Minds & Machines (2009) 19:1–24
DOI 10.1007/s11023-008-9114-6

Semantic Externalism and the Mechanics of Thought

Carrie Figdor

Received: 25 February 2008 / Accepted: 5 August 2008 / Published online: 26 August 2008

Abstract I review a widely accepted argument to the conclusion that the contents of our beliefs, desires and other mental states cannot be causally efficacious in a classical computational model of the mind. I reply that this argument rests essentially on an assumption about the nature of neural structure that we have no good scientific reason to accept. I conclude that computationalism is compatible with wide semantic causal efficacy, and suggest how the computational model might be modified to accommodate this possibility.

Keywords Computationalism · Semantic externalism · Syntax · Mental causation · Mental representation · Language of Thought

Introduction

It is a truth universally acknowledged that wide semantic properties cannot be causally efficacious in a classical computational model of the mind.[1]

[1] The irony of Jane Austen's original sentence notwithstanding, the overwhelming majority of writers on the problem of the causal efficacy of wide content claim that such properties cannot be causal powers (or, at the very least, find it highly problematic to see how they could be) (Stich 1983; Block 1990; Fodor 1987, 1994, 2000; Kim 1982, 1998; Jackson and Pettit 1988). The classical (Language of Thought) computational theory of mind provides a particular, and revealing, setting for this received view. Exceptions include Burge (1989, 1995) and Baker (1995), who defend a causal-explanatory role for semantics in psychology, although not specifically in a computational model of mind. A number of philosophers have defended alternative roles for content in classical computationalism, such as in the individuation of computational states or systems (Crane 1990; Bontly 1998; Shagrir 2001), in "content-involving" explanations of relational or intentional properties of events (Peacocke 1993, 1999), or in 'bridging' the gap between intentional and computational explanations (Egan 1995). Others (e.g., Stich op. cit.) argue that it plays no role at all. To a first approximation, classical computationalism explains the

C. Figdor (✉)
Department of Philosophy, University of Iowa, 260 EPB, Iowa City, IA 52242, USA
e-mail: carrie-figdor@uiowa.edu

However little known opposing claims may be upon first encounter, this truth is so well fixed in the minds of many that it may be considered hopeless—even incoherent—to argue against it. Still, I want to argue that it follows from a premise that we have no good scientific reason to accept. As a result, the universally acknowledged claim loses much of its credibility. We then have no strong reason to think that a computational model of mind cannot accommodate semantic causal powers, and it becomes an empirical matter—as it should be—as to whether the semantic properties of our beliefs, desires and other psychological states are as causally efficacious as they seem to be.[2] This result, moreover, can be generalized to any information-processing model of the mind whose explanation of the causes of intelligent behavior includes representational states but leaves their semantics out of the causal loop for reasons similar to those ascribed to classical computationalism. I focus on classical (Language of Thought) computationalism because its explanation of the causal mechanisms of thinking is familiar and provides a reference point for similar explanations in alternative models of cognition. Throughout, my references to computationalism (without any qualifier) are to that model.

I'll begin by sketching the basic argument against semantic causal powers, and then show how current scientific evidence provides no justification for a crucial premise of that argument. This premise involves a claim about the nature of structural properties of the brain or central nervous system (or of any implementing mechanism, if we assume multiple realizability). It is assumed that all such properties are intrinsic—that is, that their individuation conditions are independent

Footnote 1 continued
mechanics of thought as the manipulation of sentence-like mental representations in virtue of their syntax by formal rules (Aydede 1997, 2004; Fodor 2000, pp. 18–19; Haugeland 1997; Horst 2005). Further discussion follows in the text.

