This is a rough, uneven, and incomplete first draft to be expanded extensively later:
Please read it accordingly!

COMMENTS, CRITICISMS AND CORRECTIONS ARE MOST WELCOME!

Language of Thought Hypothesis:
State of the Art

MURAT AYDEDE

The University of Chicago, Department of Philosophy
1010 East 59th Street, Chicago, IL 60637, USA
EMAIL: m-aydede@uchicago.edu, (773) 702-8513

CONTENTS

* The sections starting with '*' can be skipped without serious loss of continuity: they are less important for a first reading of this monograph.

0 Introduction

[[To be completed…]]

1 *Common Sense Conception of Beliefs and Other Propositional Attitudes

Common sense attributes a variety of mental states — e.g., beliefs, desires, hopes, fears, regrets, expectations, etc. — to people (and sometimes even to non-humans) to make sense of their behavior — including their verbal behavior. Philosophers standardly call such states propositional attitudes, because they seem to be mental attitudes towards propositions. Since *Beliefs* and *desires*, in particular, seem to play an especially pivotal role, I will, following standard practice, focus on them in what follows.

(i) Common sense seems to take beliefs to be relations between an agent and something else. Typically, this something else is characterized by the complements of what I will call *belief sentences* — sentences used to attribute beliefs to agents, e.g.,

- John believes that Brutus killed Caesar.

Similarly for other attitudes: e.g., John regrets that Brutus killed Caesar; Smith hopes that his father loves his fiancée, but fears that he doesn't; Mary desires that John come to tonight's dinner alone, etc. Let us call the complements of such sentences, i.e., the embedded sentences following *that*-clauses, *complement sentences*.

Common sense seems to take beliefs to be relations between agents and what is described or referred to by the complement sentences. The latter are *what is believed*: we may conveniently call them the *objects of beliefs* whatever they turn out to be. The folk conception seems to be neutral about their ontological status, but paradigmatically takes them to be semantically evaluable, i.e., capable of being true and false. It appears that the following relation between complement sentences and the object of beliefs holds:

- John's belief that Brutus killed Caesar is true if and only if (iff) the complement sentence 'Brutus killed Caesar' is true.

Furthermore, we say that John's belief is true in virtue of the fact that what he believes, i.e., the object of his belief, is true. But *that* is true just in case the complement sentence is true.

Moreover, different agents can believe the same thing, or more generally can have different attitudes towards the same thing. Consider:

- John believes that permitting free use of marijuana will be beneficial, and hopes that one day marijuana use will be free of legal sanctions. His friend, Smith, agrees and has the same hope. Their belief and hope should not surprise anyone given their life style and liberal education. But some of his

friends dispute their beliefs, and fear what they hope. Mary, in particular, believes that what they believe is, not only plainly false, but totally preposterous and morally offensive.

It would be very difficult to make sense of such talk if we didn't take belief sentences, and sentences ascribing propositional attitudes in general, as expressing relations to "objects" that are semantically evaluable and sharable in some intuitive sense by different people and by different kinds of attitudes.

Vendler (1972) draws interesting parallels between verbs of propositional attitudes, especially believing, and verbs of saying (asserting, stating, etc.). He shows that there are very detailed isomorphisms if we classify verbs of saying and propositional attitude verbs on the basis of the syntax of their complement sentences. At a minimum, we may here observe the following phenomena:

- John asserts (sincerely says, states, etc.) that Brutus killed Caesar only if he believes that Brutus killed Caesar.
- What John asserts is true iff what he believes is true.
- John's assertion is true iff John's belief is true.

We also say things like "John asserts what he believes" or "John believes what he asserts."

As noted above, the traditional philosophical gloss on this observation is to say that a belief is a relation between an agent and a *proposition*, conceived as some sort of abstract object, expressed by the complement sentence occurring in the belief sentence attributing the relevant belief to the agent.[1]

(ii) Common sense also takes beliefs as capable of standing to each other in various semantic, evidential, and inferential relations: If John believes that all police officers are corrupt and comes to believe that Smith's brother is a police officer, the folk, all things being equal, expect John to come to believe (to infer) that Smith's brother is corrupt. Notice that the folk license talk about *entailment* relations holding among *beliefs*, not just among the *objects* of beliefs, i.e. what are believed by John:

- What John believes is contradicted by Smith's belief, and confirmed by Marvin's experience with police officers, etc.

This situation seems to be quite standard with respect to other attitudes and other semantic and epistemic relations such as logical equivalence, synonymy, disconfirmation, etc.

---

[1] The contrary view is what is sometimes called the *fusion view*, according to which beliefs are monadic states, not relations. On this view, the typical attributions of belief have a misleading surface grammar. Less misleading (but a bit artificial) attributions would be like 'John believes-that-Brutus-killed-Caesar' just as 'John is bald' attributes a unary property to John. According to this view, it is a pure accident that 'Brutus' occurs both in 'John believes that Brutus killed Caesar' and in 'John believes that Brutus was loved by Romans', just as it is an accident that 'cat' occurs in 'catalogue'. There are very serious difficulties with the fusion theory for which the discussion in Fodor (1978), Field (1978), Stich (1983), Lycan (1981) is quite useful.

(iii) Moreover, common sense typically takes this kind of inference engaged in by John as a causal process: it is in virtue of his previous two beliefs that John now comes to have the third belief. According to the folk, this 'in virtue of' is to be read as causal in a quite literal and robust sense. As we will see below when discussing what I will call the problem of thinking, common sense capitalizes on the fact that these causal relations have the specific character they do because of the specific content of the beliefs involved, or more properly, because of their logico-syntactic structure. For instance, John's first two beliefs of John above cause the belief that Smith's brother is corrupt, but not the belief that Mary is corrupt, or for that matter, the belief that two plus two is four. This is no accident according to the folk: the beliefs causally interact with each other in ways sensitive to their content.

Again, *practical reasoning* and *production of behavior* are typically responsive to the content of the beliefs and desires involved. If John has the specific belief and hope about marijuana mentioned above, and if there is a public referendum as to whether marijuana use should be legalized and John believes that his vote can make a difference, then, ceteris paribus, he will typically form a desire to vote 'yes' in the referendum, which will result (ceteris paribus) in a certain kind of yes-voting behavior. Such means-ends reasoning is paradigmatically responsive to what is wanted and what is believed. It is *because* I believe that drunk driving is potentially life-threatening, and desire not to take a risk that I form the desire not to drink at the party, which in turn is causally involved in my ensuing behavior of sober driving. Again, according to the folk, I formed the desire I did at that point, but not, for instance, the desire to eat chocolate ice cream, *because* I had the specific belief and desire I did: their content was relevant to the causal explanation of why I formed the particular desire not to drink and why I behaved the way I did. On the folk view, what is believed and desired appear to have overlapping (shared) parts, conceptual elements, and this fact is what the folk seem to appeal to as part of the causal story underlying inference, practical reasoning and production of behavior.

The folk explanation of such mental and behavioral phenomena appeals to specific contents of propositional attitudes. But the *form* of the causal explanation seems to involve generalizations *over* propositional attitudes. For instance, it is apparently by appeal to some such generalization as

- For any subject *S*, and for any three beliefs of the form belief that *P*, belief that if *P* then *Q*, and the belief that *Q*, respectively: if *S* comes to have the first two, then, all else being equal, *S* tends to have the third as a causal consequence of having the first two

that the folk explain why and how John came to believe that Smith's brother was corrupt on the basis of his two previous beliefs. Here we seem to have a folk psychological generalization somewhat mirroring the logical rule called *modus ponens*: from any two propositions of the form 'P', and 'if P then Q', 'Q' may be validly inferred. This inference rule is valid in virtue of the fact that it is truth-preserving: if the premises are true then the conclusion, what is inferred, will necessarily be true.

It is presumably in virtue of many such psychological generalizations quantifying over the objects of propositional attitudes that the folk are able to subsume many different agents and to causally explain their behavior and thought processes falling under them. In fact, many have argued that folk psychology**,** with its stock of such (mostly implicit) generalizations and with the pattern of causal explanations they engender forms an implicit psychological theory about those psychological/cognitive aspects of people that seem to be especially relevant to their interactions.

Following Fodor (1985:5, 1987:10), I will call any psychological theory (scientific or philosophical) an *intentional realist* theory if

1. it recognizes as its central theoretical posits mental states with intentional content (in particular, propositional content),
2. it takes them to causally efficacious, and
3. its explanations (thus the generalizations involved therein) of behavior and mental phenomena are recognizably similar to those of common sense belief-desire explanations (and generalizations).

Any scientific or proto-scientific psychological theory that is intentional realist in just this sense is poised to *vindicate* folk psychology conceived in the way characterized above. Not all psychological theories have been intentional realist. For instance, theories that were the product of behaviorism of the early part of this century were not intentional realist. Behaviorism is now dead, and modern (cognitive) psychology is mentalistic through and through, and appears to embrace intentional realism in general. Consequently, many philosophers of mind are already convinced that contemporary cognitive psychology is essentially intentional realist and thus is a clear vindication of folk psychology. Not everyone agrees however. There are theorists who think that, contrary to initial appearances, the intentional idioms of the folk will not survive the advance of scientific psychology, and folk's intentional categories are radically ill suited to become the natural kinds of a successful science. Thus, eliminativism, it is claimed, is a live option.[2] Intentional Realism, thus, constrains the extent to which a psychology is said to vindicate folk psychology.

2 What is the Language of Thought Hypothesis?

The Language of Thought Hypothesis (LOTH — a.k.a. Computational/Representational Theory of Mind, CRTM, see below) is a view about the specific way in which folk psychology will be vindicated by scientific cognitive psychology. More specifically, LOTH is, among other things, an attempt to show that the general framework and intentional categories of folk psychology can be pressed into scientific use in such a way that intentional realism is preserved in the resulting scientific psychology. Moreover, as we will see below, many claim as an argument

[2] See, for instance, Stich (1983) who defends a purely Syntactic Theory of Mind (STM); Dennett's (1981, 1987) instrumentalist proposals; Churchlands' eliminativist theses to replace folk psychology with brain science (P.M. Churchland 1981, 1990; P.S. Churchland 1986).

in favor of LOTH that it is at the foundations of, and thus presupposed by, much of modern cognitive psychology. LOTH as such is a *naturalistic* attempt to show how intentional realism, and thus folk psychology, will turn out to be true.

LOTH is an empirical thesis about the nature of thought and thinking; in particular, it is an hypothesis about the nature of *propositional attitudes* and the processes involving them. It can be characterized as the conjunction of the following three main theses:

(A) **Representational Theory of Mind** (RTM): (cf. Field 1978:37, Fodor 1987:17)

(1) *Representational Theory of Thought*:

For each propositional attitude *A*, there is a unique and distinct (i.e. dedicated)[3] psychological relation *R* and for all propositions *P* and subjects *S*, *S A*s that *P* if and only if there is a mental representation #*P*# such that

(a) *S* bears *R* to #*P*#, and
(b) #*P*# means that *P*.

(2) *Representational Theory of Thinking*:

Mental processes, thinking in particular, consists of causal sequences of tokenings of mental representations.

(B) Mental representations, which, as per (A1), constitute the direct "objects" of propositional attitudes, belong to a representational or symbolic *system* which is such that (cf. Fodor and Pylyshyn 1988: 12-3)

(1) representations of the system have a combinatorial syntax and semantics: structurally complex (molecular) representations are systematically built up out of structurally simple (atomic) constituents, and the semantic content of a molecular representation is a function of the semantic content of its atomic constituents together with its syntactic/formal structure, *and*

(2) the operations on representations (constituting, as per (A2), the domain of mental processes, thinking) are causally sensitive to the syntactic/formal structure of representations defined by this combinatorial syntax.

(C) **Functionalist Materialism**. Mental representations so characterized are, at some suitable level, functionally characterizable entities that are *realized* by the physical properties of the subject having propositional attitudes (if the subject is an organism, then the realizing properties are presumably the neurophysiological properties in the brain or the central nervous system of the organism).

The relation *R* in (A1), when RTM is combined with (B), is meant to be understood as a *computational/functional* relation. The idea is that each attitude is identified with a characteristic computational/functional role played by the mental sentence

[3] This is to convey the basic idea that each type of attitude (e.g., believing) is realized by the same type of computational relation (e.g., being inside the computationally defined B-Box) and by no others. So the mapping from attitudes *A* into computational relations *R* is meant to be injective.

that is the direct object of that kind of attitude. For instance, what makes a certain mental sentence an (occurrent) belief might be that it is characteristically the output of perceptual output systems and input to an inferential system that interacts decision-theoretically with desires to produce further sentences or actions. Or equivalently, we may think of belief sentences as those that are accessible only to certain sorts of computational operations appropriate for beliefs, but not to others. Similarly, desire-sentences (and sentences for other attitudes) may be characterized by a different set of operations that define a characteristic computational role for them. In the literature it is customary to use the metaphor of a "belief-box" (cf. Schiffer 1981) as a blanket term to cover whatever specific computational role belief sentences turn out to have in the mental economy of their possessors. (Similarly for "desire-box," etc.)

The Language of Thought Hypothesis is so-called because of (B): token mental representations are like *sentences* in a language in that they have a syntactically and semantically regimented *constituent structure*. Put differently, mental representations that are the objects of attitudes are structurally complex symbols whose complexity lends itself to a syntactic and semantic analysis. This is also why the LOT is sometimes called *Mentalese*.

As we will see later on, the striking parallelisms between (B1) and the character of *formation* rules that define the well-formedness of expressions in a *formal language* on the one hand, and the one between (B2) and the character of *transformation* rules defined over the well-formed-formulas (wffs) of *formal systems*, on the other, is more than a mere analogy. In fact, according to the defenders of LOTH, LOTH claims explanatory advantages over its competitors precisely because the postulated LOT can be claimed to *constitute* (or be literally characterizable as) an (interpreted) *formal system* in the logician or mathematician's sense of the phrase.

It is important to note that it is (B2) that makes LOTH *a species of* the so-called Computational Theory of Mind (CTM) (about which more below). This is why LOTH is sometimes called the Computational/Representational Theory of Mind or Thought (CRTM/CRTT) (cf. Rey 1991, 1997). Indeed, LOTH seems to be the most natural product when RTM is combined with a view that would treat mental processes or thinking as computational when computation is understood traditionally or *classically* (as it has recently come to be called to emphasize the contrast with connectionist processing or "computation," which we will discuss later).

Before moving on to comment on different aspects of LOTH, let me briefly outline how LOTH proposes to accommodate the observations we have made above in § 1 about the folk conception of propositional attitudes. I hope that from the statement of LOTH much of it is already clear. When someone believes that *P*, there is an obvious sense in which the immediate object of her belief, what she believes, can be said to be a complex symbol, according to LOTH, a sentence in her LOT physically realized in the neurophysiology of her brain, that has both syntactic structure and a semantic content, namely the proposition that *P*. So, contrary to the orthodox view that takes the belief relation as a dyadic relation between an agent and a proposition,

LOTH takes it to be a triadic relation among an agent, a Mentalese symbol, and a proposition. The Mentalese sentence can then be said to have the proposition as its semantic/intentional content. It is only in this indirect/derivative sense can it be said that what is believed is a proposition.[4]

This triadic view seems to have an advantage over the orthodox view in that it is a puzzle in the dyadic view how what are thought to be purely physical organisms can stand in direct relation to abstract objects like propositions in such a way as to influence their causal powers. Remember, according to the folk, it is because those states have the propositional content they do that they have the causal powers they do. LOTH makes this relatively non-mysterious by introducing a physical intermediary that is capable of having the relevant causal powers in virtue of its syntactic structure that encodes its semantic content. Another advantage of this is that the thought processes can be causally guided by the syntactic forms of the sentences in a way that respect their semantic contents. This is the virtue of (B). But then, the observations we made in (ii) and (iii) above about the folk explanation of thinking, practical reasoning, and behavior have a powerful explication in terms of LOTH. (We will come back to this in greater detail below in § 9 and to some of the problems it may involve in § 10.)

According to many LOT theorists who have embraced a more or less Gricean program (Grice 1957) of explaining the semantics of natural languages by appeal to the speaker's intentions (and other propositional attitudes), it is natural to expect that the relation between what one (sincerely) says, asserts, etc. and what one believes is very much what the folk assume it to be: paradigmatically, one asserts what one believes. The asserted proposition is generally taken to be the proposition expressed by the belief (i.e., by the Mentalese sentence that the agent stands in belief relation to) that is implicated in the etiology of the assertion.[5]

Folk psychology is thus vindicated if LOTH turns out to be true.

3 Status of LOTH

Notwithstanding some recent attempts to establish the truth of LOTH on a priori or conceptual grounds (given, of course, the natural conceptual contours of folk psychology — see Davies 1989, 1991; Lycan 1993; Rey 1995), LOTH has primarily been advanced as an *empirical* thesis. It is not meant to be taken as an analysis of what the folk *mean* (or, for that matter, what the scientists ought to *mean*) when they talk

[4] Strictly speaking, talk of propositions can be avoided by a LOT theorist who prefers to talk about the semantic properties of Mentalese symbols, simple as well as complex. This requires that a presumably naturalistic account can be given for the conditions of a Mentalese symbol's having semantic properties. This is indeed what many LOT theorists assume — see below. This may appeal to those who think along with Quine (e.g. Quine 1960) that propositions are not well understood entities, or simply "creatures of darkness." See Devitt (1996) who is a semantic realist but generally avoids introducing propositions.

[5] For more elaboration about this kind of closely interrelated arguments for LOTH whose appreciation draw upon critically reflecting on the common sense notion of propositional attitudes as well as their typical use and ascription by the folk, see Vendler (1972), Fodor (1978, 1980), Field (1978), Stich (1983: chp.3).

about various propositional attitudes and their role in thinking. In this regard, LOT theorists typically view themselves as engaged in some sort of a proto-science, or at least in some empirical research program continuous with scientific psychology or more generally with empirical inquiry. Indeed, as we will see in more detail below, when Jerry Fodor first explicitly articulated and elaborated LOTH in some considerable detail in his (1975), he basically defended it on the ground that it was assumed by our best scientific theories or models in cognitive psychology and psycholinguistics. This empirical status accorded to LOTH should be kept firmly in mind when assessing its plausibility and especially its prospects in the light of new evidence and developments in scientific psychology.[6]

When viewed in this way, LOTH is not, *strictly speaking*, committed to preserving the folk taxonomy of the mental states in any very exact way. Notions like belief, desire, hope, fear, etc. are folk notions and, as such, it may not be utterly plausible to expect (eliminativist arguments aside) that a (completed) scientific psychology will preserve the exact contours of these concepts. On the contrary, there is every reason to believe that scientific counterparts of these notions will carve the mental space somewhat differently. For instance, it has been noted that the folk notion of belief harbors many distinctions. It is noted for example that it has both a dispositional and an occurrent sense. In the occurrent sense, it seems to mean something like consciously entertaining and accepting a thought (proposition) as true. There is quite a bit of literature and controversy on the dispositional sense.[7] Beliefs are also capable of being explicitly stored in long term memory as opposed to being merely dispositional or tacit. Compare, for instance: I believe that there was a big surprise party for my 24th birthday vs. I have always believed that lions don't eat their food with forks and knifes, or that 13652/4=3413, even though until now these latter two thoughts had never occurred to me. There is furthermore the issue of degree of belief: while I may believe that George will come to dinner with his new girlfriend even though I wouldn't bet on it, you, thinking that you know him better than I do, may nevertheless go to the wall for it. It is unlikely that there will be one single construct of scientific psychology that will exactly correspond to the folk notion of belief in all these ways.

