process by which concepts are attained is causal rather than rational. Empiricists can happily agree. I acquire a concept of aardvarks when exemplars leave an impression on my sensorium. Fodor’s new strengthened argument may look more momentous, but it is not. Fodor is not saying we have mastery of
descriptive concepts prior to experience; he merely establishes that bringing together the features that align in such concepts might be described as a brute causal process. For example, if I see a blue sphere, it might cause me to form a concept of blue spheres. I did not have this concept before, but its acquisition was brute causal, hence that acquisition does not count as learning. If that is what radical nativism amounts to, sign me up, along with Locke and Hume. Curiously, Fodor does admit that prototype acquisition is a learning process, perhaps because the abstraction of a central tendency is an abductive inference that involves seeing exemplars as evidence for a structure that need never be experienced. I am inclined to call all of these cases learning, because (a) we end up with a concept that we did not have use of before, and (b) we did so on the basis of experiencing entities that fall under those concepts. But I am not one to quibble about terminology. The main point is that Fodor’s acquisition story does not deserve the ire and incredulity that it has elicited, least of all from an empiricist.

But Fodor’s positive view does raise one issue of concern. Recall that, for him, concepts are not prototypes or stored impressions; they are words in Mentalese, which, like English words, are arbitrarily related to their referents. So he, unlike the empiricist, owes us an explanation of why we seem to attain concepts by means of encounters with their instances. Here things get a little weird. In *Concepts* (1998), Fodor offered a bizarre metaphysical solution. Concepts, he said, refer (at least initially) to mind-dependent properties, and, in particular, C refers to the property of being the kind of thing typical instances of which cause attainment of C. This idealist theory of reference does not ring true even for concepts that are widely held to refer to response-dependent properties. FUNNY does not refer to things that lead us to acquire FUNNY; it refers to things that make us laugh (these are modally distinct properties). Moreover, Fodor’s solution will not work. Suppose we type a Mentalese token by its syntax (something Fodor still needs to explain), and ask why tokens of that syntactic type get acquired by means of encounters with a certain category. If syntactic types are arbitrary and Mentalese symbols are syntactically individuated, then it is sheer chance that a given object in the world causes us to acquire a particular concept. Thus, Fodor does not have an explanation for why any given concept is acquired by means of encounters with its instances.

In *LOT 2*, Fodor re-asserts the mind-dependence proposal, despite objections in the literature (e.g. in Prinz 2004), and he also adds
something even more startling. The addition comes in response to the question: Why does a given prototype lead to one concept, rather than another? Fodor’s answer is that the distance between Mentalese labels (which are concepts, on his theory) and prototypes is innately specified. As I understand the proposal, when a new prototype is acquired, it is not given an arbitrary Mentalese label, but rather is innately predisposed to get one particular label, rather than another (Fodor uses the language of attractor landscapes here, but that anachronistic concession to the forces of evil need not distract from the basic picture). This really is a radically nativist proposal, because it assumes that arbitrary labels are predestined to have certain features associated with them.

This ‘picture’, as Fodor calls it, is needlessly radical. On the face of it, there are two simpler answers to the question of why a given prototype leads to one concept rather than another. One possibility is that concepts are prototypes. Fodor rejects this because he thinks prototypes do not compose. Another possibility is that prototypes are the mechanisms that get concepts to detect reliably the properties to which they refer, as Margolis (1998) might have it. The story might go something like this. The mind picks up prototypes by statistical inference over exemplars, and then, once a prototype is formed, it appends an arbitrary label to the prototype. The concept then refers to the property that reliably causes the prototype to be tokened. This at least might be the initial stage. Over time, the semantics might change by grounding the prototype in theories. For example, initially a dog prototype might detect the property of having typical doggy appearances, and the label on that prototype will refer to that property. But, over development, the concept possessor might start to treat the concept as referring to a natural kind by adopting a functional role that allows tokenings in the absence of doggy appearances, provided there is reason to believe that the animal in question has the same kind of inner properties as a typical dog. This would comport with the developmental evidence presented by Keil in his discussion of the characteristic-to-defining shift (Keil 1989).