[2] My discussion concerns what is variously called representational, intentional, semantic or informational content: that feature of a mental representation (paradigmatically, a belief) in virtue of which it is about something (paradigmatically, a worldly state of affairs) and is truth-evaluable. I set aside perceptual content, and so defer the possible application of my solution to computational theories of perception and perceptual states, such as Marr (1982). I focus specifically on extrinsically individuated ("wide") contents. Many philosophers (including myself) hold that the thought experiments by Putnam (1975) and Burge (1979) show that at least some semantic contents are individuated at least partly in terms of the nature of the thinker's physical or social environments; those who disagree may recast the problem as a conditional: if externalism were true, then such contents could not play a causal role in a computational model of the mind. A causally efficacious property, or causal power, is a property of an entity in virtue of which it participates in a singular causal relation; an explanation is causal iff it is grounded in singular causal relation (Strawson 1985; Kim 1993, pp. 254–260). Since my argument does not turn on whether (causal) laws are necessary for (causal) explanations, I will remain neutral regarding that debate. I do hold that semantic causal powers, if there be any, are as genuine as any other kind; others relegate them to proleptic or second-class status (Pettit 1986, p. 19; Jackson and Pettit 1988; Stich 1983; Kim 1984). Burge (op. cit.) and Baker (op. cit.) both adopt a view of causation in which putatively causal claims in any science, including in particular psychology, may be accepted as causal at face value. Although this radically autonomous view of psychological causation might dissolve the causal efficacy problem (or, at the very least, make defending semantic causal powers much easier), I do not rely on it. In any case my conclusion is not that semantic properties *are* causally efficacious within computationalism, but that the model is compatible with them.

of the nature of the thinker's environment.[3] I will argue that at least some neural structures may be individuated in ways that depend in part on what the thinker's environment is like. If some semantic properties supervene on them, then the basic argument fails.[4]

This response leaves untouched other problems facing a complete account of mental causation, including most obviously the ongoing debate regarding whether supervenient properties can be causal. But it does resolve the problem of extrinsic semantic properties, which is one of the main difficulties facing any complete account (Kim 1998, pp. 35–37). In effect, it establishes parallel explanations of the causal efficacy of syntax and semantics; we cannot consistently deny this sort of explanation for the one type of property while affirming it for the other. Moreover, it invites a new conception of mental representations that could help clarify the vexed issue of what a computational theory of mind is committed to.[5] For some may wonder whether my solution is compatible with a truly *computational* model, given the claim long associated with it of the brain's being a syntactic engine and not a semantic one. I address this worry in my concluding remarks. What I hope should become clear overall is that classical computationalism has more untapped explanatory resources that either its defenders or its detractors have given it credit for.

The Problem of Semantic Causal Efficacy in Classical Computationalism

The problem of the causal efficacy of content—itself a special case of what Kim calls the problem of extrinsic mental properties—may be stated as follows:

[3] To a first approximation, an extrinsic property is one that a thing cannot have unless its environment has a certain contingent character; an intrinsic property is one that a thing can have independently of the nature of its environment. More formally, F is an extrinsic property of x = df. x has F, and necessarily, x has F only if some contingent object wholly distinct from x exists; F is intrinsic iff it is not extrinsic (Kim 1982; Lewis 1983; Langton and Lewis 1998; Humberstone 1996; Vallentyne 1997). Some writers on the problem of semantic externalism and causal efficacy sometimes seem to use 'internal' (or 'local') and 'intrinsic' as synonyms, or at least without making clear which of two distinctions are intended (Wilson 1995 is a clear exception). The internal/external distinction is a *spatial* distinction ("where in the world is it?"); the intrinsic/extrinsic distinction is a *modal* distinction ("in which worlds could it be had?")—just as supervenience relations are defined modally, not spatially. An entity can be internal (spatially) to a system and extrinsic (individuated partly in terms of the nature of things external to the system): semantic externalism is one thesis, the extended mind (Clark and Chalmers 1998; Rupert 2004) another. (That said, extrinsic individuation in this context is called externalistic individuation.) In this paper I will assume that the distinction of interest is the intrinsic/extrinsic one unless another interpretation seems warranted.

[4] In this debate, supervenience is strong supervenience, in which there can be no difference in mental-state type without a difference of brain-state type. In global supervenience, there can be no difference in mental type without a difference in physical type, which need not be a brain-state type difference. Wide contents do not supervene strongly, but may supervene globally. The internal nature of strongly supervenient mental states is due to the brain's being inside the skull, not supervenience; if your left hemisphere were removed and kept viable in a vat, your mental states could still strongly supervene on your brain-states but your brain states would not be wholly internal in the usual sense.