So, again, it is an *open empirical* issue to what extent the notions of folk psychology will be preserved in a completed scientific psychology. No one expects scientific psychology to reproduce or reconstruct all and only those concepts employed by folk psychology with exactly the same or very similar features. But many LOT theorists are willing to bet that when the dust settles, scientific psychology will turn out to be an *intentional realist* theory in the way LOT theorists more or less envisage it, and thus will vindicate folk psychology to some significant degree, rather than replacing or eliminating it. For LOTH to vindicate folk psychology it is sufficient that a

[6] This, of course, assumes that the hypothesis is conceptually or otherwise *coherent*. There are objections and arguments accusing LOT theorists of being theoretically irresponsible by proposing a completely fortuitous hypothesis or even of somehow proposing an incoherent thesis. We will discuss some of these more conceptually motivated objections later on in § 8.
[7] Lycan (1986), Davies (1989, 1995), Cummins (1986), Hadley (1995). A parallel discussion is going on in AI: Kirsh (1990).

scientific psychology with a LOT architecture come up with scientifically grounded psychological states that are recognizably like the propositional attitudes of folk psychology, and that play more or less similar roles in psychological explanations.[8]

4 Scope of LOTH

LOTH is an hypothesis about the nature of thought and thinking with propositional content. As such, it may or may not be applicable to other aspects of mental life. Officially, it is silent about the nature of some mental phenomena such as experience, qualia,[9] sensory processes, mental images, visual and auditory imagination, sensory memory, perceptual pattern-recognition capacities, dreaming, hallucinating, etc. To be sure, many LOT theorists hold views about these aspects of mental life that make it seem that they are also to be explained by something similar to LOTH.[10]

For instance, Fodor seems to think that many modular input systems (Fodor 1983) have their own LOT to the extent to which they can be explained in representational and computational terms. Indeed, many contemporary psychological models treat perceptual input systems in just these terms.[11] There is indeed some evidence that this kind of treatment is appropriate for many perceptual processes. But it is to be kept in mind that a system may employ representations and be computational without necessarily satisfying either or both of the clauses in (B) above in any full-fledged way. Just think of finite automata theory where there are plenty of examples of a computational process defined over states or symbols that lack full-blown syntactic and/or semantic structural complexity. Whether sensory or perceptual processes are to be treated within the framework of full-blown LOTH is again an open empirical question. It may well be that the answer to this question is affirmative. If so, there may be more than one LOT realized in different subsystems or mechanisms in the mind/brain. So LOTH is not committed to there being a single representational system realized in the brain, nor is it committed to the claim that all mental representations are complex or language-like, nor would it be falsified if it turns out that most aspects of mental life other than the ones involving propositional attitudes don't require a LOT.

Similarly, there is strong evidence that the mind also exploits an image-like representational medium for certain kinds of mental tasks.[12] LOTH is non-committal about the existence of an image-like representational system for many mental tasks other than the ones involving propositional attitudes. But it *is*

---

[8] See Fodor (1985, 1986, 1987: chp.1), Devitt (1990).
[9] But see Rey (1992, 1993) for an attempt to expand LOTH to sensations and qualia.
[10] See for instance the controversy involved in the so-called imagery debate. The literature here is huge but the following sample may be useful: Block (1981, 1983b), Dennett (1978), Kosslyn (1980), Pylyshyn (1978), Rey (1981), Sterelny (1986), Tye (1991).
[11] E.g., Marr (1982), or any textbook on vision or language comprehension and production.
[12] E.g., Kosslyn (1980, 1994); Shepard and Cooper (1982). In fact, some theorists even go so far as to claim that *all* cognition is done in an image-like symbol system — early British empiricists from Locke to Hume held something like this view, but more recently, see L. Barsalou and his colleagues who have been developing models to that effect (Barsalou 1993a, Barsalou et al 1993b, Barsalou and Prinz 1997).

committed to the claim that propositional thought and thinking cannot be successfully accounted for in their entirety in purely imagistic terms. It claims that a combinatorial sentential syntax is necessary for propositional attitudes and a purely imagistic medium is not an adequate medium to capture that.[13]

There are in fact some interesting and difficult issues surrounding these claims. The adequacy of an imagistic system seems to turn on the nature of syntax at the *sentential* level. For instance, Fodor, in Chapter 4 of his (1975) book, allows that many lexical items in one's LOT may be image-like; he introduces the notion of a *mental image/picture under description* to avoid some obvious inadequacies of pictures (e.g., what makes a picture a picture of a fat woman rather than a pregnant one, or vice versa, etc.). This is an attempt to combine discursive and imagistic representational elements at the *lexical* level. There may even be a well defined sense in which pictures can be combined to produce structurally complex pictures (as in British Empiricism: image-like simple ideas are combined to produce complex ideas, e.g., the idea of a unicorn) But what is absolutely essential for LOTH, and what Fodor insists on, is the claim that there is no adequate way in which a *purely* image-like system can capture what is involved in making judgments, i.e., in judging *propositions* to be true. This seems to require a discursive syntactic approach at the sentential level. The general problem here is the inadequacy of pictures or image-like representations to express *propositions*. I can judge that the blue box is on top of the red one without judging that the red box is under the blue one. I can judge that Mary kisses John without judging that John kisses Mary, and so on for indefinitely many such cases, concrete as well as abstract. It is hard to see how images or pictures can do that *without using any syntactic structure or discursive elements*, to say nothing of judging, e.g., conditionals, disjunctive or negative propositions, quantifications, negative existentials, etc.[14]

Moreover, there are difficulties with imagistic representations arising from demands on *processing* representations. As we will see below, (B2) turns out to provide the foundations for one of the most important arguments for LOTH: it makes it possible to mechanize thinking understood as a semantically coherent thought process, which, as per (A2), consists of a causal sequence of tokenings of mental representations. It is not clear, however, how an equivalent of (B2) could be

[13] The controversial issue here is not the absurdity of the claim that there are literally pictures or images in the brain. Probably no one believes this claim these days. Rather, postulating picture-like representations is to be cashed out in functionalist terms. Pictures as mental representations presumably bear some non-arbitrary isomorphisms to what they represent, although it is hard to make this sort of claim crystal-clear in purely functionalist terms. See, for instance, Kosslyn (1980, 1981), Block (1983a, 1983b), Tye (1984).

[14] The issues here are too complex and difficult to go over here in any useful detail, but for a general criticism of pictures as mental representations, see the critical essays in Block (1981) and Rey (1981); for an attempt to overcome many such criticisms, see Barsalou and Prinz (1997) and Prinz (1997). The contemporary debate about the adequacy of a purely imagistic medium for capturing what is involved in making a judgment and discursive thinking seem to parallel some of Kant's critique of British Empiricism in general and of Hume's associationism in particular, as indeed emphasized by many classicists like Fodor and Pylyshyn (1988), Rey (1997).

provided for images or pictures in order to accommodate operations defined over them, even if something like an equivalent of (B1) could be given. On the other hand, there are truly promising attempts to *integrate* discursive symbolic theorem proving with reasoning with image-like symbols. They achieve impressive efficiency in theorem proving or in any deductive process defined over the expressions of such an integrated system. Such attempts, if they prove to be generalizable to psychological theorizing, are by no means threats to LOTH; on the contrary, such systems have every features to make them a species of a LOT system: they satisfy (B).[15]

5 *Natural Language as Mentalese?

What is the exact relation of an organism's language of thought to the natural language she speaks (if she speaks one)? Some theorists (e.g. Harman 1973, one of the earliest defenders of LOTH) take one's language of thought to be identical to one's natural language in the sense that sentences of LOT are quite generally isomorphic to the syntactic and semantic constitution of the sentences of the natural language. There are of course variations and nuances. For instance, Devitt and Sterelny (1987) think that the innately determined Mentalese the organism begins its development with is expressively and grammatically quite weak, but strong enough to allow the organism to begin to acquire the first elements of its natural language; then the rest of the development is a strategy of bootstrapping. (It appears that Field 1978 has also a similar story in mind.) Fodor (1975), and perhaps the majority of the theorists who accept LOTH, think that Mentalese is quite distinct from one's natural language. There are various reasons for thinking so: some of them are pretty strong (see Fodor 1975; Pinker 1994, 1997). But the point to emphasize here is that LOTH comes in various forms in this respect. Officially, LOTH should be thought as being neutral on the character of the mental language. To repeat, the claim of LOTH, minimally but crucially, is that the mental representations constitute a system of which (B) is true. So to the extent to which natural languages satisfy at least (B1), it is (at least partly) an open empirical issue as to whether one's representational system is identical, in some suitable sense of isomorphism, to one's natural language. In what follows, however, I will have the more popular Fodor's version of LOTH in mind.

6 *Nativism and LOTH

Fodor's reason for thinking that Mentalese cannot be identical to any natural language has important connections to his notoriously strong version of nativism (Fodor 1975:Chps.2-3; 1981). He reasoned, very roughly, as follows. Learning a natural language, is, among other things, learning certain T-sentences that express the truth conditions of the sentences of the target language to be learned (e.g. 'Snow is white' is true iff snow is white). But if so, learning a language (be it one's native or second language) essentially involves forming and confirming hypotheses (T-sentences). But this involves already possessing a representational medium, essentially a system, in which the hypotheses are formed and confirmed. Thus LOT

[15] See, e.g., Barwise and Etchemendy's *Hyperproof* (1995).

cannot be one's natural language, since acquiring a natural language presupposes already possessing an inner representational system to do the forming and confirming hypotheses, hence LOT. He thought that learning a natural language is basically learning how to translate natural language sentences into one's LOT.

He had a similar story about concept learning. He thought of concept *learning* (as opposed to *acquiring* a concept, for which a nativist can make some room), again, as essentially involving forming and confirming hypotheses about the (extensions of) concepts to be learned. But this seems to imply that if one is capable of forming a hypothesis expressing the target concept, then there is a non-trivial sense in which one already has that concept — albeit potentially. Hence Fodor's notorious nativism about all concepts.[16]

It is historically unfortunate that Fodor ran his arguments for the innateness of all concepts in the same book (1975) in which he first elaborated and defended LOTH in such a way that the connection between LOTH and an implausibly strong version of nativism looked very much internal. As a result, this historical coincidence has led some people to think that LOTH is essentially committed to a very strong version of nativism, so strong in fact that it seems to make a *reductio* of itself (see, for instance, P.S. Churchland 1986, Putnam 1988).

However, it should be emphasized that LOTH is *not* per se committed to such a strong version of nativism, especially about *concepts*. It is certainly plausible to assume that LOTH will turn out to have some empirically (as well as theoretically / a priori) motivated nativist commitments especially about the structural organization and dynamic management of the entire representational system. But this much is to be expected especially in the light of recent empirical findings and trends of contemporary cognitive and developmental psychology as well as psycholinguistics. This, however, does not constitute a *reductio*. On the other hand, LOTH is by no means committed to the innateness of all *concepts* or even some of them. It is an open empirical question how much nativism is true about concepts, and LOTH should be so taken as to be capable of accommodating whatever turns out to be true in this matter. LOTH, therefore, when properly conceived, is independent of any specific proposal about *conceptual* nativism.[17]

7 Naturalism and LOTH

Cartesian mind-body dualism can certainly be seen as an instance of intentional realism, as would many theories of early British Empiricists like Locke's and Hume's. Yet, they would not be taken seriously as candidates for providing the foundations of modern scientific psychology. The difference is the commitment of present day's theorizing to naturalism, or perhaps a bit more precisely, to physicalism (understood here as neutral between token or type identity theories).

[16] Fodor in his (1998) seems to have changed his mind on conceptual nativism: he now seems to think that most of our concepts, though may not be innate, are such that their extensions are essentially mind-dependent!

[17] For a non-nativist but otherwise quite Fodorian account of concept acquisition, see Margolis (forthcoming).

To demand of a theory that it be naturalistic is to demand something not very clear and precise. Nevertheless, the intuition goes very deep: the entities, events, processes and mechanisms postulated by a naturalistic theory are not allowed to have properties that could not in principle be explained in completely physical/functional terms. By this prohibition, Cartesian thinking substance, for instance, is certainly out, and so are disembodied souls, spirits, fairies and other spooky stuff. More illuminatingly, however, the demand of naturalism is meant to disallow postulating properties that are irreducibly psychic or mental as belonging to the primitive ontological structure of the world. We may illustrate the point with a parallel case of attempts to naturalize life: until recently, life had been thought to be just such a basic element, an irreducible constituent of the universe. At one of its best theoretical articulations, the force behind it was simply called *élan vital*, and was left unanalyzed as a simple primitive force to be reckoned as one of the essential constituents of ontological order. But this had always been an uneasy thought for scientifically oriented minds. We can now say with confidence that life has been naturalized: we now know the physico-chemical bases of cellular life. There is no scientific mystery left about how life can arise out of mere matter.

Similarly, there has always been a strong tendency to think of consciousness (qualia, conscious experience) and intentionality (aboutness, representational aspect) of mental phenomena as somehow ontologically and explanatorily special, belonging to an order radically different from the mere material or physical. These mental phenomena have puzzled some people so much so that even the very feat of conceiving how the mental *could possibly* arise out of the mere physical was sometimes declared to be impossible.[18] Minimally, this is the challenge to be met by naturalists: to show in some plausible way how the conscious and intentional mind could possibly arise out of mere matter.

One of the most attractive features of LOTH is that it is a central component of an ongoing research program in philosophy of psychology to naturalize the mind, to give a theoretical framework in which the mind could naturally be seen as part of the physical world without postulating irreducibly psychic entities, events, processes or properties. Fodor, the most ardent defender of LOTH, once identified the major mysteries in philosophy of mind thus:

> How could anything material have conscious states? How could anything material have semantical properties? How could anything material be rational? (where this means something like: how could the state transitions of a physical system preserve semantical properties?). (1991: 285, Reply to Devitt)

LOTH is a full-blown attempt to give a naturalist answer to the third question, is an attempt to solve at least *part of* the problem underlying the second one, and is almost completely silent about the first.[19]

---

[18] See, for example, Descartes (1637/1970; 1649/1970), Brentano (1874/1973), McGinn (1991).

[19] But, again, see Rey (1992, 1993) for an attempt to extend LOTH in this direction.

According to RTM, propositional attitudes are relations to meaningful mental representations whose tokenings constitute the domain of thinking. This much can, in principle, be granted by an intentional realist who would nevertheless reject LOTH. Indeed, there are plenty of theorists who accept RTM in some suitable form (and *also* happily accept (C) in many cases) but reject LOTH either by explicitly rejecting (B) or simply by remaining neutral about it. Some prominent people who chose the former option are Searle (1984, 1990, 1992), Stalnaker (1984), Lewis (1972), Barwise and Perry (1983).[20] Among the latter, we might include Loar (1982a, 1982b), Dretske (1981); Armstrong (1980), and many contemporary functionalists.[21]

But RTM per se doesn't so much propose a naturalistic solution to intentionality and mechanization of thinking as simply assert a framework to emphasize intentional realism and, perhaps, with (C), a declaration of a commitment to naturalism or physicalism at best. How, then, is the addition of (B) supposed to help? Let us first try to see in a bit more detail what the problem is supposed to be in the first place to which (B) is proposed as a solution. So let us start by reflecting on thinking and see what it is about thinking that makes it a mystery in Fodor's list. This will give rise to one of the most powerful (albeit still nondemonstrative) arguments for LOTH.

7.1 The Problem of Thinking

RTM's second clause (A2), in effect, says that thinking is at least the tokenings of states that are (a) intentional (i.e. have representational / propositional content) and (b) causally connected. But, surely, thinking is more. There could be a causally connected series of intentional states that makes no sense at all. Thinking, therefore, is causally proceeding from states to states that would make semantic sense: the transitions among states must preserve some of their semantic properties to count as thinking. In the ideal case, this property would be the truth value of the states. But in most cases, any interesting intentional property like warrantedness, degree of confirmation, semantic coherence given a certain practical context like satisfaction of goals in a specific context, etc. would do. In general, it is hard to spell out what this requirement of "making sense" comes to. The intuitive idea, however, should be clear. Thinking is not proceeding from thoughts to thoughts in arbitrary fashion: thoughts that are causally connected are in some fashion

[20] Also, Hubert Dreyfus and John Haugeland's many writings indicate that they are hyperrealist about propositional attitudes but would reject LOTH nevertheless.

[21] Almost all British empiricists might be put in this latter category too, but they were in fact closer to LOTH by having embraced something like (B1) in some imagistic version. But it looks as if they could not be better than being associationist regarding *thought processes*: they could not exploit the clear implications of modern symbolic logic and the advancement of computers — they did not have their Frege and Turing, though Hobbes came close. This rendering of RTM relies on a broad interpretation of the notion of mental representation, of course, which has not always been the intended interpretation of Fodor: there are many places where he defends RTM (by that name) meaning to include (B) by default (Fodor 1981b, 1985, 1987, 1998). This should cause no confusion. Here I have chosen to stick to the literal meaning of the phrase rather than to its historically more accurate use — this has become necessary, at any rate, in the light of the recent classicism / connectionism debate to which we will return below.

semantically connected too. If this were not so, there would be little point and gain in thinking. Thinking couldn't serve any useful purpose. Call this general phenomenon, then, the *semantic coherence* of causally connected thought processes. But thinking seems to be something still more. For you can have causally connected state transitions that would make semantic sense, but nevertheless wouldn't, intuitively, count as thinking. Any scenario under which a series of semantically coherent state transitions would be causally connected to each other but "in the wrong sort of way" would illustrate the point. There are some nice illustrations of this kind of scenario in Rey (1995), but one is particularly striking. Rey quotes Davidson as worrying about how to capture intentional causation of action as a species of practical inference brought about "in the right sort of way":

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally. (Davidson 1980: 79)

Here we have causation and semantic coherence (understood broadly as indicated above), nevertheless the action causally comes about in the wrong sort of way. Here, intuitively, the set of properties of this belief / desire pair explaining why the process makes semantic / practical sense appears not to be the set of properties that would also be causally responsible for the ensuing behavior. Intuitively, we want the very same properties that would make the transitions coherent to be the ones that are also causally implicated in the state transitions (or the causation of the purposeful action), as indeed demanded by intentional realism. But what are those properties of states by virtue of which we see them as "making sense"? The answer seems to be the logico-semantic properties of thoughts. What is good about Davidson's argument is that it seems to show that we (not just the theorist but the folk as well) do indeed have this intuition: namely, we indeed distinguish actions, which would make perfect sense in a given context, that are brought about in the wrong way from the ones brought about in the right way. Now apply this case to pure thought processes. The situation seems parallel. We do seem to care about how thoughts are caused in order for them to count as thinking in this (perhaps strong) sense. Thinking involves tokening of thoughts that are causally brought about in the right sort of way. In brief, there seems to be a robust sense of 'thinking' according to which the very same properties of thoughts that explain the semantic coherence of a thought process, i.e. the logico-semantic properties, are to be causally implicated in the state transitions that constitute the process, just as intuition, and with it, intentional realism, demand.