Fodor resists this simple answer, though it is not clear why. One possibility has to do with the issue raised earlier. He may assume that he needs to explain why each specific Mentalese label gets assigned to one prototype rather than another, as if specific labels were predestined to have particular semantic roles. The arbitrary labelling story that I just told would not work this way. But that is a gratuitous
assumption. All that matters is that the label, whatever it happens to be, ends up locking onto a property unifying the exemplars on the basis of which it is acquired. Fodor’s second reason for resisting the labelling story may have to do with his contention that prototypes can play very different roles with respect to concepts. He says, for example, that the cabbage prototype bears a different relationship to the cabbage concept than the king prototype bears to the king concept. I do not think this is a serious worry. First, there may not be much variation: in almost all cases prototypes can serve as detectors for a range of exemplars that are unified by some shared property. Second, the way prototypes determine reference can be influenced by the adjoining theories. In the case of natural kinds, the reference rule might be: refer to those things that have the biological or chemical microstructure shared by the things that are reliably detected by the prototype. In the case of the king prototype, the rule might be: refer to the social role uniting the individuals who are reliably detected by the prototype. In each case, the prototype detects a class of individuals, and then semantic rules, which can each apply to multiple concepts, are applied to determine how to generalize those individuals to a reference class.

In chapter six, Fodor takes on the vexed topic of non-conceptual content. In his work on modularity, Fodor argued for functional differences between perception and thought (see Fodor 1983), but, for all he says there, modules use concepts. Here Fodor presents an argument for the conclusion that some content is non-conceptual. He begins with a novel way of distinguishing conceptual structures from non-conceptual structures, which does not depend on Gareth Evans’s dubious suggestion that non-conceptual content is not compositional (pace Evans, percepts must be made of re-combinable parts in order to facilitate recognition of arbitrarily novel scenes). The difference has to do with the nature of the parts, according to Fodor. For non-conceptual representations, every collection of parts counts as a component, but not so with thoughts (the thought that John loves Mary does not contain a component of the form ‘John … Mary’). Fodor uses this distinction to derive a whole set of ways in which non-conceptual representations, which he calls ‘iconic’, differ from their conceptual counterparts (they lack truth values, they do not become more complex as items are added, they cannot contain quantification, etc.). Finally, he argues that empirical evidence, including Sperling’s famous array studies on iconic memory prove that such representations exist.
Still, there are bones one might pick in this discussion. Fodor gives the impression that early vision is something like a pixel array, and he might be wrong about this. When we move past the retina, vision may parcel the percept into something like proto-objects (identified by bounded contours) or at least into regions of space with features placed in those regions (Clark 2004). Such visual representations violate some of Fodor’s strictures; for example, a visual representation with many proto-objects or occupied regions is more functionally complex than one with few, and the relationship between features and regions might be best construed as a kind of predication. Also, given that vision has the function of representing items in the world, it might be said to produce representations that have truth-values. Perhaps an after-image says something false. Sperling’s studies show that we can see without appending certain categorical labels, not that vision lacks structure. In the end, I think Fodor cannot be right about how to draw the distinction, because he relies too heavily on a snapshot theory of visual representation.

Fodor’s final chapter turns, at last, to the question of how mental states represent. Playing on a Nagel quote, he says that this question is what makes the mind–body problem so hard (consciousness, he continues, makes it impossible). Readers will be relieved to discover that the chapter does not offer further technical refinements to Fodor’s asymmetric dependency theory. Fodor still accepts that theory, but does not get bogged down in details here. Instead he informally addresses some nagging worries about the programme. He writes dismissively about those who think naturalized semantics cannot handle normativity. His most substantive point against norm-mongers is that they over-emphasize the social in their semantic theories. I think this is a misdiagnosis. After all, there are social norms, and these too can be naturalized. Moreover, it is plausible that natural kind concepts — Fodor’s favourite examples — depend on deference to scientific practices; does TIGER refer to a genotype or a clade? Ask biologists (cf. Fodor 1998, Ch. 7). I think Fodor is right that semantics bottoms out in individuals; social patterns of sanctioning are easier to explain by appeal to the underlying individual psychologies than conversely. But this does not mean we can ignore the social world in explaining how ordinary concepts refer.