[5] I assume the standard, if disputed, view that mental representations in classical computationalism are individuated wide-semantically. Whether they are or ought to be individuated non-semantically or narrow-semantically is a distinct issue, one bound up closely with the ongoing debate over what it means for a system to be computational. I touch on this debate in my concluding remarks.

(GA) (1) The contents of ordinary psychological states do not supervene on intrinsic physical properties. (semantic externalism)

(2) Only intrinsic physical properties, or those that supervene on them, can be causally efficacious.

(3) The contents of ordinary psychological states are not causally efficacious.

I call this the General Argument because it arises in contexts that have no essential connection with classical computationalism or the syntactic/semantic distinction that has been associated with that information-processing paradigm (Kim 1998; Stich 1983). For example, Fodor (1987) claims, as a Metaphysical Principle, that the identity or supervenience of mental properties on 'local microstructure'—specified explicitly as neural structure—is the only plausible materialist explanation of mental causal powers that we have. He then argues that since the contents of ordinary mental states are extrinsic, they do not supervene on neural structure; the possibility of extrinsically individuated brain states is dismissed as 'mad' and 'grotesque'. It follows that semantic properties cannot play causal roles in a materialist theory of mind, computational or otherwise. But seeing how a version of (GA) arises within the computational context is important for two reasons. First, doing so reveals an unnoticed degree of freedom for elaborating the computational model; second, if the problem can be solved within that paradigm, then we will have made room for semantic causal powers within a materialist theory of thinking that once monopolized, and remains a dominant part of, the cognitive scientific landscape.

As a first step towards stating a computational version of (GA), consider Block's expression of a related difficulty, which I'll call Block's Paradox (BP):

(BP) [T]he 'Paradox of the Causal Efficacy of Content' [is] that the following claims all seem to be true, yet incompatible:

1. The intentional content of a thought (or other intentional state) is causally relevant to its behavioral (and other) effects.
2. Intentional content reduces to meanings of internal representations.
3. Internal processors are sensitive to the 'syntactic forms' of internal representations, not their meanings.

> The first claim is meant to be part of the common-sense view of the mind. The third is plausibly taken to be a basic claim of the computer model of the mind, and the second is a useful and plausible way of thinking how common-sense psychology meshes with the computer model. (1990, pp. 138–139)[6]

Block calls this a paradox because computationalism is supposed to be a model of thought, and the contents of our thoughts certainly seem to be causally efficacious: Socrates remained in jail because he believed it would be unjust to

[6] Block does seem to mean "internal" here, not "intrinsic"; at least, I will take his use of "internal" at face value. I do not dispute that the processors and representations are internal, as I am not defending the extended mind thesis.

break the laws of Athens and he desired to act justly. If his beliefs or desires had differed, he might have eagerly accepted Crito's plan of escape. The paradox is that the computer model of the mind seems to rule out (and not merely not account for) the causal-explanatory efficacy of the contents of those very mental states that it seeks to model: beliefs, desires and other ordinary or "folk" psychological states.

It is of course debatable whether computationalism, or any information-processing model of cognition for that matter, must vindicate "folk" psychological mental states, rather than, say, eliminate them (Stich op. cit.; Ramsey, Stich and Garon 1990/1995). So it may seem unwarranted to require that an adequate, or better, computational (or information-processing) model ought to account for the causal efficacy of such states and thereby not give rise to (BP) or an information-processing analogue. This issue is complicated by the fact that while classical computationalism posits a fairly straightforward isomorphism between ordinary thoughts and sentence-like Mentalese symbols, the relation between ordinary thoughts and semantically interpretable activation patterns in connectionist networks is not straightforward (Smolensky 1991/1995; Macdonald and Macdonald 1995; Fodor and Pylyshyn 1988/1995; Fodor and McLaughlin 1990/1995; Ramsey W. 1997). Nevertheless, if we had a computational model for which the paradox did not arise, we would have one less reason to puzzle over the relation between common-sense thoughts and not-so-common-sense mental representations, and one less reason to question either the theoretical adequacy of the model or the causal efficacy of our beliefs. So, *ceteris paribus*, it is preferable to have a computational model that clearly allows for the possibility of semantic causal powers.