To be sure, in some sense, there are other, less stringently characterized, thought processes that fall under the heading of 'thinking' not only in the ordinary folk parlance but also in cognitive psychology. But whatever else may qualify as thinking, it is thinking in this more stringent but perfectly respectable sense that is used as an argument for LOTH. LOTH is offered as a solution to this puzzle: how is

thinking, conceived in this (strong) sense, physically possible? This is the problem of thinking, thus the problem of mechanization of rationality in Fodor's version.[22]

How does LOTH propose to solve this problem and bring us one big step closer to the naturalization of the mind?

7.2 Syntactic Engine Driving a Semantic Engine: Computation

The two most important achievements of the 20th century that are at the foundations of LOTH as well as most of modern Artificial Intelligence (AI) research and the so-called information processing approaches to cognition (practically all of contemporary cognitive psychology) are (i) the developments in modern symbolic (formal) logic, and (ii) Alan Turing's idea of a Turing Machine and Turing computability. It is putting these two ideas together that gives LOTH its enormous explanatory power within a naturalistic framework.

(i) *Formal Logic*. Twentieth Century logic, building on the pioneering work of Russell, Frege, and Gödel, was able to successfully formalize most of deductive reasoning. Formalization, or the so called proof-theoretic approach to logic, showed that the important semantic concepts like validity, entailment, consistency, etc., can be explicated completely in formal / syntactic terms, without using any semantic notion whatsoever. Proof-theory is the branch of logic and mathematics that consists in the attempt to construct formal languages capable of capturing logical / mathematical facts like valid arguments, tautologies, etc. To this end, the languages are constructed solely on the basis of non-semantic, formal features of symbols. In this regard, what is important are not the intended meanings of symbols but their formal properties like their physical shape, spelling, or syntactic categories. The typical stages in constructing a formal language is to provide an alphabet, then rules for putting the items in the alphabet together to form well-formed formulas (wffs), or grammatically / syntactically correct sentences. Theoretically, the most interesting and important formation rules are combinatorial or recursive. Here is an example:

Alphabet of Sentential Logic, SL:

Atomic sentences: p, q, r, etc.
Connectives: ~, v, &, ->
Parentheses: (,)

Formation Rules:

1. If ø is an atomic sentence, then it is a sentence, i.e., wff.
2. If ø and µ are wffs then '~ø', '(øvµ)', '(ø&µ)', '(ø->µ)' are wffs.
3. Nothing else is a wff of SL.

Note that Formation Rules define infinitely many wffs, since the rules can be reiterated and applied to previously formed formulas. Once we have the syntactic

[22] I cannot help here but add that this strong way of characterizing thinking may be weakened without removing the mystery and thus without argumentatively underpowering LOTH as long as the *semantic coherence* of thought processes still needs to be explained.

conditions on wffs, we can use them to manipulate these wffs purely on the basis of their form or syntax by adding *Transformation Rules*.

Transformational Rules:

1. ~ø, (øvμ) :— μ and ~μ, (øvμ) :— ø
2. (ø&μ) :— μ and (ø&μ) :— ø
3. ø, (ø->μ) :— μ
4. μ := ~~μ and ~~μ :— μ
5. ø :— øvμ

etc.

Take (3) for instance: it says that given any two wffs of the form 'ø' and '(ø->μ)', it is permissible to derive 'μ'. Similarly for other rules.

What is to be noted here is that the specification of this simple formal system has not used any semantic notions like truth, validity, or what the wffs mean, etc. The conditions on being grammatical are all aspects of shape, spelling, spatial order, etc. The rules for manipulating the wffs are specified, again, purely on the basis of their shape and form. But, as is well known, this system has a systematic semantic interpretation: wffs of SL may be interpreted as expressing propositions that are *true* or *false*, in terms of which *validity* of arguments can be defined. There are two important points to be made here.

First, the truth-value of a complex wff can be exhaustively determined by the truth-values of the atomic sentences it contains: the connectives of SL are said to be *truth-functional*. So for instance, a *conjunction*, i.e. a wff of the form '(ø&μ)', is true just in case all its *conjuncts*, i.e. 'ø' and 'μ', are true, and false otherwise. Similarly, a *conditional*, i.e. a wff of the form '(ø->μ)', is false just in case its *antecedent*, i.e. 'ø', is true and *consequent*, i.e. 'μ', is false, and true otherwise. The rules of assigning truth-values to complex wffs can be specified in so-called truth tables like:

ø	μ	(ø->μ)
T	T	T
T	F	F
F	T	T
F	F	T

SL is said to have a compositional semantics because the truth-values of complex wffs can be uniquely determined given the truth-values of its constituent wffs together with its syntactic/grammatical form.

The second important point is that transformation rules can be interpreted as rules of *inference*, i.e. rules that determine what *validly* follows from what, so that *valid argument forms* can be established. A valid argument form is one whose

conclusion can't possibly be false if the premises are all true. Let's have a look at an example, *Modus Ponens*: 'ø, (ø->µ) :— µ'

ø	µ	ø	(ø->µ)	µ
T	T	T	T	T
T	F	T	F	F
F	T	F	T	T
F	F	F	F	F

The rows containing the truth-values exhaust all the possible semantic interpretations, yet there is no row in which the premises, 'ø' and '(ø->µ)', are true and the conclusion, 'µ', is false. But by formally specifying this rule, we forget that it has a semantic interpretation and thus can be interpreted as reflecting a valid argument form.

This is elementary sentential (propositional) logic, but it illustrates rather nicely the general principles of division of labor exploited by LOTH. The formal system SL is capable of completely capturing all the semantic facts in propositional logic. The *proofs* of *theorems* will be conducted purely on the basis of *formal* properties of SL, but what is thus proved will all be *tautologies*, i.e. wffs that are *true* under every interpretation (e.g., '(pv~p)'). The derivations or transformations of wffs will be conducted all formally / syntactically, but the formulas thus derived will be true if the previous wffs (premises) are all true, i.e. the derivations will capture valid argument forms.

One of the major results of logical research in this century was the discovery that first-order predicate logic (which is an extremely powerful extension of sentential logic) is *complete* (Gödel 1930): that is, all its tautologies are *theorems* (i.e. wffs that are provable formally) and all the *valid argument forms* are capable of being captured by purely syntactic *transformation rules*. The converse also turned out to be the case: all *theorems* are *tautologies*, and all *formally permissible derivations* correspond to *valid arguments*. This discovery meant that all semantic aspects of first order logic can be captured purely syntactically (formally, proof-theoretically) and vice versa. There is a useful analogy to be made here: the semantic aspects of first-order logic mirrors or mimics its syntactic or formal aspects, and vice versa.

Now, why is this important from the perspective of philosophy of psychology? Suppose that this proof-theoretic approach in logic can be successfully extended to other languages (formal or otherwise) and all the forms of "good" arguments that can be couched in them (expressing not just deductive reasoning, but also inductive and abductive reasoning, practical reasoning, using perhaps versions of what the AI researchers called non-monotonic logic combined with relevance logic, etc. — along with powerful heuristic programming perhaps). This would surely be a major

theoretical achievement all by itself in formal semantics, but we can appreciate its consequences for the philosophy of psychology when we combine it with the implications of Turing's results. This is the second important idea I mentioned above.

(ii) *Turing Machines*. Turing showed that all intuitively computable functions, when properly regimented, are Turing Machine computable. There is an intuitive sense in which Turing can be said to have theoretically reduced the intuitive notion of a computable function to Turing Machine computability. Now forming wffs and formally manipulating them on the basis of rules can be characterized in terms of computable functions. Ordinarily and somewhat loosely, when people talk about the formalization of representational processes, that is what they have in mind: capturing the processes in formal terms in such a way that their computability is revealed. This is partly why formalization is so important. But then, by Turing's results, they are Turing Machine computable. The significance of this, in turn, lies in the fact that these processes can then be realized by physically realized machines or systems (and if LOTH is true, in organisms, i.e. on the assumption that the brain of thinking organisms is a kind of computer).

A Turing Machine is in fact an idealization, a simple abstract device. But it can be thought of in terms of a configuration of a certain set of simple physical devices consisting of an indefinitely long tape divided into small cells on each of which a finite set of simple symbols (like *1, 0, X*) can be written one at a time by a scanner-printer. The scanner-printer can see only one cell of the tape at a time, it can read, erase, or write a symbol in the cell and can move one cell to the right or the left depending on its "instructions" in the machine table. The machine table of a Turing Machine can be characterized completely in terms of a finite set of conditionals like:

- If the symbol on the current cell is *1* and you're in state S1 then replace it with *X*, and move to the right.
- If the symbol on the current cell is *0* and you're in state S2, then leave it as such and move to the left.
- If the symbol on the current cell is *0* and you're in state S3, then replace it with *1* and move to the left. Etc.

The historical importance of the idea of a Turing Machine was in its naturalization of the intuitive idea of an algorithm or more precisely of a computable function. It was naturalistic partly because Turing machines were physically realizable (not that anyone would actually bother to realize them). The physical/engineering requirements for building such machines are trivial, but their computational power was universal and complete in that any intuitively computable function, and thus any formalizable representational process, could be computed by a Turing Machine, thus by a physical device. Moreover, Turing showed that there are *universal* Turing Machines, machines that can take as their own program a properly regimented description of another Turing Machine, i.e. its machine table together with its

symbols, and compute the function *it* computes, hence the idea of a programmable computer.

Turing's idea of a Turing Machine was a simple abstract device to make an extremely important theoretical point. For our purposes, we may conveniently put it thus: every formally regimented representational process can in principle be realized by physical devices.

We can now appreciate the implications of (i) and (ii) for the philosophy of psychology more explicitly: if thinking consists in processing representations physically realized in the brain (in the way the internal data structures are realized in a computer) and these representations form a formal system, i.e. a language with its proper combinatorial syntax (and semantics) and a set of derivations rules formally defined over the syntactic features of those representations (allowing for specific but extremely powerful programs to be written in terms of them), then the problem of thinking, as I described it above, can in principle be solved in completely naturalistic terms: thus the mystery surrounding how a physical device can ever have semantically coherent state transitions (processes) can be removed. Thus, given the commitment to naturalism, the hypothesis that the brain is a kind of computer trafficking in representations in virtue of their syntactic properties is the basic idea of LOTH (and the AI vision of cognition).

Computers are environments in which symbols are manipulated in virtue of their formal features, but what is thus preserved are their semantic properties, hence the semantic coherence of symbolic processes. Slightly paraphrasing Haugeland (cf. 1985:106), who puts the same point nicely in the form of a motto:

> THE FORMALIST MOTTO:
> If you take care of the syntax of a representational system, its semantics will take care of itself.

This is in virtue of the mimicry or mirroring relation we have seen above between the semantic and formal properties of symbols. As Dennett once put it in describing LOTH, we can view the thinking brain as a syntactically driven engine preserving semantic properties of its processes, i.e. driving a semantic engine. What is so nice about this picture is that if LOTH is true we have a naturalistically adequate causal treatment of *thinking* that respect the semantic properties of the *thoughts* involved: it is in virtue of the physically coded syntactic/formal features that thoughts cause each other while the coherence of their semantic properties is preserved precisely in virtue of this. Here is how Fodor makes the same point in a concise but powerful manner:

> You connect the causal properties of a symbol with its semantic properties via its syntax. The syntax of a symbol is one of its higher-order physical properties. To a metaphorical first approximation, we can think of the syntactic structure of a symbol as an abstract feature of its [geometric or acoustic — Fodor 1985:22] shape. Because, to all intents and purposes, syntax reduces to shape, and because the shape of a symbol is a potential determinant of its causal role, it is fairly easy to see how there could be environments in

> which the causal role of a symbol correlates with its syntax. It's easy, that's to say, to imagine symbol tokens interacting causally in virtue of their syntactic structures. The syntax of a symbol might determine the causes and effects of its tokenings in much the same way that the geometry of a key determines which locks it will open. (1987: 18–9)

Whether or not LOTH actually turns out to be empirically true in the details or in its entire vision of rational thinking, this picture of a syntactic engine driving a semantic one can at least be taken to be an important *philosophical* demonstration of how Descartes' challenge *can* be met (cf. Rey 1997: chp. 8). Descartes claimed that rationality in the sense of having the power "to act in all the contingencies of life in the way in which our reason makes us act" couldn't possibly be possessed by a purely physical device:

> The rational soul ... could not be in any way extracted from the power of matter ... but must ... be expressly created. (1637/1970: 117-18)

He was completely puzzled by just this rational character and semantic coherence of thought processes so much so that he failed to even imagine a possible mechanistic explication of it. He thus was forced to appeal to Divine creation. But we can now see/imagine at least a possible mechanistic/naturalistic scenario. I think this is enough to dissolve the mystery and remove Descartes' bafflement.[23]

But LOTH is not merely a philosophical response on the part of the naturalist, as indicated before: it is also advanced as a bold empirical hypothesis, claimed to underlie almost all scientific research in contemporary cognitive psychology.

*7.3 *Intentionality and LOTH*

But where do the semantic properties of the mental representations come from in the first place? How is it that the syntactically structured symbols represent anything? How can they mean anything? How is (original) intentionality possible in a world composed of pure matter? This is Brentano's challenge to a naturalist. Brentano's bafflement was with the intentionality of the human mind, its apparently mysterious power to represent things, events, properties in the world. More than a century ago, Brentano wrote :

> Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or, mental) inexistence of an object, and what we would call, although not in entirely unambiguous terms, the reference to a content, a direction upon an object (by which we are not to understand an actually existing reality in this case) or an immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not do so in the same way. In presentation, something is presented, in judgment something is affirmed or denied, in love loved, in hate hated, in desire desired and so on. (Brentano, 1874/1973: 88-9)

[23] For a powerful elaboration of this line of thought, see Rey (1997).

The peculiar thing that Brentano emphasizes here is that contentful mental states can be said to have an object even though the object they are said to be about or directed toward does not exist in reality. In the words of a contemporary American philosopher, Roderick Chisholm, who elaborated on Brentano's thesis: "Diogenes could have looked for an honest man even if there hadn't been any honest men. The horse can desire to be fed even though he won't be fed. James could believe that there are tigers in India, and *take* something there to be a tiger, even if there aren't any tigers in India" (1957:169). Similarly, you can have beliefs about future events, even about events that are said to fall outside of our light cone. In just this sense, it seems an amazing and miraculous feat that we can have false beliefs! When John believes that the Earth is flat he represents the world as being a certain way, i.e. as being flat. That is *what* he thinks, *that* is the object of his thought, apparently an intentional object or fact, namely the Earth's being flat, but the fact isn't there, so to speak: it doesn't obtain.

More generally, from the fact that somebody believes that *a* is *F*, we can't routinely infer that there is an *a* that one believes to be *F*, for any *a* and *F*: Again, existential generalization is said to fail to apply in belief contexts generally. Compare the situation to purely "physical relations." Again in Chisholm's words: "In order for Diogenes to sit in his tub, there must be a tub for him to sit in: in order for the horse to eat his oats, there must be oats for him to eat: and in order for James to shoot a tiger, there must be a tiger there to shoot." (1957: 169)

But if Brentano is right, for a representation to represent *X* at a time, there doesn't need to be any *X* at that time, which is not to say that representations typically don't represent things that exist. They typically do. This is in fact what their fundamental utility consists in. But what seems peculiar to representational mental states like beliefs and desires, i.e. mental states with intentional / semantic content, is that their current representational power doesn't seem to depend essentially on the actual current existence of what they represent, or on any current causal interaction, or for that matter any interaction at all, with what they represent.

As representational physical organisms, how do we manage to do that? How do our mental states manage to do that? Brentano thought that intentionality in this sense is exactly what makes minds so special and puts them outside of the natural order of things. He thought that nothing physical can have this property of intentionality:

> The reference to something as an object is a distinguishing characteristic of all mental phenomena. No physical phenomenon exhibits anything similar. (1874 / 1973: 97)

This problem of intentionality is the second problem or mystery in Fodor's list that I quoted above. I said that LOTH officially offers only a partial solution to it and perhaps proposes a framework within which the remainder of the solution can be couched and elaborated in a naturalistically acceptable way.

As characterized at the beginning, RTM contains a clause (A1b) that says that the immediate object of a propositional attitude that *P* is a mental *representation #P#* that *means* that *P*. Again, (B1) attributes a compositional *semantics* to the

syntactically complex symbols belonging to one's LOT that are, as per (C), physically realized in the brain of a thinking organism. According to LOTH, the semantic content of propositional attitudes is inherited from the semantic content of the mental symbols. So Brentano's question for a LOT theorist becomes: how do the symbols in one's LOT get their meanings in the first place?

There are two levels or stages at which this question can be raised and answered:

(1) At the level of *atomic* (simple) symbols: how do the atomic symbols represent what they do?

(2) At the level of *molecular* (phrasal complexes or sentences) symbols: how do molecular symbols represent what they do?

There have been at least two major lines LOT theorists took regarding these questions. The one that is least committal might perhaps be usefully described as the official position regarding LOTH's treatment of intentionality. Most LOT theorists seem to have taken this line.

The official line doesn't propose any theory about the first stage, but simply assumes that the first question can be answered in a naturalistically acceptable way. In other words, officially LOTH simply *assumes* that the *atomic* symbols/expressions in one's LOT have whatever meanings they have. The LOT theorist therefore assumes that providing an answer to the first question is a different project, a project that LOTH ultimately assumes can be naturalistically pursued and successfully completed, but she insists that this project of naturalizing intentionality at the atomic level, is not to be confused with viewing the mind as a sort of computer. Put differently, it is claimed that LOTH in its bare outline cannot and ought not to be credited with a claim to the solution of intentionality. Naturalizing intentionality is something about which LOTH is officially silent at least at the level of atomic symbols (but see below).