Another problem that Fodor addresses here concerns the causal chain leading up to the tokening of a concept. On a causal theory of reference, a concept must represent something in that chain, but which thing? It could be anything from the big bang to activity in the
cortex. Fodor rules out the big bang and cortical activity by saying that these things are not visible. That is a cheat, I think, since visibility looks like a semantic term and it is debatable what is visible. Is tiger- hood a visible property, for instance, just because I can see tigers? In any case, there are many numerous intuitively visible properties in any causal chain. When I see a tiger, I also see its fur, the air between us, and so on. Fodor offers an account of when such an episode counts as TIGER-perception, and it goes like this: if I see a tiger, the concept I token refers to the property of being a tiger if that property is the common denominator in the causal chains of counterfactual versions of myself who token instances of the same concept on the same occasion from different vantage points. Graphically, Fodor thinks of this as triangulation (a term he co-opts, with intentional irony, from Davidson). Me and my modal counterpart both stare at the tiger from different positions, and the lines from our eyes converge on the tiger and nothing else. The problem with the solution is it does not seem to rule out the candidates I just mentioned. My counterpart and I both see the fur and the air as well.

Readers might be puzzled about Fodor’s triangulation proposal for another reason. In other works (e.g. Fodor 2000) he offers another solution: asymmetric dependence. Recalling that account, we might have expected him to say that a TIGER token refers to tigers, not fur, because the counterfactual linking tiger experiences to tiger fur depends on the counterfactual linking tigers to TIGER-tokenings, but not conversely. It is not entirely clear why Fodor does not make this move here, but it is probably a good thing that he resists. After all, the claim that TIGER tokens would be caused by tigers if they were not caused by tiger fur is a bit hard to swallow. Fodor’s semantic theory is synchronic, that is, to hold by it, we must keep psychology fixed, and not invoke scenarios in which the concept-learner grew up under different conditions. So, when we pick the closest world in which tiger fur does not cause TIGER tokens, it must be a world in which our minds have not changed. I, for one, can only identify tigers by their fur. So when the fur-TIGER link breaks, I am hopeless. For example, in a world were tigers had polka-dotted fur, they would have no tendency to cause TIGER tokens in me. Conversely, if I went to a world in which leopards had tiger fur, the tiger fur-TIGER link would remain, but the tiger-TIGER link would be lost. Fodor’s old theory stipulates counterfactuals that we have no reason to believe are true.

I think the best strategy for dealing with these cases will involve something about how we use the concept, because no externally
specified condition will do. For example, we use the TIGER concept as an animal concept, not as an appearance concept, as reflected in the inferences we are inclined to draw. Fodor may think this is a concession to pragmatism (gasp!), but I do not see why an appeal to functional roles is any more horrifying than an appeal to counterfactual selves. Moreover, Fodor (1994) appeals to functional roles elsewhere, for instance in his treatment of Quinean indeterminacies (see Fodor 1994) — he endeavours to explain how a causal semantics can differentiate a tiger from tiger parts, and other properties with which they are necessarily co-instantiated. I do not think Fodor’s treatment of Quinean indeterminacy succeeds (see Gates 1996), but I do think such indeterminacies fail to bother us precisely because, when they are brought to our attention, we can stipulate which order of nature we are referring to, and that stipulation alters the inferential roles of our concepts, even if the laws linking mind and world remain unchanged.