(GA) can be located within computationalism if we unpack the assumptions within the third proposition of (BP), which Block identifies as a basic claim of the model. It is a version of what Fodor (1980) calls the Formality Condition: the claim that computational processes only have access to the formal—which he specifies as non-semantic—properties of representations.[7] Although Block does not say so explicitly, it is in virtue of the internal processors' sensitivity to 'syntactic forms' that the causal efficacy of these 'syntactic forms' arises. For if the internal processors are sensitive to the syntax of mental representations, not their semantics, and if this feature of computationalism (given the other assumptions) gives rise to the paradox of semantic causal efficacy, then it must be the case that a mental representation's causally efficacious properties just are those to which the internal processors are sensitive.

So we can extract from (BP) an argument against semantic causal powers (Block's Argument, for convenience) in computational terms, substituting 'syntax' for 'syntactic forms' and 'semantics' for 'meanings':

(BA) (1) Internal processors are sensitive to the syntax of internal representations, not their semantics.

(2) A mental representation's causally efficacious properties are those to which the internal processors are sensitive.

[7] Smith (2002) notes that there are 'almost a dozen' meanings of 'formal' that can be gleaned from the computational literature. I adopt its common usage in this context as 'formal' in the sense of 'syntactic'—which itself is ambiguous. Further discussion follows in the text.

(3) Semantic properties of mental representations are not causally efficacious.

As yet, however, a version of (GA) has not appeared in computational form, for (BA) says nothing (at least not explicitly) about extrinsic properties. It rules out narrowly individuated semantic properties as causal powers just as much as it does externalistically individuated ones, since its validity requires only that syntax have some feature to which internal processors are sensitive that semantic properties of any sort lack. So while (BA) argues against semantic causal powers in specifically computational terms, it does not argue against wide semantic causal powers in particular.

Still, (BA) is still a step in the right direction. For by locating the problem within computationalism, (BA) also invites us to look specifically to computationalism to understand the support for each premise. Note that premise (2) allows that semantic properties can be causally efficacious; it only requires that causally efficacious properties be those to which internal processors are sensitive. So the crucial premise is (1). To accept this premise, we will want to know: what is this feature that syntactic properties have, and that semantic properties lack, that makes only the former those to which internal processors are sensitive? It is not sufficient to support this premise, and hence for (BA) to be valid, merely for syntax and semantics to be distinct. For it is logically possible that they be distinct and yet both possess the feature to which internal processors are sensitive. We need to know what this feature is, according to computationalism, that syntax has and semantics does not.

Perhaps we can glean an answer from Fodor's explanation of how computers explain the causal powers of mind:

> Computers show us how to connect semantical with causal properties for symbols. … You connect the causal properties of a symbol with its semantic properties via its syntax. The syntax of a symbol is one of its second-order physical properties. To a first approximation, we can think of its syntactic structure as an abstract feature of its (geometric or acoustic) *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 imagine symbol tokens interacting causally *in virtue of* their syntactic structures. The syntax of a symbol might determine [its] causes and effects … in much the same way that the geometry of a key determines which locks it will open. (1985/1990, p. 22)

This expansive, if still metaphorical, computational explanation of mental causal powers involves two different kinds of syntactic properties, one of which (in Fodor's terms) is 1st order and the other of which is 2nd order; the term "syntax" is used ambiguously to denote both the reducing (or subvenient) and the reduced (or supervenient) property.[8] The explanation can be made a bit more precise as follows. The causal role of mental representations, conceived of as sentences in the

[8] Properties whose definitions quantify over properties are often referred to as 2nd-order properties (e.g. a functional property is the property of having a property that plays such-and-such causal role). It is often implicit which type of syntax is being referred to. Take the following quotation from Fodor:

Language of Thought (Fodor 1975), is explained in terms of what Devitt (1996) has called the representations' logico-functional and physico-structural properties.[9] The former class of properties—2nd-order syntax—includes, *inter alia*, mental analogues of functional properties in natural language (such as being a noun) and rules for composing complex representations (simple or complex sentences) from simpler representations or concepts. (The syntactic rules are just those which do not advert to semantic relations—e.g., relations such as those between the symbol 'cat' and cats or between the symbol 'cat' and the symbol 'mammal'.) This group of properties—call it syntax$^{\mathbf{L}}$—is posited to explain such cognitive features as the systematicity and productivity of thought. The latter class—1st-order syntax, from which the 2nd-order syntactic$^{\mathbf{L}}$ properties are abstracted—are physico-structural (and perhaps physico-functional) properties of mental representations, such as being a certain neural activation pattern. This class of as-yet-unknown properties—call it syntax$^{\mathbf{P}}$—is posited to explain the causal efficacy of thought. Such properties are held to be analogous to the physical shapes of words or sentences, and are often referred to as 'shape' properties, where 'shape' is a metaphor envisioned (as in the above quotation) in physical (geometric) terms.

One can easily picture, given the 'shape' metaphor, how machines can manipulate symbols in virtue of their syntax$^{\mathbf{P}}$. When it is said (as it is often said) that a computational system's symbols are manipulated (composed, decomposed, substituted, transposed) in virtue of their syntax, and also that a system's syntax comprises the rules for performing these manipulations, what is meant appears to be something like this: each mental representation has syntactic$^{\mathbf{P}}$ properties in virtue of which it can be manipulated via rules specified in terms of its syntactic$^{\mathbf{L}}$ properties, which are realized by, or abstracted from, its syntactic$^{\mathbf{P}}$ properties. For example, all 'cat' symbols have a 'c' shape, an 'a' shape, and a 't' shape, in that order (syntax$^{\mathbf{P}}$), and these 'cat' symbols can become components of noun-phrases and sentences, themselves constructed according to syntactic$^{\mathbf{L}}$ rules, partly in virtue of the fact that all symbols with the syntax$^{\mathbf{P}}$ of 'cat' satisfy, or at least are compatible with, the syntactic$^{\mathbf{L}}$ rules that determine what it is to function as—that is, *be*—a noun or

Footnote 8 continued

> How could a process which, like computation, merely *transforms one symbol into another* guarantee the causal relations between symbols and the world upon which … the meanings of symbols depend? I can … transform the symbol 'dog' any way I like: I can write the word backwards, or cut off the 'd' or replace it by the word 'cat'. (1994, p. 13)

Notice that these sample syntactic transformations are one and all transformations in a symbol's physical ('shape') properties, not transformations from one functionally-defined syntactic category to another (as in nominalization). On the other hand, when Fodor (2000) describes computational psychology as involving 'an innate Turing architecture of syntactically structured mental representations and syntactically driven computational operations defined on these representations', either reading seems apt.

[9] See also Aydede (op. cit.), Devitt (1991), Wilson (1994). Devitt (1996, pp. 258–265) introduces these terms to distinguish between the syntactic properties that are analogous in some way to geometric structure (as in Fodor's lock and key metaphor; see also the previous footnote) and syntactic properties that a symbol may have in virtue of its role within a cognitive structure (as in the syntactic categories of linguistics). Any additional distinctions within the category of syntax will not matter for the purposes of my argument.

noun-phrase in Mentalese. One way to think about this explanation is to say that syntactic$^{\mathbf{L}}$ properties are causal-functional roles and syntactic$^{\mathbf{P}}$ properties are the realizers of those roles. But the arguments against semantic causal efficacy are usually worded in terms of supervenience (or lack thereof), not realization, so I will simply assume that the above explanation holds for whatever syntax$^{\mathbf{L}}$–syntax$^{\mathbf{P}}$ relation is posited.

Now the differences between syntax$^{\mathbf{L}}$ and syntax$^{\mathbf{P}}$ may not matter greatly in some or even most contexts, especially if syntax$^{\mathbf{L}}$ supervenes on syntax$^{\mathbf{P}}$; this would abet the ambiguous use of 'syntax'. However, in the context of the problem of the causal efficacy of content, it is important to distinguish the two kinds.