On the other hand, a number of proposals have been offered by contemporary theorists (who are not necessarily defenders of LOTH as opposed to being mere RTM theorists but whose proposals can be adapted by LOT theorists) about how exactly to pursue that project.[24] Of course, if this project cannot be pursued successfully, then LOTH is in trouble as a *completely* naturalistic program: it quantifies over things, i.e. symbols, that have semantic properties that it doesn't know how to naturalize. But the hope is that LOTH is only a step forward in an overall project of naturalization of the mind. This is nothing but a perspicuous application of the divide-and-conquer strategy employed by naturalists in a difficult job.

But, the official line continues, LOTH has a lot to say about the second stage, the stage where the semantic contents are computed or assigned to complex (molecular) symbols on the basis of their combinatorial syntax or grammar together with whatever meanings atomic symbols are assumed to have in the first stage. This

[24] See, for instance, Fodor (1987, 1990a), Dretske (1981, 1988), Millikan (1984, 1993), Papineau (1987), Devitt (1996), Loar (1982a), Field (1972, 1978), Block (1986).

procedure is familiar from a Tarski-style[25] definition of truth conditions of *sentences*. Above, when I described the truth functional character of SL's connectives, I have already illustrated the idea in a very simple way: the truth-value of complex sentences are to be determined by the truth-values of the atomic sentences they contain together with the rules fixed by the truth-tables of the connectives occurring in the complex sentences. This process is similar but more complex in first-order languages, and even more so for natural languages — in fact, we don't have a completely working compositional semantics for the latter at the moment. So, *if* we have a semantic interpretation of atomic symbols (*if* we have symbols whose reference and extension are fixed at the first stage by whatever naturalistic mechanism turns out to govern it), *then* the combinatorial syntax will take over and effectively determine the semantic interpretation (truth-conditions) of the complex sentences they are constituents of. So officially LOTH would only contribute to a complete naturalization project if there is a naturalistic story at the atomic level.

Early Fodor (1975, 1978, 1978a, 1980), for instance, envisaged a science of psychology which, among other things, would reasonably set for itself the goal of discovering the combinatorial syntactic principles of LOT and the computational rules governing its operations, without worrying much about semantic matters, especially about how to fix the semantics of atomic symbols (he probably thought that this was not a job for LOTH). Similarly, Field (1978) is very explicit about the combinatorial rules for assigning truth-conditions to the sentences of the internal code. In fact, Field's major argument for LOTH is that, given a naturalistic causal theory of reference for atomic symbols, about which he is optimistic (Field 1972), it is the only naturalistic theory that has a chance of solving Brentano's puzzle. For the moment, this is not much more than a hope, but, according to the LOT theorist, it is a well-founded hope based on a number of theoretical and empirical assumptions and data. Furthermore, it is a framework defining a naturalistic research program in which there have been promising successes.[26]

As I said, this official and, in a way, least committal line has been overall the more standard way of conceiving LOTH's role in the project of naturalizing intentionality.

---

[25] Tarski (1956), Field (1972), Davidson (1984).

[26] Although I described the line above as official and presented it as requiring a compositional semantics, and although almost all the defenders of LOTH conceive of it in this way because they think that is what empirical facts about thought and language demand, nevertheless it is perhaps important to be pedantic about exactly what LOTH is minimally committed to. Minimally, it is *not* committed to regarding the internal code as having a compositional semantics, namely a semantics where the meaning of complex sentences are determined by the meanings of its constituents together with their syntax; this, in effect, requires that the atomic expressions always make (approximately) the same semantic contributions to the whole of which they are constituents (idioms excepted). But strictly speaking LOTH can live without having a strictly compositional semantics if it turns out that there are other ways of explaining those empirical facts about the mind to which I will come below. Admittedly, in such a case LOTH would lose some portion of its appeal and interest. But even if this scenario turns out to be the case, there are still a lot of facts for LOTH to explain. Having said this, however, I will simply forget it in what follows.

But some have gone beyond it and explored the ways in which the resources of LOTH can be exploited even in answering the first question (1) about the semantics of atomic symbols.

Now, there is a weak version of an answer to (1) on the part of LOTH and a strong version. On the weak version, LOTH may be untendentiously viewed as inevitably providing *some* of the resources in giving the ultimate naturalistic theory in naturalizing the meaning of atomic symbols. The basic idea is that whatever the ultimate naturalistic theory turns out to be true about atomic expressions, computation as conceived by LOTH will be part of it. For instance, it may be that, as with nomic covariation theories of meaning (Fodor 1987, 1990a; Dretske 1981), the meaning of an atomic predicate may consist in its potential to get tokened in the presence of (or, in causal response to) something that instantiates the property the predicate is said to express. A natural way of explicating this potential may partly but ultimately rely on certain computational principles the symbol may be subjected to within a LOT framework, or principles that in some sense govern the "behavior" of the symbol. Insofar as computation is naturalistically understood in the way LOTH proposes, a complete answer to the first question about the semantics of atomic symbols may plausibly involve an explicatory appeal to computation within a system of symbols. This is the weak version because it doesn't see LOTH as proposing a complete solution to the first question (1) above, but only helping it.

A strong version would have it that LOTH provides a *complete* naturalistic solution to both questions: given the resources of LOTH we don't need to look any further to meet Brentano's challenge. The basic idea lies in so-called functional or conceptual role semantics, according to which a concept is the concept it is precisely in virtue of the particular causal/functional potential it has in interacting with other concepts. The intuitive idea is that each concept has a certain peculiar set of epistemic/semantic relations or liaisons to other concepts. We can conceive of this set as determining a certain "conceptual role" for each concept. We can then take these roles to determine the semantic identity of concepts: concepts are the concepts they are because they have the conceptual roles they have; that is to say, among other things, concepts represent whatever they do precisely in virtue of these roles. The idea then is to reduce each *conceptual* role to *causal/functional* role of atomic symbols (now conceived as primitive concepts on LOTH), and then use the resources of LOTH to reduce it in turn to *computational* role. Since computation is naturalistically well-defined, the argument goes, and since causal interactions between thoughts and concepts can be understood completely in terms of computation, we can completely naturalize intentionality if we can successfully treat meanings as arising out of thoughts/concepts' internal interactions with each other. In other words, the strong version of LOTH would claim that atomic symbols in LOT have the content they do in virtue of their potential for causal interactions with other tokens, and cashing out this potential in mechanical/naturalistic terms is what, among other things, LOTH is for. LOTH then comes as a naturalistic rescuer for conceptual (and even functional) role semantics.

It is not clear whether any one holds this strong version of LOTH in this rather naive form. But certainly some people have elaborated the basic idea in quite subtle

ways, for which Cummins (1989:Chp.8) is perhaps the best example. (But also see Block 1986 and Field 1978.) But even in the best hands, the proposal turns out to be very problematic and full of difficulties nobody seems to know how to straighten out. In fact, some of the most ardent critics of taking LOTH as incorporating a functional role semantics turn out to be some of the most ardent defenders of LOTH understood in a weak, non-committal sense we have explored above — see Fodor (1987: Chp.3), Fodor and Lepore (1991), Fodor's attack (1978b) on AI's way of doing procedural semantics is also relevant here. Haugeland (1981), Searle (1980, 1984), and Putnam (1988) quite explicitly take LOTH to involve a program for providing a complete semantic account of mental symbols, which they then attack accordingly.[27]

As indicated previously, LOTH is almost completely silent about consciousness and the problem of qualia, the third mystery in Fodor's list in the quote above. But the naturalist's hope is that this problem too will be solved, if not by LOTH, then by something else. On the other hand, it is important to emphasize that LOTH is neutral about the naturalizability of consciousness / qualia. If it turns out that qualia cannot be naturalized, this would by no means show that LOTH is false or defective in some way. In fact, there are people who *seem* to think that LOTH may well turn out to be true even though qualia can perhaps not be naturalized (e.g., Block 1980, Chalmers 1996, McGinn 1991).

Finally, it should be emphasized that LOTH has no particular commitment to every symbolic activity's being conscious. Conscious thoughts and thinking may be the tip of a computational iceberg. On the other hand, to the extent to which thought and thinking are conscious, to that extent LOTH can be viewed as providing the necessary means for a naturalistic account — perhaps in the form of higher order thoughts about first-order thoughts cashed out in a LOT framework, for an elaboration see Rosenthal (1997) and Lycan (1997).

8 Objections to LOTH

There have been numerous arguments against LOTH. Some of them are directed more specifically against the RTM (A), some against the functionalist nature of LOTH, (C). Here I will concentrate only on those arguments specifically targeting (B) — the most controversial component of LOTH.

8.1 Regress Arguments against the LOTH

These arguments rely on the explanations offered by LOTH defenders for certain aspects of natural languages. In particular, many LOT theorists advert to LOTH to explain (1) how natural languages are learned, (2) how natural languages are understood, or (3) how the utterances in such languages can be meaningful. For instance, according to Fodor (1975), natural languages are learned by forming and

[27] For fairness I should add that Searle's and Haugeland's criticisms were directed against AI community at large, and there, it was more or less common to conceive the computational model of mind as potentially involving a complete solution to semantic worries among others. Thus, Haugeland termed his target 'GOFAI' (the Good Old Fashion Artificial Intelligence). Similarly, Searle's famous Chinese Room Argument was directed against what he called 'Strong AI'.

confirming hypotheses about the translation of natural language sentences into Mentalese such as: 'Snow is white' is true in English if and only if *P*, where '*P*' is a sentence in one's LOT. But to be able to do that, one needs a representational medium in which to form and confirm hypotheses. The LOT is such a medium. Again, natural languages are understood because, roughly, such an understanding consists in translating their sentences into one's Mentalese. Similarly, natural language utterances are meaningful in virtue of the meanings of corresponding Mentalese sentences.

The basic complaint is that in each of these cases, either the explanations generate a regress because the same sort of explanations ought to be given for how the LOT is learned, understood or can be meaningful, or else they are gratuitous because if a successful explanation can be given for LOT that does not generate a regress then it could and ought to be given for the natural language phenomena without introducing a LOT (see, e.g. Blackburn 1984). Fodor's response in (1975) is (1) that LOT is not learned, it's innate, (2) that it's understood in a different sense than the sense involved in natural language comprehension, and (3) that LOT sentences acquire their meanings not in virtue of another meaningful language but in a completely different way, perhaps by standing in some sort of causal relation to what they represent (see above) or by having certain computational profiles. For many who have a Wittgensteinian bent, these replies are not likely to be very convincing. But then the issues here tend to concern RTM rather than (B).

Laurence and Margolis (1997) point out that the regress arguments depend on the assumption that LOTH is introduced only to explain (1)–(3). If it can be shown that there are lots of other empirical phenomena for which the LOTH provides good explanations, then the regress arguments fail because LOTH then would not be gratuitous. In fact, as we'll see below, there are plenty of such phenomena. But still it is important to realize that the sort of explanations proposed for the understanding of one's LOT (computational use / activity of LOT sentences with certain meanings) and how LOT sentences can be meaningful (computational roles and / or nomic relations with the world) cannot be given for (1)–(3): it's unclear, for example, what it would be like to give a computational role and / or nomic relation account for the meanings of natural language utterances.

8.2 Propositional Attitudes without Explicit Representations (cf. Fodor 1987: 21–3)

Dennett in his review of Fodor's (1975) has raised the following objection:

> In a recent conversation with the designer of a chess-playing program I heard the following criticism of a rival program: "it thinks it should get its queen out early." This ascribes a propositional attitude to the program in a very useful and predictive way, for as the designer went on to say, one can usefully count on chasing that queen around the board. But for all the many levels of explicit representation to be found in that program, nowhere is anything roughly synonymous with "I should get my queen out early" explicitly tokened. The level of analysis to which the designer's remark belongs describes features of the program that are, in an entirely innocent way,

> emergent properties of the computational processes that have "engineering reality." I see no reason to believe that the relation between belief-talk and psychological talk will be any more direct. (Dennett 1981a: 107)

The objection, as Fodor (1987: 22) points out, isn't that the program has a *dispositional*, or *potential*, belief that it will get its queen out early. Rather, the program actually operates on this belief. There appear to be lots of other examples: e.g. in reasoning we pretty often follow certain inference rules like modus ponens, disjunctive syllogism, etc. without necessarily explicitly representing them.

The standard reply to such objections is to draw a distinction between rules on the basis of which Mentalese data-structures are manipulated, and the data-structures themselves (intuitively, the program / data distinction). LOTH is not committed to every rule's being explicitly represented. In fact, as a point of nomological fact, in a computational device not every rule can be explicitly represented: some *have to* be hard-wired and, thus, implicit in this sense. In other words, LOTH permits but doesn't require that rules be explicitly represented. On the other hand, data structures *have to* be explicitly represented: it is these that are manipulated formally by the rules. No causal manipulation is possible without explicit tokening of these structures. According to Fodor, if a propositional attitude is an actual episode in one's reasoning that plays a causal role, then LOTH is committed to explicit representation of its content, which is as per (A2 and B2) causally implicated in the physical process realizing that reasoning. Dispositional propositional attitudes can then be accounted for in terms of an appropriate principle of inferential closure of explicitly represented propositional attitudes (cf. Lycan 1986).

Dennett's chess program certainly involves explicit representations of the chess board, the pieces, etc. and perhaps some of the rules. Which rules are implicit and which are explicit depend on the empirical details of the program. Pointing to the fact that there may be some rules that are emergent out of the implementation of explicit rules and data-structures does not suffice to undermine LOTH.

8.3 Explicit Representations without Propositional Attitudes (cf. Fodor 1987: 23–6)

In any sufficiently complex computational system, there are bound to be many symbol manipulations with no obviously corresponding description at the level of propositional attitudes. For instance, when a multiplication program is run through a standard conventional computer, the steps of the program are translated into the computer's machine language and executed there, but at this level the operations apply to 1's and 0's with no obvious way to map them onto the original numbers to be multiplied or to the multiplication operation. So, it seems that at the levels that, according to Dennett, have engineering reality there are plenty of explicit tokenings of representations with appropriate operations that don't correspond to anything like the propositional attitudes of folk psychology. In other words, there is plenty of symbolic activity which it would be wrong to say a *person* engages in. Rather, they are done by the person's subpersonal computational *components* as opposed to the person. How to rule out such cases?

They are ruled out by an appropriate reading of (A1) and (B1): (A1) says that the person herself must stand in an appropriate computational relation to a Mentalese sentence, which, as per (B1), has a suitable syntax and semantics. Only then, will the sentence constitute the person's having a propositional attitude. Not all explicit symbols in one's LOT will satisfy this. In other words, not every computational routine will correspond to a processing appropriately described as storage in, e.g., the "belief-box." Furthermore, as pointed out by Fodor (1987), LOTH would vindicate the common sense view of propositional attitudes if they turn out to be computational relations to Mentalese sentences. It may not be further required that every explicit representation correspond to a propositional attitude.

There have been many other objections to LOTH, in recent years raised especially by connectionists: that LOT systems cannot handle certain cognitive tasks like perceptual pattern recognition, that they are too brittle and not sufficiently damage resistant, that they don't exhibit graceful degradation when physically damaged or as a response to noisy or degraded input, that they are too rigid, deterministic, so are not well-suited for modeling humans' capacity to satisfy multiple soft-constraints so gracefully, that they are not biologically realistic, and so on. For useful discussions of these and many similar objections, see Rumelhart, McClelland and the PDP Research Group (1986), Fodor and Pylyshysn (1988), Bechtel and Abrahamsen (1991), Horgan and Tienson (1996), and McLaughlin and Warfield (forthcoming).

9 Arguments for LOTH

We have already seen two major arguments, perhaps the historically most important ones, for LOTH: First, we have seen in § 2 that if LOTH is true then all the essential features of the common sense conception of propositional attitudes will be explicated in a naturalistic framework which is likely to be co-opted by scientific cognitive psychology, thus vindicating folk psychology. Second, we discussed in § 7 that, if true, LOTH would solve one of the mysteries about thinking minds: how is thinking (as characterized above) possible? How is rationality mechanically possible? Then we have also seen a third argument that LOTH would partially contribute to the project of naturalizing intentionality by offering an account of how the semantic properties of whole attitudes are fixed on the basis of their atomic constituents.

But there have been many other arguments for LOTH — in fact, a whole range of different kinds of arguments. Some are more empirically motivated, others more speculative, still others are more relevant to the defense of RTM rather than LOTH per se, some are more technically powerful than others, etc. In this section, I will try to describe only those arguments that have been historically more influential and controversial.

9.1 Argument from Contemporary Cognitive Psychology

When Fodor first formulated LOTH with significant elaboration in his (1975), he introduced his major argument for it along with its initial formulation in the first chapter. It was basically this: our best scientific theories and models of different aspects of higher cognition assume a framework that requires a

computational/representational medium for them to be true. More specifically, he analyzes the basic form of the information processing models developed to account for three types of cognitive phenomena: *perception* as the fixation of perceptual beliefs, *concept learning* as hypothesis formation and confirmation, and *decision making* as a form of representing and evaluating the consequences of possible actions carried out by the agent in a situation with a preordered set of preferences. He rightly points out that all these models treat mental processes as computational processes defined over representations. Then he draws what seems to be the obvious conclusion: if these models are right in at least treating mental processes as computational, even if not in detail, then there must be a LOT over which they are defined, hence LOTH.

In Fodor's (1975), the arguments for different aspects of LOTH are diffused and the emphasis, with the book's slogan "no computation without representation," is put on the RTM rather than on (B) or (C). But all the elements are surely there.

9.2 Argument from the Productivity of Thought

People seem to be capable of entertaining an indefinite number of thoughts, at least in principle, although they *in fact* entertain only a finite number of them, of course. Indeed adults who speak a natural language are capable of understanding sentences they have never heard uttered before. Here is one: there is a big lake of melted gold on the dark side of the moon. I bet that you never heard this sentence before, and yet, you have no difficulty in understanding it: it is one you're in fact likely to believe false. But this sentence was arbitrary, there are infinitely many such sentences I can in principle utter and you can in principle *understand*. But understanding a sentence is to entertain the thought/proposition it expresses. So there are in principle infinitely many thoughts you are capable of entertaining. This is sometimes expressed by saying that we have an unbounded *competence* in entertaining different thoughts, even though we have a bounded *performance*. But this unbounded capacity must be achieved by finite means. For instance, storing an infinite number of representations in our heads is out of the question: we are finite beings. If human cognitive capacities (capacities to entertain an unbounded number of thoughts, or to have attitudes towards an unbounded number of propositions) are productive in this sense, how is this to be explained on the basis of finitary resources?

The explanation LOTH offers is straightforward: postulate a representational system that satisfies at least (B1). Indeed, recursion is the only known way to produce an infinite number of symbols from a finite base. In fact, given LOTH, productivity of thought as a competence mechanism seems to be guaranteed.[28]

9.3 Argument from the Systematicity and Compositionality of Thought

Systematicity of thought consists in the empirical fact that the ability to entertain certain thoughts is intrinsically connected to the ability to entertain certain others.