More generally, it might be added that there is something deeply peculiar about Fodor’s ‘naturalism’, when it comes to semantics. First, he is happy to appeal to counterfactuals and modal selves, which are on shaky scientific footing, but he does not want to appeal to human inferences, which clearly exist. Second, like many semantic naturalists, he never proposes a single experiment that could test between his and competing theories, suggesting that these are not empirical theories at all. (Notice that Fodor’s various theories of content over the years make no different predictions about behaviour and posit no different inner states.) Third, he insists that we explain intentionality without reference to psychological states, but he is a materialist about psychology, so this is a metaphysically unmotivated demand. The worry might be that we need to pin down our psychological states somewhere, but why think we need to do it in one step? Perhaps we can explain concepts by appeal to inferences that are grounded, ultimately, in perception, with perception then grounded in sensory features, which can be explained, at last, by appeal to mind–world relations. Fodor probably resists this for fear that it will not secure publicity; inferential roles between people differ. But we have seen that he himself has no adequate theory of how concepts are shared. Finally, it is not clear that we need a psychosemantic theory in Fodor’s sense: a theory that assigns a unique reference to our concepts without indeterminacy. Once we establish what inferences people draw when they token a concept, and what real-world conditions cause concepts to be tokened, there is nothing left to explain, even if these two factors do
not establish a unique referent for every concept. The only reason we might want a theory that determines a unique referent is that some people believe their concepts refer to natural kinds, but this itself is just another piece of behaviour; and like any behaviour, it can be explained by appeal to psychological states.

The foregoing has been a bit of a snarky review, in honour of the Snark character who appears impetuously throughout Fodor’s book, and as a tribute to Fodor, whose bellicose wit has provoked more than one generation of philosophers. Reading Fodor is valuable in part because he can be exasperating, and he is exasperating because he defends provocative views using arguments that deserve and demand attention. This book is no exception, and that makes it a rewarding experience for both long-time readers and new initiates. In the spirit of LOT 1 (as he calls the first book), let me end with one final barb that summarizes the main theme of this commentary. Fodor has little patience for a priori philosophy, and he seeks a theory that will be consistent with empirical facts and guide empirical research. For all that, we see scant few research findings, nor proposals about how to test the theories. The arguments are largely of the in-principle sort. This is unsurprising, in a way, because of Fodor’s theoretical commitments. He wants to divorce semantics from psychology and then focus on semantic questions. At the same time, he wants his semantics to dictate which parts of psychology get to fall under the label ‘concept’, leaving the rest outside the scope of his theory. As a result, he leaves out most of what psychologists study. He offers no account of categorization, inference, belief fixation, deliberation, or decision. One might suppose that this disconnect between what mainstream cognitive scientists investigate and his own project stems from the fact that he is doing foundational work on how minds operate in general, while others are studying applications. But Fodor’s foundations provide little support. Foundational theories are supposed to help with ‘how-possible’ questions, but Fodor’s account does not explain how categorization or inference are possible. Indeed, he insists that psychological capacities are too holistic and anarchic to be studied scientifically. Happily, that grim prognosis is counterbalanced by his interesting discussion of the format used in early perception, but he does not apply the same empirical methodology when positing his language of thought. In fact, his focus on psychosemantics is designed to bypass psychology by focusing on mind–world relations while ignoring the mechanisms that make such relations possible.
To sum up this overarching concern, one might pointedly ask: After three decades, has LOT earned its keep as a research program? Is it guiding research into how the mind works? Could it? How much would it explain even if it were true? At present, there is no serious empirical evidence for LOT (as Fodor says, we would not know what to look for), and the in-principle arguments fall short. It is widely accepted that the mind uses combinatorial representations, but the claim that these are language-like in any stronger sense (amodal, unstructured symbols that represent determinate referents), remains empirically elusive. Even if we could confirm that Mentalese exists, it would shed little light on many of the psychological capacities, such as categorization, for which concepts are postulated in the first place. If that is the price of keeping semantics apart from psychology, so much the worse for semantics. As naturalists, should we not be more interested in how organisms coordinate behaviour with the world than some abstract counterfactual relationship that happens to exist between items in the head and properties in Plato’s Heaven?¹

References


¹ Many thanks to Eric Mandelbaum, David Pereplyotchik, Barry Lee, and four anonymous referees, three of whom later identified themselves as Marat Aydede, Edouard Machery, and Olga Markic.