[28] See Fodor (1985, 1987), Fodor and Pylyshyn (1988) for an elaborate presentation of this argument for LOTH.

Which ones? Thoughts that are related in a certain way. In what way? There is a certain initial difficulty in answering such questions. I think, partly because of this, Fodor (1987) and Fodor and Pylyshyn (1988), who are the original defenders of this kind of argument, first argue for the systematicity of language production and understanding: the ability to produce / understand certain sentences is intrinsically connected to the ability to produce / understand certain others. Given that a mature speaker is able to produce / understand a certain sentence in her native language, by psychological law, there always appear to be a cluster of other sentences that she is able to produce / understand. For instance, you don't seem to find speakers who know how to express in their native language the fact that John loves the girl but not the fact that the girl loves John. This is apparently so, moreover, for expressions of any n-place relation.

Fodor and Pylyshyn bring out the force of this psychological fact by comparing learning languages the way we actually do with learning a language by memorizing a huge phrase book. In the phrase book model, there is nothing to prevent someone learning how to say 'John loves the girl' without learning how to say 'the girl loves John.' In fact, that is exactly the way some information booklets prepared for tourists help them to cope with their new social environment. You might, for example, learn from a phrase book how to say 'I'd like to have a cup of coffee with sugar and milk' in Turkish without knowing how to say / understand absolutely anything else in Turkish. In other words, the phrase book model of learning a language allows arbitrarily punctate linguistic capabilities. In contrast, a speaker's knowledge of her native language is not punctate, it is *systematic*. Accordingly, you do not find, by nomological necessity, native speakers whose linguistic capacities are punctate.

How is this empirical truth, in fact, a law-like generalization to be explained? Obviously if this is a general nomological fact, then learning one's native language cannot be modeled on the phrase book model. What is the alternative? The alternative is well known. Native speakers master the grammar and vocabulary of their language. But this is just to say that sentences are not atomic, but have syntactic constituent structure. If you have a vocabulary, the grammar tells you how to combine *systematically* the words into sentences. Hence, in this way, if you know how to construct a particular sentence out of certain words, you automatically know how to construct many others. If you view all sentences as atomic, then, as Fodor and Pylyshyn say, the systematicity of language production / understanding is a mystery, but if you acknowledge that sentences have syntactic constituent structure, systematicity of linguistic capacities is what you automatically get; it is guaranteed. This is the orthodox explanation of linguistic systematicity.

From here, according to Fodor and Pylyshyn, establishing the systematicity of thought as a nomological fact is one step away. If it is a law that the ability to understand a sentence is systematically connected to the ability to understand many others, then it is similarly a law that the ability to think a thought is systematically connected to the ability to think many others. For to understand a sentence is just to think the thought / proposition it expresses. Since, according to RTM, to think a certain thought is just to token a representation in the head that expresses the relevant proposition, the ability to token certain representations is systematically

connected to the ability to token certain others. But then, this fact needs an adequate explanation too. The classical explanation LOTH offers is to postulate a system of representations with combinatorial syntax exactly as in the case of the explanation of the linguistic systematicity. This is what (B1) offers.[29] This seems to be the only explanation that does not make the systematicity of thought a miracle, and thus argues for the LOT hypothesis.

However, thought is not only systematic but also compositional: systematically connected thoughts are also always semantically related in such a way that the thoughts so related seem to be composed out of the same semantic elements. For instance, the ability to think 'John loves the girl' is connected to the ability to think 'the girl loves John' but not to, say, 'protons are made up of quarks' or to '2+2=4.' Why is this so? The answer LOTH gives is to postulate a combinatorial semantics in addition to a combinatorial syntax, where an atomic constituent of a mental sentence makes (approximately) the same semantic contribution to any complex mental expression in which it occurs. This is what Fodor and Pylyshyn call 'the principle of compositionality'.[30]

In brief, it is an argument for LOTH that it offers a cogent and principled solution to the systematicity and compositionality of cognitive capacities by postulating a system of representations that has a combinatorial syntax *and* semantics, i.e., a system of representations that satisfies at least (B1).

9.4 Argument from the Systematicity of Thinking (Inferential Coherence)

Systematicity of thought does not seem to be restricted solely to the systematic ability to entertain certain *thoughts*. If the system of mental representations does have a combinatorial syntax, then there is a set of rules, syntactic formation rules, so to speak, that govern the construction of well-formed expressions in the system. It is this fact, (B1) that guarantees that if you can form a mental sentence on the basis of certain rules, then you can also form many others on the basis of the same rules. The rules of combinatorial syntax determine the syntactic or formal structure of complex mental representations. This is the *formative* (or, *formational*) aspect of systematicity. But inferential *thought processes* seem to be systematic too: the ability to make certain *inferences* is intrinsically connected to the ability to make certain many others. For instance, you do not find minds that can infer '*A*' from '*A&B*' but

---

[29] It should be noted however that (B1) is a meta-architectural condition that needs to be satisfied by any *particular* grammar for Mentalese, just as an analogue for (B1) is a condition upon the *specific* grammar of all systematic languages (see below § 11).

[30] It is somewhat confusing that Fodor and Pylyshyn called this *empirical cognitive regularity* "compositionality" of cognitive capacities. In particular, the empirical phenomenon — i.e., the *fact* that systematically connected thoughts are also always semantically related or semantically close to each other — that needs to be explained is explained by LOT theorists by what is also called semantic compositionality: namely, the semantic value of a complex expression is a function of the semantic value of its atomic constituents such that each atomic constituent makes approximately the same semantic contribution to the context in which it occurs. This is what the postulation of a combinatorial semantics in conjunction with a combinatorial syntax buys for LOT-theorists in adequately explaining the empirical regularity in question. See Fodor and Pylyshyn (1988: 41–5).

cannot infer '*C*' from '*A&B&C*.' It seems to be a psychological fact that inferential capacities come in clusters that are homogeneous in certain aspects. How is this fact (i.e., the *inferential* or *transformational* systematicity) to be explained?

As we have seen, the explanation LOTH offers depends on the exploitation of the notion of logical form or syntactic structure determined by the combinatorial syntax postulated for the representational system. The combinatorial syntax not only gives us a criterion of well-formedness for mental expressions, but it also defines the logical form or syntactic structure for each well-formed expression. The classical solution to inferential systematicity is to make the mental operations on representations sensitive to their form or structure, i.e. to insist on (B2). Since, from a syntactic view point, similarly formed expressions will have similar forms, it is possible to define a single operation which will apply to only certain expressions that have a certain form, say, only to conjunctions, or disjunctions. This allows the LOT theorist to give homogeneous explanations of what appear to be homogeneous classes of inferential capacities. This is one of the greatest virtues of LOTH, hence provides an argument for it.

The solution LOTH offers for what I called the problem of thinking, above, is connected to the argument here because the two phenomena are connected in a deep way. Thinking requires that the logico-semantic properties of a particular thought process (say, inferring that John is happy from knowing that if John is at the beach then John is happy and coming to realize that John is indeed at the beach) be somehow causally implicated in the process. The systematicity of inferential thought processes then is based on the observation that if the agent is capable of making *that* particular inference, then she is capable of making many other somehow *similarly organized* inferences. But the idea of similar organization in this context seems to demand some sort of classification of thoughts independently of their *particular* content. But what can the basis of such a classification be? The only basis seems to be the logico-syntactic properties of thoughts, their form. Although it feels a little uneasy to talk about syntactic properties of thoughts common-sensically understood, it seems that they are forced upon us by the very attempt to understand their semantic properties: how, for instance, could we explain the semantic content of the thought that if John is at the beach then he is happy without somehow appealing to its being a *conditional*? This is the point of contact between the two phenomena. Especially when the demands of naturalism are added to this picture, inferring a LOT (= a representational system satisfying B) realized in the brain becomes indeed almost irresistible. Indeed Rey (1995) doesn't resist and claims that, given the above observations, LOTH can be established on the basis of arguments that are not "merely empirical." I leave it to the reader to evaluate whether mere critical reflection on our concepts of thought and thinking could, all by itself, establish LOTH.[31]

The last three arguments have played a central role in the recent debate between the defenders of LOTH, sometimes called classicists, and connectionists, which we will take up later. They have therefore assumed a well-publicized and important role in

[31] For a prioristic arguments of this sort, see also Lycan (1993) and Davies (1989, 1991).

discussions of LOTH. But there are many other arguments some of which appear to be equally important at least from a more philosophical perspective, and thus deserve close attention. The following argument, in particular, draws upon a set of closely related problems that have traditionally been the concern of philosophers more than psychologists.

*9.5 *Argument from the Opacity of Propositional Attitudes*

I mean to use the term 'opacity' rather loosely. In particular, I want to consider three phenomena that are somewhat importantly related to a standard understanding of opacity of attitudes. After going over them and stating what problems they pose for any naturalistic theory of propositional attitudes like LOTH, I will show how LOTH proposes a framework within which they can be handled in a distinctive way.[32] It will become clear that the distinctive and promising nature of this framework LOTH offers for explaining propositional attitudes is in fact a powerful argument for it. Again, like the previous ones, this argument takes the form of an argument from explanatory power. The three phenomena may be labeled as (I) Frege-Kripke cases, (II) the hyper-opacity of the attitudes, and (III) Perry cases (the problem of the essential indexical).

(I) *Frege-Kripke Cases*. It is certainly not foreign to common sense that the following inference is not generally valid:

1. John believes that Bob Dylan is the conscience of American youth.

2. Bob Dylan = Robert Zimmerman.

3. Therefore, John believes that Zimmerman is the conscience of American youth.

Common sense typically regards the matter of whether (3) may be validly concluded on the basis of (1)–(2) as depending on whether John believes that Dylan = Zimmerman. Moreover, as we have seen before, it is only when John believes this identity that we expect John to come to have the belief expressed in (3).

---

[32] What follows is nothing but a very sketchy account of some of the well-known puzzles about propositional attitudes and their ascription. Even more sketchy is the account a LOT-theorist might give about how to start providing a solution to these puzzles by using the resources of LOTH and its treatment of propositional attitudes. What I would like to emphasize here is that the account I'll sketch below is by no means a proposal about the semantics of propositional attitude ascription. The literature on the latter has grown immensely in the last twenty years or so, and has become very sophisticated and technical. One increasingly prominent trend in the literature is the attention given to mental states (qua realized in the brain of the agents to whom the attitudes are ascribed). Indeed, it seems to have become clear in the beginning of the eighties that a successful semantics for propositional attitude ascriptions would inevitably advert, in the phrase of Perry (1979), to the "belief-states" of agents as distinct from propositions they are said to be belief-related to. To many, LOTH has seemed a useful framework in which to cash out these "states." For useful elaborations in more or less this direction, see, among others, Boër and Lycan (1986), Fodor (1989), Crimmins and Perry (1989), Richard (1990), Crimmins (1992), Boër (1995), Zalta (forthcoming). For criticism of Richard and Crimmins, see Saul (1993). Perry and Israel's (1991) discussion is also very useful in this context.

Since Russell and Frege, it has been abundantly observed by philosophers that *belief contexts* are referentially opaque in that substituting a term with a co-referring or a co-extensional expression in a complement sentence embedded in a belief sentence may not generally preserve the truth-value of the belief sentence in which it occurs (similarly for any propositional attitude sentence). Similarly, existential generalizations may fail in belief contexts: from the fact that little David believes that Santa Clause came down the chimney last night, we cannot validly infer that there is someone who came down the chimney. This feature of belief sentences has produced quite a bit of amazement and philosophical embarrassment. Accordingly, a lot of literature has been produced about how to best treat them, in particular, about how to give a semantics for belief sentences.

Perhaps the earliest most explicit articulation of the phenomenon was by Frege (1892/1949) who observed that sentences of the form 'a=a' and 'a=b', where both 'a' and 'b' refer to the same thing, radically differ in their epistemic status: although the truth of the former, he said, is a priori, the truth of the latter is only a posteriori knowable. Similarly for their cognitive significance for an agent who understands the sentences, hence grasps their meanings: the former is a trivial self-identity, but the latter may constitute an important empirical discovery, hence is not trivial at all. How is this possible given that both 'a' and 'b' refer to the same entity, and thus both sentences seem to have exactly the same truth conditions? The puzzle becomes especially acute if the relevant referring expressions are proper names, for which postulating Fregean senses as distinct from reference (as Frege did in order to explain opacity) seems to be less plausible.

As is well known, Kripke (1979) generated interesting puzzles about our practices of attributing beliefs and other attitudes on the assumption that proper names have no senses and refer rigidly. Kripke observes that someone, say John, who is ignorant of the fact that Dylan = Zimmerman, can have the following beliefs simultaneously:

- that Dylan is politically cool.
- that Zimmerman is not politically cool.

Thus he can conjoin and express them by saying "I believe that Dylan is politically cool but Zimmerman isn't." Kripke observes the following: (a) anyone who is a competent speaker of English, by sincerely asserting the sentence 'Dylan is politically cool but Zimmerman isn't', expresses the belief that Dylan is politically cool but Zimmerman isn't; (b) thus if we take the objects of beliefs as propositions, John believes the proposition that Dylan is politically cool but Zimmerman isn't; (c) assuming what seems to be the dominant and plausible view that proper names are rigid and refer to their referents directly (i.e., without the mediation of Fregean senses), the proposition believed by John is false in every possible world, hence, by this account, is a contradiction; (d) yet, along with common sense, we may assume that John is not irrational, he would certainly not explicitly ascent to any contradiction (he may be guilty of ignorance but not of irrationality). According to Kripke, observations (a) though (d) are in great tension with each other, or form an inconsistent set of claims at worst. Moreover, common sense seems to countenance

each of them separately (excepting perhaps (c) about which we may regard the folk as neutral, if not already endorsing). Kripke also asks the following question as requiring a definite answer: "Does John believe that Dylan is politically cool or does he not?" (or, "Does John believe that Zimmerman is not politically cool, or does he not?"). This is the puzzle of belief generated by Kripke, assuming the rigidity of proper names.[33]

(II) *Hyper-opacity*. Even more interestingly, it is clear that we sometimes distinguish even among *logically equivalent beliefs*. For instance, we certainly process logically more complex propositions with more difficulty: e.g., people usually have more difficulty in processing propositions of the form, '~(~pv~q)', compared to those with its simpler equivalent form, '(p&q)'. There are well-established psychological experiments and arguments that show that most people are susceptible to make mistakes depending on how the logical forms of the propositions are represented describing the problem for which the subjects are asked to produce a solution.[34]

There also seem to be important psychological differences between synonymous thoughts: e.g., thinking that this is a triangle and thinking that this is a closed three (straight-)sided figure creating three interior angles whose sum equals to 180°. Indeed, according to Rey (1997:254-5), this sort of case is paradigmatic about how we gain deep theoretic/conceptual insight into the same phenomena. Examples can certainly be multiplied even with the most banal cases like the difference between 'bachelor' and 'adult unmarried man.' There seem to be psychological contexts we may indeed want to distinguish between them. Some of Fodor's examples may illustrate:

> it seems to me (as it seemed to Mates' 1952) that it is possible for me to doubt (/deny) that everybody who believes that Oedipus is a bachelor believes that Oedipus is an unmarried man even though I don't doubt (/deny) that everybody who believes that Oedipus is a bachelor believes that Oedipus is a bachelor. At a minimum, it's surely possible for it *to seem to me* that [it's possible for me to doubt (/deny) that everybody who believes that Oedipus is

[33] Of course, a Fregean would not probably be much moved by this puzzle over and above the puzzles the standard opacity cases generate. On the contrary, she would take it to show that names are not rigid but refer through mediating senses. Following Lycan (1981), I have chosen to present Kripke's puzzle by this example rather than his standard 'Londres'/'London' and Paderewski examples in order to tie it with the standard opacity of belief contexts introduced above. But Kripke, of course, thinks that for a Fregean it is much more difficult to appeal to senses in the case of his examples about the bilingual Frenchman, Peter, on the one hand, and Paderewski the musician/politician on the other, about which he seems right to me. It seems indeed less plausible to try to explain these examples by arguing that 'Londres' and 'London' differ in senses when uttered by Pierre in the relevant contexts, and similarly for 'Paderewski' in two different apparently contradictory uses. These examples and those of Mates (1952), according to Fodor (1989), show that belief contexts distinguish even between those expressions with the same sense and thus sometimes require taxonomies that are even more fine-grained than senses. But as we will see, a LOT-theorist, as such, need not take sides between Fregean and Kripkean approaches in this regard.

[34] E.g., Wason and Johnson-Laird (1972), Wason (1981), Kahneman, Slovic and Tversky (1982).

> a bachelor believes that Oedipus is an unmarried man] even though it doesn't seem to me that [it's possible for me to doubt (/ deny) that everybody who believes that Oedipus is a bachelor believes that Oedipus is a bachelor]. For as a matter of fact, it does seem to me that it seems to me that all of this is so; and I would seem to be in about as good a position as anyone can be to say how things seem to me to be, *nicht wahr*? So maybe substitution of synonyms salva veritate fails in the context 'it seems to me that ... ', or in iterations of that context. (1989: 164, brackets and emphases in the original.)

If this set of observations about the psychological significance of representing logically equivalent propositions and synonymous concepts is right, then propositional attitude contexts are not only *intensional*, but, in the words of Rey (1997: 254-5), are also *hyperintensional*. We may perhaps equally say that they are hyper-opaque: Attitude contexts fail not only to be referentially transparent, they also fail to be transparent with respect to synonyms and logical equivalents, i.e., they don't always allow for substitution of synonyms or logically equivalent expressions salva veritate. How is this phenomenon to be explained?

(III) *The problem of the essential indexical.* John Perry (1979) argues that there are certain kinds of belief reports containing indexicals (like 'I', 'he', 'now', etc.) whose explanatory force cannot be preserved if the indexicals are replaced by co-referring expressions. According to Perry, this casts doubt upon the traditional dyadic view of beliefs as relations between agents and propositions, in whatever ways propositions are understood. He proposes that the traditional idea should be rejected in favor a triadic view where the three relata are agents, propositions believed and belief states. He wants to remain neutral about what belief states are, and how they can be accounted for by a naturalistic philosophy of psychology. As we will see, LOTH can be viewed as an improvement over his suggestion in just that direction. But first, let us see the phenomenon in an example. I will use Perry's:

> I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with a torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. But I seemed unable to catch up. Finally it dawned on me. I was the shopper I was trying to catch. (1979: 33)

Then, we may continue, I stopped and rearranged the torn sack in my cart. There is certainly a change in belief here that explains the change in my action. What is that change? Before I realized I was the culprit, I believed that

(1) the shopper with a torn sack of sugar was making a mess.

Then I came to believe that

(2) I was making a mess

after I realized that I was the shopper with a torn sack. Perry argues that no substitution of 'I' in (2) with any co-designative term, *t*, will account for the change in my behavior unless I also believe that I am identical to what *t* refers to. Hence

there is no way of getting rid of the indexical, 'I', in the belief report involved in the explanation of the change in my behavior. Perry calls such indexicals *essential indexicals*.

To see this posing a real problem for the dyadic view, let us ask: What propositions are expressed by (1) and (2)? The answer depends on how we conceive of propositions. Perry argues plausibly that a Fregean conception of propositions is inadequate to accommodate the problem of the essential indexical, where propositions are roughly conceived to consists of concepts (~ senses) that pick up referents that "fit" them, rather than the referents themselves. (For the full discussion, the reader should consult Perry's discussion in his (1979).)

Let us conceive of propositions more as Russell did. I will follow Kaplan's refinement (1989), and consider the so-called *singular propositions*. Singular propositions consist, again roughly, of an object or an ordered set of objects and a property or a relation. Conceived this way, the proposition I believed when I believed (2) is

(3) <Murat, the property of making a mess>.

But this seems to be the very same proposition as when I believed (1). There are, of course, complications involving the fact that (1) contains a definite description 'The shopper with torn sack'. In order to avoid them, and to sharpen the problem of the essential indexical, let us note that even if I believed that

(4) Murat Aydede is making a mess,

or,

(5) *he* is making a mess,

where I utter 'he' ostending a mirror image of myself down the aisle without realizing that I am referring to myself in the mirror, the explanation of the change in my behavior still essentially depends on identifying *myself* as Murat Aydede, or as the referent of 'he'. (Just try to imagine how such identifications may be broken.) Otherwise, there would be no change in my behavior bending over my cart, except accidentally. But what proposition(s) do (4) and (5) express? This time it seems more plausible to identify the proposition expressed by them as (3), the very same one when I believed (2). But if so, appealing to propositions in this manner won't explain the causal difference that believing in the way (2), (4), and (5) describe makes to my behavior: they all have the same propositional object. There is something special about believing the proposition (3) *in the way (2) describes,* not captured by (1), (4) and (5) or, for that matter, by any sentence not containing 'I'. But what could the difference be if it is not the proposition believed?[35]

[35] Perry's discussion is rich, and covers a lot of terrain about the different ways in which a proponent of a dyadic view can attempt to explain Perry's examples by offering different conceptions of propositions. Perry argues that none of them is successful. I have certainly not discussed all the initially plausible variations on how to conceive propositions, but instead focused on the phenomenon itself. This is enough for our purposes here, but the reader should consult Perry's (1979) article for more detail. His (1977) is also helpful. Perry himself

Argument for LOTH from the explanation of (I)-(III). It is now time to see how LOTH proposes a framework within which the three phenomena just described can be successfully handled. Given our discussion of LOTH so far, the general outline of their explanation can be anticipated. As we have seen, LOTH proposes an analysis of propositional attitudes as triadic relations. To repeat, if we focus on belief, the explication of '*S* believes that *P*' involves postulation of a triadic relation between the agent *S*, a syntactically structured mental sentence #*P*# physically realized in *S*'s brain with the computational role appropriate for beliefs, and a proposition *P* that #*P*# expresses, or has as its semantic content. Obviously, this gives LOTH more degrees of freedom to maneuver than the ones the traditional dyadic view has. The details of how LOTH can accommodate the phenomena described in (I)–(III) are complicated and vary among LOT theorists depending on their views of semantics in general and of propositions in particular (some of them are neo-Fregean, some are not, some are in between, etc.), but let me sketch the skeleton of the general solution LOTH gives.

Let us start with the standard opacity cases: John's believing that Dylan is cool is not the same thing as his believing that Zimmerman is cool in that (i) he may have one of the beliefs without having the other, or (ii) he may have both, but he may still be said to have two different beliefs if he doesn't believe (lacks the belief) that Dylan = Zimmerman, or, even more radically, (iii) he may have the one while explicitly denying the proposition involved in the other, despite the fact that Dylan = Zimmerman. According to LOTH the explanation may roughly go something like this:

(i) John has a token of the mental sentence #Dylan is cool# in his belief box (see § 2). This sentence expresses the proposition:

(6) <Dylan, being cool>

But John doesn't have a token of the mental sentence #Zimmerman is cool# in his belief box, which would express the same proposition (6) (we are assuming that proper names don't have Fregean senses for the moment, but this is not essential, we may assume that the two mental sentence tokens have exactly the same truth-conditions).

(ii) John has both sentence tokens in his belief box, but he doesn't have a token of #Dylan = Zimmerman#

(iii) In his belief-box John has a token of #Dylan is cool#, and a token of #Zimmerman is not cool# expressing the proposition

(7) <Dylan, not being cool>

In this case we may assume that John is rational and (at least) lacks a token of #Dylan = Zimmerman#.

---

acknowledges that he was inspired for the articulation of the puzzle by Castenada's writings, and his advice about which direction the solution lies in was influenced by Kaplan's distinction between the content and character of an indexical.

In all the three cases, LOTH, of course, assumes that the two token mental sentences, #Dylan is cool# and #Zimmerman is cool#, belong to different "syntactic" types. On LOTH, John can be saved from an accusation of irrationality despite the fact that, as in the case of Kripke's example, he can be said to believe a contradictory proposition, namely the conjunction of (6) and (7) if he tokens the mental sentence #Dylan is cool but Zimmerman isn't# in his belief box. The trick is to define rationality over sentences (along with their propositional content), not just over propositions believed. But then, John is not irrational. Intuitively, he would be irrational if he simultaneously harbored in his belief-box two sentence tokens of the forms '#a is F#' and '#a is not F#' with the corresponding contradictory propositions assigned to them as their semantic content respectively. (Note that this is in harmony with a proof-theoretic notion of contradiction.) But this is patently not the case with John: the two sentences with contradictory propositions he has in his box don't fit into that pattern: #Dylan# and #Zimmerman# are of different symbol types despite their being co-referential. Below, we will take up the issue of what makes two tokens belong to the same "syntactic" type. For the moment, let us extend this sketchy solution to (II) and (III).

That there may be psychological differences between logically equivalent beliefs poses no special problems for LOTH. The solution lies in the appeal to the syntactic form of the mental sentences underlying the beliefs with the same truth-conditions. Since the processing of these sentences is sensitive to their logical form, LOTH in fact predicts the existence of potential computational differences in processing logically equivalent sentences with the same propositional content (where propositions may be understood in this case as even more fine-grained than truth conditions insofar as the logical equivalence consists of transforming the same atomic symbols as in '~~*Fa*' and '*Fa*'). Depending on the variable resources and capacities people have of computational memory, processor speed, buffer conditions, time pressures, etc. it is quite natural to expect such psychological differences according to LOTH. Similarly for synonyms: the differences in different psychological contexts should pose no problem because the symbols expressing synonymous contents are syntactically type-distinct. LOTH can accommodate those differences by appealing to their different computational profiles arising from their different syntactic properties.

Perry's example seems to show that there is a special Mentalese vehicle underlying the beliefs reported by an essential use of indexicals. It is only when I am belief related to a token of #I am making a mess#, that the ensuing typical behavior occurs. This is as it should be. Intentional explanation of behavior, according to LOTH, involves appeal to computational processes defined over sentences underlying those intentional states. There should thus be no surprise if it turns out that our computational organization requires special symbols underlying self-beliefs: they may be the ones reserved for certain computational jobs but not for others. In this case, in order to ascribe a property to myself in such a way that it would make a causal difference in *my* behavior, I need to represent that property in my Mentalese as predicating of a special symbol, #I#. The details here are certainly

empirical. Indeed, there are some well-known psychological abnormalities that seem to require some such computational treatment of the psychology of self [[refs]].

So although both #I am making a mess# and #Murat Aydede is making a mess# (or, #*He* is making a mess#) may have the same propositional content expressed by (3) above, they remain syntactically type-distinct symbols; hence the difference in *my belief states*. And the ensuing change in my behavior comes about only when I come to believe that *I myself* am making a mess.[36]

10 *Individuation of Mentalese Symbols

[[To be completed later... But see my "On the Type/Token Relation of Mental Representations",*Facta Philosophica: International Journal of Contemporary Philosophy*, Vol. 2, No. 1, pp. 23–49, March 2000. Available at: http://humanities.uchicago.edu/faculty/aydede/Typing.pdf]]

11 *The Connectionism/Classicism Debate

When Jerry Fodor published his influential book, *The Language of Thought*, in (1975), he called LOTH "the only game in town." As we have seen, it was the philosophical articulation of the assumptions that underlay the new developments in "cognitive sciences" after the demise of behaviorism. Fodor argued for the truth of LOTH on the basis of the successes of the best scientific theories we had then. Indeed most of the scientific work in cognitive psychology, psycholinguistics, and AI assumed the framework of LOTH. This remains true for most of the present scientific work in these disciplines.

In the early 1980's, however, Fodor's claim that LOTH was the only game in town was beginning to be challenged by some people who were working on so-called connectionist networks. They claimed that connectionism offered a new and radically different alternative to classicism in modeling cognitive phenomena. The name 'classicism' has since then become to be applied to the LOTH framework. On the other hand, many classicists like Fodor thought that connectionism was nothing but a slightly more sophisticated way with which the old and long dead associationism, whose roots could be traced back to early British empiricists, was being revived. In 1988 Fodor and Pylyshyn (F&P) published a long article, "Connectionism and Cognitive Architecture: A Critical Analysis", in which they launched a formidable attack on connectionism, which largely set the terms for the ensuing debate between connectionists and classicists.

F&P's forceful criticism consists in posing a dilemma for connectionists: They either fail to explain the law-like cognitive regularities like systematicity and productivity in an adequate way or the connectionist models are nothing but mere implementation models of classical architectures; hence, they fail to provide a

[36] See Rey (1997: 290–92), Lycan (1981) for more elaboration of this kind of treatment of essential indexicals.

radically new paradigm as connectionists claim. This conclusion was also meant to be a challenge: Explain the cognitive regularities in question without postulating a LOT architecture.

First, let me present F&P's argument against connectionism in a somewhat reconstructed fashion. It will be helpful to characterize the debate by locating the issues according to the reactions many connectionists had to the premises of the argument.

F&P's Argument against Connectionism in their (1988):

> **(i)** Cognition essentially involves representational states and causal operations whose domain and range are these states; consequently, any scientifically adequate account of cognition should acknowledge such states and processes.
>
> **(ii)** Higher cognition (specifically, thought and thinking with propositional content) conceived in this way, has certain scientifically interesting properties: in particular, it is a law of nature that cognitive capacities are *productive*, *systematic*, and *inferentially coherent*.
>
> **(iii)** Accordingly, the architecture of any proposed cognitive model is scientifically adequate only if it guarantees that cognitive capacities are productive, systematic, etc. This would amount to explaining, in the scientifically relevant and required sense, how it could be a law that cognition has these properties.
>
> **(iv)** The only way (i.e. necessary condition) for a cognitive architecture to guarantee systematicity (etc.) is for it to involve a representational system for which (B) is true (see above). (Classical architectures necessarily satisfy (B).)
>
> **(v)** Either the architecture of connectionist models does satisfy (B), or it does not.
>
> **(vi)** If it does, then connectionist models are implementations of the classical LOT architecture and have little new to offer (i.e., they fail to compete with classicism, and thus connectionism does not constitute a radically new way of modeling cognition).
>
> **(vii)** If it does not, then (since connectionism does not then guarantee systematicity, etc., in the required sense) connectionism is empirically false as a theory of the *cognitive* architecture.
>
> **(viii)** Therefore, connectionism is either true as an implementation theory, or empirically false as a theory of cognitive architecture.

The notion of *cognitive architecture* assumes special importance in this debate. So it is useful to say a few words on this. F&P's characterization of the notion goes as follows:

> The architecture of the cognitive system consists of the set of basic operations, resources, functions, principles, etc. (generally the sorts of properties that would be described in a "user's manual'' for that architecture if it were

available on a computer) whose domain and range are the *representational states* of the organism. (1988: 10)

Their emphasis here is on what makes an architecture a cognitive one. But let us first focus on what an architecture is.

As suggested by the parenthetical remark, what F&P seem to have in mind here is whatever notion of architecture is involved when we consider current high-level computer programming languages like BASIC, PASCAL, PROLOG, LISP, etc. These languages have different architectures in that their syntax and organization (e.g., some may require ample use of "GO TO'' statement, whereas others not, thus forcing the programmer to write highly "structured'' programs, etc.), primitive operations (e.g., the square root function might be primitive in one but not in others, etc.), use of computational resources (e.g., memory, processor time), and the like, are different. In this sense, the architecture of these universal languages is indeed what is being described in their "user's manual'' (e.g., when you buy an over-the-counter compiler for one of these languages).[37]

So, if the notion of a (computational) architecture is to be understood in this way, i.e. on analogy to what is described in the "user's manual'' of programming languages, what makes it cognitive? According to F&P, when we talk about the *cognitive* architecture of the (computational) mind/brain, we are talking about a computational level whose primitive operations, functions, etc. have, as their domain and range, representational states, i.e., data structures (symbols) that, at a minimum, represent the states of affairs in the world. So, an architecture is *cognitive* if, and only if, what is being processed in this architecture has such representational content.

F&P want to say, then, of any such cognitive architecture that it is *classical* if, and only if, (B1) is true of what is being thus processed (i.e., representations) and the processing architecture does actually exploit the (syntactic/formal) structural features of the representations in processing them (hence, B2).

Also, it is important to note that (B1) and (B2) are abstract meta-architectural properties in that they are themselves conditions upon any proposed specific architecture's being classical. There are indefinitely many possible classical architectures. To illustrate the point, consider, for instance, different formulations of sentential logic: in one, the only formally complex sentences may be negations and conditionals in which case the transformation rules that are appropriate for these would define the primitive processing operations; in others, all the five standard logical forms of sentences and different sets of primitive rules for transforming them might be given. But (B) would come out to be true of any

[37] Robert Cummins has criticized Pylyshyn's (1984) notion of functional architecture and proposed a more specific notion of cognitive architecture: "Pylyshyn often makes it sound as if the primitive operations of a programming language define a functional architecture, but this cannot be right. The functional [cognitive] architecture of the mind is supposed to be that aspect of the mind's structure that remains fixed across data structures (i.e., in what is represented). This is the [hardwired] program itself, including its control structure, not the primitive operations of a language we might write in" (Cummins 1989: 165–6).

different formulation of sentential logic if considered as a representational system run in a computational architecture. Similarly, any architecture (LISP, PROLOG, etc.) that would process such representations in a structure-sensitive way would count as a classical one. This is the sense in which (B1) and (B2) are abstract meta-architectural properties. They define classicism per se, but not any particular way of being classical. Classicism as such, then, is not committed to any particular architecture or to any particular B-like representational system in advance. It simply claims that whatever the *particular* cognitive architecture of the brain might turn out to be (whatever the *specific* grammar of Mentalese turns out to be), (B) must be true of it. F&P claim that this is the only way an architecture can be said to guarantee the nomological necessity of cognitive regularities like systematicity, etc. This seems to be the relevant and required sense in which a scientific explanation of cognition is required to guarantee the regularities — hence the third premise in their argument.

Let us turn to the premises themselves. The acceptance of premise (i) of their argument, as F&P point out, draws a general line between two radically different traditions in the philosophy of mind; namely, between eliminativism and representationalism (or, representational realism), and places the connectionists within the representationalist camp. However, not every connectionist or philosopher who views connectionism as a radical and promising theory would like to see herself placed in this camp. Indeed, there has been a considerable controversy going on as to whether connectionism is a new theory with the necessary resources to constitute a serious challenge to the fundamental tenets of folk psychology.[38] For this reason, those connectionists who reject premise (i) are viewed to promote an approach that is sometimes called *radical* or *eliminativist connectionism*. It seems, however, that it is too early to assess the potential of connectionism in terms of the support it gives to the elimination of folk psychology.[39] On the other hand, many

[38] For example, Churchlands, who have been the champions of eliminativism, hope that connectionism is the long waited theory which will provide the scientific foundations of the elimination of folk psychological constructs in "psychology" (P.S. Churchland 1986, 1987; Churchland and Sejnowski 1989; P.M. Churchland 1990; P.S. Churchland and P.M. Churchland 1990). Ramsey, Stich and Garon (1991) have recently argued that if certain sorts of connectionist models turn out to be right then the elimination of folk psychology will be inevitable. Dennett (1986), and Cummins and Schwartz (1987) have also pointed out the potential of connectionism in the elimination of at least certain aspects of folk psychology.

[39] In fact, it is not at all clear, how connectionism can genuinely give support to intentional eliminativism as far as the units (or collections of units) in connectionist networks are treated as representing. If they are not treated as such, it is hard to see how they could be models of *cognitive* phenomena, and thus hard to see how they can present any eliminativist challenge. However, there appear to be two vague strands among eliminativists in this regard. One stems from the intuition that it is unlikely that there are really any concrete, isolatable, and modularly identifiable symbol structures realized in the brain that would correspond to what Stich has called (1983: 237ff.) functionally discrete beliefs and desires of folk psychology, and connectionist networks, it is claimed, will vindicate this intuition. For similar remarks, among others, see Dennett (1986, 1991a), Clark (1988, 1989b). The second trend seems to be that connectionism will vindicate that the explanation of mental phenomena doesn't require a full-blown semantics for such higher-order states as propositional attitudes. Rather, all that is

connectionists *do in fact* advance their models as having causally efficacious representational states, and explicitly endorse F&P's first premise. In this regard, they seem to accept intentional realism.[40]

Connectionists who accept premise (i) can be divided into three groups in their reactions to F&P's argument. One group may be seen as more or less accepting the cogency of the entire argument; this group characterizes itself as implementationalist. According to this group, the appropriate niche for neural networks is closer to neuroscience than to cognitive psychology. They seem to view the importance of the program in terms of its prospects of closing the gap between the neurosciences and high-level cognitive theorizing. In this, many seem content to admit premise (vi). However, it seems that this implementationalist outlook in no way diminishes the significance of connectionist research. On the contrary, its importance is emphasized by McLaughlin in just these terms:

> ... think of implementation this way: if connectionism implements classicism, connectionism gets to be the quantum mechanics and classicism only gets to be chemistry. If there were a Nobel Prize in psychology, an account of how a connectionist network in the brain implements a classical cognitive architecture would surely win it. (McLaughlin 1993a: 184)

It is also plausible to claim that implementational models would put just the right sort of theoretical pressure on high-level cognitive modeling (and vice versa, of course). This would result in a healthy co-evolution of these two levels. The prediction of such co-development would surely be supported by parallel cases from the history of science.

Of course, the major portion of the ongoing debate has been generated by the remaining two connectionist groups. As I said, F&P's dilemma was also meant as a challenge to connectionists: Adequately explain cognitive regularities like systematicity (etc.) without postulating a classical architecture. One group, who took

---

needed is an account of some form of information processing at a much lower level, which, it is hoped, will be sufficient for the whole range of cognitive phenomena. Again, it is not clear what the proposals are. But see Paul Churchland (1990).

[40] It seems clear from some of the so far proposed models that many connectionists have been developing their models ultimately with an eye to capture the generalizations in their respective psychological domain. To see this it is enough to look at some of the papers in the second PDP volume (Rumelhart, McClelland and the PDP Research Group, 1986) among which Rumelhart and McClelland's paper on modeling learning the past tenses of English verbs is particularly celebrated. At the end, it is of course an open empirical question whether connectionist models will ultimately be able to capture them, or whether the generalizations they come up with will be compatible with or be the ones implicitly recognized by the folk, just as it is an open question whether classical models will ultimately be successful in this respect. Moreover, it is also empirically possible that connectionists might at the end be forced to give up interpreting the states of their models as representational or as causally efficacious. Whatever the final outcome might be, however, it is prima facie the case that many connectionists intend their models to be taken as contributions within the intentional realist tradition. Smolensky (1988) is the most articulated defense of something like this position. He calls his position "the Proper Treatment of Connectionism" (PTC) and clearly separates it from various eliminativist positions.

the challenge very seriously, attempted to meet it by developing quite interesting connectionist models. The other group declined to meet the challenge on a number of grounds.

What unites this latter group is their rejection of premise (ii) or (iv), or both.[41] What follows is a small sample with a few glosses on each. Some connectionists, and philosophers sympathetic to connectionism, have seriously questioned whether human cognitive capacities are productive and/or systematic.[42] Those who reject that cognitive capacities are productive tend also to reject a robust distinction between competence and performance: most are just unwilling to accept what appear to be quite strong assumptions that go into making such a distinction. With respect to systematicity, they seem to think that if human cognitive capacities are systematic through and through as classicists maintain, that would be a very surprising fact from the perspective of the evolutionary biology of our species.

Andy Clark (1989b, 1991) has claimed that productivity and systematicity are artifacts of natural language usage. In a somewhat Dennettian way, he seems to think that the logic of propositional attitude ascriptions makes it a conceptual necessity that we treat those we attribute propositional attitudes to as having productive/systematic cognitive capacities. So he claims that it is a *conceptual* truth that cognitive capacities are productive/systematic. Hence, there is no need to explain this fact as there would be if it is an *empirical* fact, i.e., by postulating a cognitive architecture that satisfies (B).

Those who have questioned premise (iv) have a motley of reasons. Some (e.g. Sterelny 1990, Braddon-Mitchell and Fitzpatrick 1990) have argued that to the extent that human cognitive capacities exhibit them the cognitive regularities can adequately be explained without recourse to a B-like architecture. They attempt to give a "diachronic" as opposed to a "synchronic" explanation of the cognitive regularities in question by appealing to the evolution of cognitive organisms that exhibit these regularities.

Another attempt to reject (iv) is made by Keith Butler (1991). He accepts that connectionists are committed to atomic representations (lacking constituent structure) and goes on to accept the productivity and systematicity of cognitive capacities. The way he attempts to reconcile the two is by appeal to etiological histories of atomic representations.

Robert Matthew (1994) complains that F&P's demand for an adequate explanation of cognitive regularities is hard for connectionists to meet, as long as the very notion of explanation insisted on by F&P is accepted. He seems to think that F&P's notion

---

[41] Premise (iii) is intimately connected to (ii) and (iv). So its rejection by itself does not mean much. As I mentioned, premise (iii), according to F&P, is there to prevent certain *ad hoc* solutions on the part of connectionists in the explanation of cognitive regularities mentioned in (ii). Premise (v) is close to being a tautology. So no one has any quarrel with it, although van Gelder (1991) comes very close to rejecting it on the ground that with every shift in scientific paradigms the conceptual apparatus of the previous and challenged paradigms becomes inadequate to correctly characterize the new and challenging paradigm.

[42] Dennett (1991b), Sterelny (1990), Rumelhart and McClelland (1986).

of explanation is highly idiosyncratic and narrow, and especially biased towards classicism: if the constraints on explanation are appropriately relaxed / corrected, which, according to Matthew, connectionists ought to insist on, then there is no reason to think that connectionists cannot adequately explain the relevant cognitive regularities.

Some (e.g., Aizawa (forthcoming); Garson (forthcoming); Wallis (forthcoming)) argued that connectionism is no less vulnerable to the same sort of criticism than classical models: classical models don't guarantee systematicity either, since they can be programmed to be unsystematic in the F&P's sense — it all depends on what kind of specific grammar / program the system employs. In other words, they claim, mere satisfaction of (B) by a representational system is not sufficient to guarantee systematicity, hence premise (iv) is problematic even for LOTH.

Classicists have responded to many of these connectionist rebuttals, and the debate between this group and classicists is still very lively.[43]

The group of connectionists who have taken F&P's challenge seriously have tended to reject premise (vi) in their argument, while accepting, on the face of it, the previous five premises (sometimes with reservations on the issue of productivity). Prominent in this group are Smolensky (1990a, 1990b, 1995), van Gelder (1989, 1990, 1991), Chalmers (1990, 1991). Some connectionists whose models gave support to this line include Elman (1989), Hinton (1990), Touretzky (1990), Pollack (1990). Smolensky, for instance, is very explicit in his rejection of premise (vi):

> ...distributed connectionist architectures, without implementing the Classical architecture, can nonetheless provide structured mental representations and mental processes sensitive to that structure. (1990a: 215)

Rejection of (vi) raises interesting and difficult questions. For it seems that once a representational model satisfies (B), there is very little room left to claim that it is not a classical or LOT model. (B) is part of the very definition of what makes a model classical. So how is it possible to reject premise (vi)? Very roughly put, the connectionists' answer comes down to this: When you devise a representational system whose satisfaction of (B) relies on a *non-concatenative* realization of structural / syntactic complexity of representations, you have a non-classical system, i.e., a system that is in no way an implementation of a classical LOT architecture. (See especially Smolensky 1990a and van Gelder 1990.) Interestingly, some classicists like Fodor and McLaughlin (1990) (F&M) seem to agree. F&M stipulate that you have a classical system only if the syntactic complexity of representations is realized *concatenatively*, or as it is sometimes put, *explicitly*:

> We... stipulate that for a pair of expression types E1, E2, the first is a *Classical* constituent of the second *only if* the first is tokened whenever the second is tokened. (F&M 1990: 186)

[43] See McLaughlin (1993a, 1993b), and McLaughlin and Warfield (forthcoming) for a partitioning of the debate similar to mine, and for extensive criticisms of connectionist rebuttals.

So, given that you have a formally specified representational system, one that apparently satisfies (B), how it is to be realized seems to make a difference to whether the system is a classical one or not.

The issues about how connectionists propose to obtain constituent structure non-concatenatively tend to be complex and technical. But they propose to exploit so called *distributed representations* in certain novel ways. Here in this short space it is impossible to do justice to the richness and potential importance of these new techniques. The essential idea behind most of them is to use vector algebra (involving superimposition, multiplication, etc. of vectors) in composing and decomposing connectionist representations which consist in coding patterns of activity across neuron-like units which can be modeled as vectors. The result of such techniques is the production of representations that have in some interesting sense a complexity whose constituent structure is largely implicit in that the constituents are not tokened explicitly when the representations are tokened, but can be recovered by further operations upon them. The interested reader should consult some of the pioneering work by Elman (1989), Hinton (1990), Smolensky (1989, 1990, 1995), Touretzky (1990), Pollack (1990).

F&M's criticism, more specifically stated, however, is this. Connectionists with such techniques only satisfy (B1) in some "extended sense", but they are incapable of satisfying (B2), precisely because their way of satisfying (B1) is committed to a non-concatenative realization of syntactic structures.

Some connectionists disagree (e.g. Chalmers 1991): they claim that you can have structure-sensitive transformations or operations defined over representations whose syntactic structure is non-concatenatively realized. So given the apparent agreement that non-concatenative realization is what makes a system non-classical, connectionists claim that they can and do perfectly satisfy (B) in its entirety with their connectionist models without implementing classical models.

The debate is still quite intense and there is a fast growing literature built around the many issues raised by it. Aydede (1997) offers an extensive analysis of the debate between classicists and this group of connectionists with special attention to the conceptual underpinnings of the debate. He argues that both parties are wrong in assuming that concatenative realization is relevant to the characterization of LOTH. On the one hand, he argues against connectionists that they failed to show that premise (vi) is false: their models are, in some interesting and potentially exciting sense, classical — to the extent to which they are adequate to explain the cognitive regularities, of course. On the other hand, he argues against F&P&M that they fail to show that only concatenatively realized representations can engage in structure sensitive processes and are unconvincing in their insistence that LOTH requires explicit tokening of constituent structure. His aim, in short, is to specify the minimal conditions for the LOTH to be true, and why.[44]

[44] A much shorter precursor of this monograph has appeared in *Stanford Encyclopedia of Philosophy* as an entry on the Language of Thought Hypothesis, edited by Edward N. Zalta, Stanford: CSLI Publications (http / / plato.stanford.edu). I would like to thank Irene

12 Bibliography

Aizawa, K. (forthcoming). "Representations without Rules, Connectionism and the Syntactic Argument."

Aydede, Murat (1995). "Connectionism and Language of Thought", *CSLI Technical Report*, Stanford, CSLI-95-195. (This is an early version of Aydede 1997 but contains quite a lot of expository material not contained in 1997.)

Aydede, Murat (1997). "Language of Thought: The Connectionist Contribution," *Minds and Machines*, Vol. 7, No. 1, pp. 57–101. Available at: http://humanities.uchicago.edu/faculty/aydede/lot.connex.pdf

Aydede, Murat (2000). "On the Type/Token Relation of Mental Representations",*Facta Philosophica: International Journal of Contemporary Philosophy*, Vol. 2, No. 1, pp. 23–49. Available at: http://humanities.uchicago.edu/faculty/aydede/Typing.pdf

Armstrong, D.M. (1980). *The Nature of Mind*, Ithaca, NY: Cornell University Press.

Barsalou, L. W. (1993a). "Flexibility, Structure, and Linguistic Vagary in Concepts: Manifestations of a Compositional System of Perceptual Symbols" in *Theories of Memory*, edited by A. Collins, S. Gathercole, M. Conway and P. Morris, Hillsdale, NJ: Lawrence Erlbaum Associates.

Barsalou, L. W., W. Yeh, B. J. Luka, K. L. Olseth, K. S. Mix, and L.-L. Wu. (1993b). "Concepts and Meaning," *Chicago Linguistics Society* 29.

Barsalou, L. W., and J. J. Prinz. (1997). "Mundane Creativity in Perceptual Symbol Systems" in *Creative Thought: An Investigation of Conceptual Structures and Processes*, edited by T. B. Ward, S. M. Smith and J. Vaid, Washington, DC: American Psychological Association.

Barwise, Jon and John Etchemendy (1995). *Hyperproof*, Stanford, Palo Alto: CSLI Publications.

Barwise, J. and J. Perry (1983). *Situations and Attitudes*, Cambridge, Massachusetts: MIT Press.

Bechtel, W. and A. Abrahamsen (1991). *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*, Oxford, UK: Basil Blackwell.

Blackburn, S. (1984). *Spreading the Word*, Oxford, UK: Oxford University Press.

Block, Ned (1980). "Troubles with Functionalism" in *Readings in Philosophy of Psychology*, N. Block (ed.), Vol.1, Cambridge, Massachusetts: Harvard University Press, 1980. (Originally appeared in *Perception and Cognition: Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science*, C.W. Savage (ed.), Minneapolis: The University of Minnesota Press, 1978.)

Block, N. (ed.) (1981). *Imagery*. Cambridge, Massachusetts: MIT Press.

Appelbaum, David Chalmers, Eric Margolis, Jesse Prinz, Philip Robbins, Brian C. Smith, Edward Zalta for their comments on an earlier draft of this piece.

Block, N. (1983a). "Mental Pictures and Cognitive Science," *Philosophical Review* 93: 499–542. (Reprinted in *Mind and Cognition*, W.G. Lycan (ed.), Oxford, UK: Basil Blackwell, 1990.)

Block, N. (1983b). "The Photographic Fallacy in the Debate about Mental Imagery," *Nous* 17: 651–62.

Block, Ned (1986). "Advertisement for a Semantics for Psychology" in *Studies in the Philosophy of Mind: Midwest Studies in Philosophy*, Vol. 10, P. French, T. Euhling and H. Wettstein (eds.), Minneapolis: University of Minnesota Press.

Boër, Steven E. and W. Lycan (1986). *Knowing Who*, Cambridge, Massachusetts: MIT Press.

Boër, Steven E. (1995). "Propositional Attitudes and Compositional Semantics" in *Philosophical Perspectives **9**: AI, Connectionism and Philosophical Psychology*, James E. Tomberlin (ed.), Atascadero, California: Ridgeview Publishing Company, 1995.

Braddon-Mitchell, David and John Fitzpatrick (1990). "Explanation and the Language of Thought," *Synthese* 83: 3–29.

Brentano, Franz (1874/1973). *Psychology from an Empirical Standpoint*, A. Rancurello, D. Terrell and L. McAlister (trans.), London: Routledge and Kegan Paul.

Butler, Keith (1991). "Towards a Connectionist Cognitive Architecture," *Mind and Language*, Vol. 6, No. 3, pp. 252–72.

Chalmers, David (1990). "Syntactic Transformations on Distributed Representations," *Connection Science*, Vol. 2.

Chalmers, David (1991). "Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation," in *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pp. 340–7.

Chalmers, David (1996). *The Conscious Mind: In Search of a Fundamental Theory*, Oxford, UK: Oxford University Press.

Chisholm, Roderick (1957). *Perceiving: A Philosophical Study*, Ithaca, NY: Cornell University Press.

Churchland, Patricia Smith (1986). *Neurophilosophy: Toward a Unified Science of Mind-Brain*, Cambridge, Massachusetts: MIT Press.

Churchland, Patricia Smith (1987). "Epistemology in the Age of Neuroscience," *Journal of Philosophy*, Vol. 84, No. 10, pp. 544–553.

Churchland, Patricia S. and Terrence J. Sejnowski (1989). "Neural Representation and Neural Computation" in *Neural Connections, Neural Computation*, L. Nadel, L.A. Cooper, P. Culicover and R.M. Harnish (eds.), Cambridge, Massachusetts: MIT Press, 1989.

Churchland, Paul M. (1990). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, Cambridge, Massachusetts: MIT Press.

Churchland, Paul M. (1981). "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy* 78: 67–90.

Churchland, Paul M. and P.S. Churchland (1990). "Could a Machine Think?," *Scientific American*, Vol. 262, No. 1, pp. 32–37.

Clark, Andy (1988). "Thoughts, Sentences and Cognitive Science," *Philosophical Psychology*, Vol. 1, No. 3, pp. 263–278.

Clark, Andy (1989a). "Beyond Eliminativism," *Mind and Language*, Vol. 4, No. 4, pp. 251–279.

Clark, Andy (1989b). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*, Cambridge, Massachusetts: MIT Press.

Clark, Andy (1990). "Connectionism, Competence, and Explanation," *British Journal for Philosophy of Science*, 41: 195–222.

Clark, Andy (1991). "Systematicity, Structured Representations and Cognitive Architecture: A Reply to Fodor and Pylyshyn" in *Connectionism and the Philosophy of Mind*, Terence Horgan and John Tienson (eds.), *Studies in Cognitive Systems* (Volume 9), Dordrecht: Kluwer Academic Publishers, 1991.

Clark, Andy (1994). "Language of Thought (2)" in *A Companion to the Philosophy of Mind*, edited by S. Guttenplan, Oxford, UK: Basil Blackwell, 1994.

Crimmins, Mark and John Perry (1989). "The Prince and the Phone Booth: Reporting Puzzling Beliefs", *Journal of Philosophy*, 86: 685–711.

Crimmins, Mark (1992). *Talk About Beliefs*, Cambridge, Massachusetts: The MIT Press.

Cummins, Robert (1986). "Inexplicit Information" in The Representation of Knowledge and Belief, M. Brand and R.M. Harnish (eds.), Tucson, Arizona: Arizona University Press, 1986.

Cummins, Robert (1989). *Meaning and Mental Representation*, Cambridge, Massachusetts: MIT Press.

Cummins, Robert and G. Schwarz (1987). "Radical Connectionism," *The Southern Journal of Philosophy*, Vol. XXVI, Supplement.

Davidson, Donald (1980). "Freedom to Act" in *Essays on Actions and Events*, D. Davidson, Oxford, UK: Oxford University Press.

Davidson, Donald (1984). *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press.

Davies, Martin (1989). "Connectionism, Modularity, and Tacit Knowledge," *British Journal for the Philosophy of Science* 40: 541–555.

Davies, Martin (1991). "Concepts, Connectionism, and the Language of Thought," in *Philosophy and Connectionist Theory*, W. Ramsey, S.P. Stich and D.E. Rumelhart (eds.), Hillsdale, NJ: Lawrence Erlbaum, 1991.

Davies, M. (1995). "Two Notions of Implicit Rules," *Philosophical Perspectives* 9: 153–83.

Dennett, D.C. (1978). “Two Approaches to Mental Images” in *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, Massachusetts: MIT Press, 1981.

Dennett, Daniel C. (1981). *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, Massachusetts: MIT Press.

Dennett, D.C. (1981a). “Cure for the Common Code” in *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, Massachusetts: MIT Press, 1981. (Originally appeared in *Mind*, April 1977.)

Dennett, Daniel C. (1986). “The Logical Geography of Computational Approaches: A View from the East Pole” in *The Representation of Knowledge and Belief*, Myles Brand and Robert M. Harnish (eds.), Tucson: The University of Arizona Press, 1986.

Dennett, Daniel C. (1987). *The Intentional Stance*, Cambridge, Massachusetts: MIT Press.

Dennett, Daniel C. (1991a). “Real Patterns,” *Journal of Philosophy*, Vol. LXXXVIII, No. 1, pp. 27–51.

Dennett, Daniel C. (1991b). “Mother Nature Versus the Walking Encyclopedia: A Western Drama” in *Philosophy and Connectionist Theory*, W. Ramsey, S.P. Stich and D.E. Rumelhart (eds.), Lawrence Erlbaum Associates.

Descartes, R. (1637/1970). “Discourse on the Method” in *The Philosophical Works of Descartes*, Vol. I, E.S. Haldane and G.R.T. Ross (trans.), Cambridge, UK: Cambridge University Press.

Descartes, R. (1649/1970). “The Passions of the Soul” in *The Philosophical Works of Descartes*, Vol. I, E.S. Haldane and G.R.T. Ross (trans.), Cambridge, UK: Cambridge University Press.

Devitt, Michael (1990). “A Narrow Representational Theory of the Mind,” *Mind and Cognition*, W.G. Lycan (ed.), Oxford, UK: Basil Blackwell, 1990.

Devitt, Michael (1996). *Coming to our Senses: A Naturalistic Program for Semantic Localism*, Cambridge, UK: Cambridge University Press.

Devitt, Michael and Sterelny, Kim (1987). *Language and Reality: An Introduction to the Philosophy of Language*, Cambridge, Massachusetts: MIT Press.

Dretske, Fred (1981). *Knowledge and the Flow of Information*, Cambridge, Massachusetts: MIT Press.

Dretske, Fred (1988). *Explaining Behavior*, Cambridge, Massachusetts: MIT Press.

Elman, Jeffrey L. (1989). “Structured Representations and Connectionist Models”, *Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society*, Ann Arbor, Michigan, pp. 17–23.

Field, Hartry H. (1972). “Tarski’s Theory of Truth,” *Journal of Philosophy*, 69: 347–75.

Field, Hartry H. (1978). “Mental Representation,” *Erkenntnis* 13, 1, pp. 9–61. (Also in *Mental Representation: A Reader*, S.P. Stich and T.A. Warfield (eds.), Oxford, UK: Basil Blackwell, 1994. References in the text are to this edition.)

Fodor, Jerry A. (1975). *The Language of Thought*, Cambridge, Massachusetts: Harvard University Press.

Fodor, Jerry A. (1978). "Propositional Attitudes" in *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, J.A. Fodor, Cambridge, Massachusetts: MIT Press, 1981. (Originally appeared in *The Monist* 64, No. 4, 1978.)

Fodor, Jerry A. (1978a). "Computation and Reduction" in *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, J.A. Fodor, Cambridge, Massachusetts: MIT Press. (Originally appeared in *Minnesota Studies in the Philosophy of Science: Perception and Cognition*, Vol. 9, W. Savage (ed.), 1978.)

Fodor, Jerry A. (1978b). "Tom Swift and His Procedural Grandmother," *Cognition*, Vol. 6. (Also in *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, J.A. Fodor, Cambridge, Massachusetts: MIT Press, 1981.)

Fodor, Jerry A. (1980). "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology," *Behavioral and Brain Sciences* 3, 1, 1980. (Also in *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, J.A. Fodor, Cambridge, Massachusetts: MIT Press, 1981. References in the text are to this edition.)

Fodor, Jerry A. (1981a). *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. (1981b), "Introduction: Something on the State of the Art" in *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, J.A. Fodor, Cambridge, Massachusetts: MIT Press, 1981.

Fodor, Jerry A. (1983). *The Modularity of Mind*, Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. (1985). "Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade Mecum", *Mind* 94, 1985, pp. 76–100. (Also in *A Theory of Content and Other Essays*, J.A. Fodor, Cambridge, Massachusetts: MIT Press. References in the text are to this edition.)

Fodor, Jerry A. (1986). "Banish DisContent" in *Language, Mind, and Logic*, J. Butterfield (ed.), Cambridge, UK: Cambridge University Press, 1986. (Also in *Mind and Cognition*, William Lycan (ed.), Oxford, UK: Basil Blackwell, 1990.)

Fodor, Jerry A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. (1989). "Substitution Arguments and the Individuation of Belief" in *A Theory of Content and Other Essays*, J. Fodor, Cambridge, Massachusetts: MIT Press, 1990. (Originally appeared in *Method, Reason and Language*, G. Boolos (ed.), Cambridge, UK: Cambridge University Press, 1989.)

Fodor, Jerry A. (1990). *A Theory of Content and Other Essays*, Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. (1990a). "A Theory of Content" (I & II) in *A Theory of Content and Other Essays*, Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. (1991). "Replies'' (Ch. 15) in *Meaning in Mind: Fodor and his Critics*, B. Loewer and G. Rey (eds.), Oxford, UK: Basil Blackwell, 1991.

Fodor, Jerry A. (1998). *Concepts: Where Cognitive Science Went Wrong*, Oxford, UK: Oxford University Press.

Fodor, Jerry A. and Ernest Lepore (1991). "Why Meaning (Probably) Isn't Conceptual Role?", *Mind and Language*, Vol. 6, No. 4, pp. 328–43.

Fodor, Jerry A. and B. McLaughlin (1990). "Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work," *Cognition* 35: 183–204.

Fodor, Jerry A. and Zenon W. Pylyshyn (1988). "Connectionism and Cognitive Architecture: A Critical Analysis" in S. Pinker and J. Mehler, eds., *Connections and Symbols*, Cambridge, Massachusetts: MIT Press (A *Cognition* Special Issue).

Forbes, Graeme (1987). "Indexicals and Intensionality: A Fregean Perspective", *Philosophical Review*, Vol. XCVI, No. 1.

Forbes, Graeme (1990). "The Indispensability of *Sinn*," *Philosophical Review*, 99: 535–64.

Forbes, Graeme (1996). "Substitutivity and the Coherence of Quantifying In," *Philosophical Review*, Vol. 105, No. 3.

Frege, Gottlob (1892 / 1943). "On Sense and Nominatum" in *Readings in Philosophical Analysis*, H. Feigl and W. Sellars (eds.), New York: Appleton-Century-Crofts, pp. 85-102. (Originally appeared in German in *Eitschr. f. Philos. und Philos. Kritik*, 100, 1892.)

Garson, J. (forthcoming). "Systematicity and Classical Architecture."

Gödel, Kurt (1930). "Der Vollstanddigkeit der Akiom des logischen Funktionenkalkuls," *Monatschefte fur Mathematik und Physik*, 37: 349–60.

Grice, H.P. (1957). "Meaning," *Philosophical Review*, 66: 377–88.

Hadley, R. F. (1995). "The 'Explicit-Implicit' Distinction," *Minds and Machines*, 5: 219–42.

Harman, Gilbert (1973). *Thought*, Princeton, NJ: Princeton University Press.

Haugeland, John (1981). "The Nature and Plausibility of Cognitivism," *Behavioral and Brain Sciences* I, 2: 215–60 (with peer commentary and replies).

Haugeland, John (1985). *Artificial Intelligence: The Very Idea*, Cambridge, Massachusetts: MIT Press.

Hinton, Geoffrey (1990). "Mapping Part-Whole Hierarchies into Connectionist Networks," *Artificial Intelligence*, Vol. 46, Nos. 1–2, (Special Issue on Connectionist Symbol Processing).

Horgan, T. E. and J. Tienson (1996). *Connectionism and the Philosophy of Psychology*, Cambridge, Massachusetts: MIT Press.

Kaplan, David (1989). “Demonstratives” in *Themes from Kaplan*, J. Almog, J. Perry and H. Wettstein (eds.), Oxford, UK: Oxford University Press.

Kahneman, D., P. Slovic and A. Tversky (1982). *Judgement under Uncertainty: Heuristics and Biases*, Cambridge, UK: Cambridge University Press.

Kirsh, D. (1990). “When Is Information Explicitly Represented?” in *Information, Language and Cognition*. P. Hanson (ed.), University of British Columbia Press.

Kosslyn, S.M. (1980). *Image and Mind*. Cambridge, Massachusetts: Harvard University Press.

Kosslyn, S.M. (1981). “The Medium and the Message in Mental Imagery: A Theory” in *Imagery,* N. Block (ed.), Cambridge, Massachusetts: MIT Press, 1981.

Kosslyn, S.M. (1994). *Image and Brain*, Cambridge, Massachusetts: MIT Press.

Kripke, Saul (1979). “A Puzzle about Belief” in *Meaning and Use*, A. Margalit (ed.) Dordrecht: Reidel.

Kripke, Saul (1980). *Naming and Necessity*, Cambridge, Massachusetts: Harvard University Press.

Laurence, Stephen and Eric Margolis (1997). “Regress Arguments Against the Language of Thought,” *Analysis*, Vol. 57, No. 1.

Lewis, David (1972). “Psychophysical and Theoretical Identifications,” *Australasian Journal of Philosophy*, 50(3): 249–58. (Also in *Readings in Philosophy of Psychology*, Ned Block (ed.), Vols. 1, Cambridge, Massachusetts: Harvard University Press, 1980.)

Loar, Brian F. (1982a). *Mind and Meaning*, Cambridge, UK: Cambridge University Press.

Loar, Brian F. (1982b). “Must Beliefs Be Sentences?” in *Proceedings of the Philosophy of Science Association for 1982*, Asquith, P. and T. Nickles (eds.), East Lansing, Michigan, 1983.

Lycan, William G. (1981). “Toward a Homuncular Theory of Believing,” *Cognition and Brain Theory* 4(2): 139–159.

Lycan, W. G. (1986). “Tacit Belief” in *Belief: Form, Content, and Function,* R. Bogdan (ed.), Oxford, UK: Oxford University Press.

Lycan, William (1993). “A Deductive Argument for the Representational Theory of Thinking,” *Mind and Language*, Vol. 8, No. 3, pp. 404–22.

Lycan, William (1997). “Consciousness as Internal Monitoring” in *The Nature of Consciousness: Philosophical Debates*, edited by N. Block, O. Flanagan and G. Güzeldere, Cambridge, Massachusetts: MIT Press.

Margolis, Eric (forthcoming). “How to Acquire a Concept?”, *Mind and Language*.

Marr, David (1982). *Vision*, San Francisco: W. H. Freeman.

Mates, B. (1952). “Synonymity” in *Semantics and the Philosophy of Language*, L. Linsky (ed.), Urbana: University of Illinois Press.

Matthew, Robert J. (1994). “Three-Concept Monte: Explanation, Implementation and Systematicity”, *Synthese*, Vol. 101, No. 3, pp. 347–63.

McGinn, Colin (1991). *The Problem of Consciousness*, Oxford, UK: Basil Blackwell.

McLaughlin, B.P. (1993a). “The Connectionism / Classicism Battle to Win Souls,” *Philosophical Studies* 71: 163–90.

McLaughlin, B.P. (1993b). “Systematicity, Conceptual Truth, and Evolution,” in *Philosophy and Cognitive Science,* C. Hookway and D. Peterson (eds.), Royal Institute of Philosophy, Supplement No. 34.

McLaughlin, B.P. and Ted Warfield (forthcoming). “The Allures of Connectionism Reexamined.”

Millikan, Ruth Garrett (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*, Cambridge, Massachusetts: MIT Press.

Millikan, Ruth Garrett (1993). *White Queen Psychology and Other Essays for Alice*, Cambridge, Massachusetts: MIT Press.

Papineau, D. (1987). *Reality and Representation*, Oxford, UK: Basil Blackwell.

Perry, John (1977). “Frege on Demonstratives,” *Philosophical Review* 86: 474–497. (Reprinted in *The Problem of the Essential Indexical and Other Essays*, J. Perry, Oxford, UK: Oxford University Press, 1993. References in the text are to this edition.)

Perry, John (1979). “The Problem of the Essential Indexical,” *Nous* 13, pp.3-21. (Reprinted in *The Problem of the Essential Indexical and Other Essays*, J. Perry, Oxford, UK: Oxford University Press, 1993. References in the text are to this edition.)

Perry, John and David Israel (1991). “Fodor and Psychological Explanations” in *Meaning in Mind: Fodor and his Critics*, B. Loewer and G. Rey (eds.), Oxford, UK: Basil Blackwell, 1991.

Pinker, Steven (1994). *The Language Instinct: How the Mind Creates Language*, New York: William Morrow and Company.

Pinker, Steven (1997). *How the Mind Works* ??

Pollack, J.B. (1990). “Recursive Distributed Representations,” *Artificial Intelligence*, Vol. 46, Nos. 1–2, (Special Issue on Connectionist Symbol Processing).

Prinz, Jesse J. (1997). *Perceptual Cognition*, Ph.D. Dissertation in philosophy, The University of Chicago.

Putnam, Hilary (1988), *Representation and Reality*, Cambridge, Massachusetts: MIT Press.

Pylyshyn, Z.W. (1978). “Imagery and Artificial Intelligence” in *Perception and Cognition*. W. Savage (ed.), University of Minnesota Press. (Reprinted in *Readings in the Philosophy of Psychology*, N. Block (ed.), Cambridge, Massachusetts: MIT Press, 1980.)

Pylyshyn, Zenon W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, Massachusetts: MIT Press.

Quine, W. (1960). *Word and Object*, Cambridge, Massachusetts: MIT Press.

Ramsey, William, Stephen Stich and Joseph Garon (1991). "Connectionism, Eliminativism and the Future of Folk Psychology," in *Philosophy and Connectionist Theory*, W. Ramsey, D. Rumelhart and Stephen Stich (eds.), Hillsdale, NJ: Lawrence Erlbaum.

Rey, Georges (1981). "What are Mental Images?" in *Readings in the Philosophy of Psychology*, N. Block (ed.), Vol. 2, Cambridge, Massachusetts: Harvard University Press, 1981.

Rey, Georges (1991). "An Explanatory Budget for Connectionism and Eliminativism" in *Connectionism and the Philosophy of Mind*, Terence Horgan and John Tienson (eds.), Studies in Cognitive Systems (Volume 9), Dordrecht: Kluwer Academic Publishers.

Rey, Georges (1992). "Sensational Sentences Switched," *Philosophical Studies* 67: 73–103.

Rey, Georges (1993). "Sensational Sentences" in Consciousness, M. Davies and G. Humphrey (eds.), Oxford, UK: Basil Blackwell, pp. 240–57.

Rey, Georges (1995). "A Not 'Merely Empirical' Argument for a Language of Thought," in *Philosophical Perspectives* 9, J. Tomberlin (ed.), pp. 201–222.

Rey, Georges (1997). *Contemporary Philosophy of Mind: A Contentiously Classical Approach*, Oxford, UK: Basil Blackwell.

Richard, Mark (1990). *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them*, Cambridge, UK: Cambridge University Press.

Rosenthal, D.M. (1997). "A Theory of Consciousness" in *The Nature of Consciousness: Philosophical Debates*, edited by N. Block, O. Flanagan and G. Güzeldere, Cambridge, Massachusetts: MIT Press.

Rumelhart, D.E. and J.L. McClelland (1986). "PDP Models and General Issues in Cognitive Science," in *Parallel Distributed Processing*, Vol. 1, D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, Cambridge, Massachusetts: MIT Press, 1986.

Rumelhart, D.E., J.L. McClelland, and the PDP Research Group (1986). *Parallel Distributed Processing*, (Vols. 1&2), Cambridge, Massachusetts: MIT Press.

Salmon, Nathan (1986). *Frege's Puzzle*, Atascadero, California: Ridgeview Publishing Company.

Saul, Jennifer M. (1993). "Still an Attitude Problem," *Linguistics and Philosophy*, 16(4): 423–435.

Schiffer, Stephen (1981). "Truth and the Theory of Content" in *Meaning and Understanding*, H. Parret and J. Bouvaresse (eds.), Berlin: Walter de Gruyter, 1981.

Searle, John R. (1980). "Minds, Brains, and Programs" *Behavioral and Brain Sciences* III, 3: 417–24.

Searle, John R. (1984). *Minds, Brains and Science*, Cambridge, Massachusetts: Harvard University Press.

Searle, John R. (1990). "Is the Brain a Digital Computer?", *Proceedings and Addresses of the APA*, Vol. 64, No. 3, November 1990.

Searle, John R. (1992). *The Rediscovery of Mind*, Cambridge, Massachusetts: MIT Press.

Shepard, R. and Cooper, L. (1982). *Mental Images and their Transformations*. Cambridge, Massachusetts: MIT Press.

Smolensky, Paul (1988). "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences* 11: 1–23.

Smolensky, Paul (1990a). "Connectionism, Constituency, and the Language of Thought" in *Meaning in Mind: Fodor and His Critics*, B. Loewer and G. Rey (eds.), : Oxford, UK: Basil Blackwell, 1991.

Smolensky, Paul (1990b). "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems," *Artificial Intelligence*, Vol. 46, Nos. 1–2, (Special Issue on Connectionist Symbol Processing), November 1990.

Smolensky, Paul (1995). "Constituent Structure and Explanation in an Integrated Connectionist/Symbolic Cognitive Architecture" in *Connectionism: Debates on Psychological Explanation*, C. Macdonald and G. Macdonald (eds.), Oxford, UK: Basil Blackwell, 1995.

Stalnaker, Robert C. (1984). *Inquiry*, Cambridge, Massachusetts: MIT Press.

Sterelny, K. (1986). "The Imagery Debate," *Philosophy of Science* 53: 560–83. (Reprinted in *Mind and Cognition,* W. Lycan (ed.), Oxford, UK: Basil Blackwell, 1990.)

Sterelny, Kim (1990). *The Representational Theory of Mind*, Cambridge, Massachusetts: MIT Press.

Stich, Stephen (1983). *From Folk Psychology to Cognitive Science: The Case against Belief*, Cambridge, Massachusetts: MIT Press.

Tarski, Alfred (1956). "The Concept of truth in Formalized Languages" in *Logic, Semantics and Metamathematics*, J.Woodger (trans.), Oxford, UK: Oxford University Press.

Touretzky, D.S. (1990). "BoltzCONS: Dynamic Symbol Structures in a Connectionist Network," *Artificial Intelligence*, Vol. 46, Nos. 1–2, (Special Issue on Connectionist Symbol Processing).

Tye, M. (1984). "The Debate about Mental Imagery," *Journal of Philosophy* 81: 678–91.

Tye, M. (1991). *The Imagery Debate*, Cambridge, Massachusetts: MIT Press.

van Gelder, Timothy (1989). "Compositionality and the Explanation of Cognitive Processes", *Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society*, Ann Arbor, Michigan, pp. 34–41.

van Gelder, Timothy (1990). "Compositionality: A Connectionist Variation on a Classical Theme," *Cognitive Science*, Vol. 14.

van Gelder, Timothy (1991). "Classical Questions, Radical Answers: Connectionism and the Structure of Mental Representations" in *Connectionism and the Philosophy of Mind*, Terence Horgan and John Tienson (eds.), Studies in Cognitive Systems (Volume 9), Dordrecht: Kluwer Academic Publishers.

Vendler, Zeno (1972). *Res Cogitans: An Essay in Rational Psychology*, Ithaca: Cornell University Press.

Wallis, C. (forthcoming). "Nomic Necessity and Systematicity."

Wason, Peter C. and P. Johnson-Laird (1972). *Psychology of Reasoning: Structure and Content*, London: Batsford.

Wason, Peter C. (1981). "Understanding and the Limits of Formal Thinking" in *Meaning and Understanding*, H. Parret and J. Bouvaresse (eds.), Berlin: Walter de Gruyter, 1981.

Zalta, Edward N. (forthcoming). "Modes of Presentation and Fregean Senses," MS., CSLI, Stanford. (This could be obtained from Zalta's homepage: http://mally.stanford.edu/zalta.html).