First, as noted, the claim that mental representations are manipulated just in virtue of their syntax—that internal processors are sensitive only to syntax—is held to be basic to the computational model (as we saw in (BP) above). But the model's more complete explanation of the causal powers of syntax involves two kinds of syntax and a relation between them. So do both types of syntax possess that feature to which internal processors are sensitive? If not—if, say, only syntax$^{\mathbf{P}}$ has this feature—then (BA) would rule out syntactic$^{\mathbf{L}}$ causal powers. So let us assume they both do, for the sake of argument. But do they possess it independently or does one type—presumably syntax$^{\mathbf{L}}$—inherit or derive it from the other? Also for the sake of argument, we may assume the following answer—one that many likely hold implicitly, although Fodor (1987), cited above, seems explicit enough: The causal efficacy of syntax$^{\mathbf{L}}$ lies in its presumed supervenience on causally efficacious neural structures (syntax$^{\mathbf{P}}$) of some sort, for it is in virtue of this relation that syntax$^{\mathbf{L}}$ possesses the feature that guarantees its manipulability by internal processors (either by supervening or by inheriting it because it supervenes).[10] We may then infer, to complete the explanation, that the requisite feature to which internal processors are sensitive is specifiable in syntactic$^{\mathbf{P}}$ terms. But this explanation makes it unclear why narrow semantic properties cannot be causally efficacious, given the way (BA) is stated. In short, for (BA) to have any genuine force, the intended, if implicit, target of premise (1) must be extrinsic semantic properties in particular, under threat of depriving the computational model of any explanation of mental causal powers at all.

Second, this explanation of the causal powers of syntax$^{\mathbf{L}}$ reveals that while classical computationalists consider their theory to be medium-independent—that is, independent of any claims about implementation-level properties—they are in fact smuggling in an implementation-level assumption via their explanation of mental causation. The assumption in question goes beyond the claim, widely associated with classical computationalism, that there really are mental representations—that intentional realism is true. It is introduced by assuming that syntax$^{\mathbf{L}}$ is intrinsic and by explaining the causal efficacy of syntax$^{\mathbf{L}}$ in terms of its

[10] Thus, when Fodor (2000) claims that "cognitive processes are causal only if they are syntactic", he could mean that cognitive processes are causal only if they are neural or only if they supervene on the neural.

supervenience on syntax$^{\mathbf{P}}$, which also must be intrinsic; semantics, which is not syntax, is extrinsic. As Fodor puts it, using the analogy of being dollar-looking (syntax) and being a dollar (semantics):

> Being a dollar is an extrinsic (causal/historical) property; whether a thing has it depends essentially upon its etiology. Being dollar-looking, by contrast, is a matter of a thing's *internal* properties (it's a property of a thing's *appearance*, whatever, exactly, that may mean.). (1994, pp. 18–19)

Now the computational model can certainly take for granted that at least some syntactic$^{\mathbf{P}}$ properties must be intrinsic if syntax$^{\mathbf{L}}$, which is intrinsic, is causally efficacious. But that *all* syntactic$^{\mathbf{P}}$ properties are intrinsic is an implementation-level assumption that rules out extrinsically individuated realizers in general for any cognitive systems to which the computational model might apply. It is this general assumption about syntax$^{\mathbf{P}}$, which is gratuitous to the computational model, that has no empirical warrant. Or so I will argue.

Before proceeding, I should also note that it is not entirely clear whether the syntactic/semantic distinction in computationalism was always, albeit implicitly, considered a case of the intrinsic/extrinsic distinction. (Of course, as Fodor (1994) shows, semantic externalism had clear implications anyway for computationalism considered as a vindication of intentional realism.) In theory, the model is consistent with a syntactic/semantic distinction that involves only narrow content. Its canonical explanation of truth-preserving thinking processes – the idea that semantic properties are "mirrored" or "mimicked" by syntax, that the brain is a syntactic engine driving a semantic engine—does not require assimilating the syntactic/semantic distinction to the intrinsic/extrinsic distinction. But in practice, things do appear otherwise. It is basic to the computational model of mind, as standardly elaborated, that a mental computational system is a *formal* system. This formality condition is often explained in this context by saying, *inter alia*, that the system operates independently of its semantics, or by saying that it operates independently of what is going on outside the system. But "independently of its semantics" and "independently of what is going on outside the system" coincide only if semantics is extrinsic (the intrinsicness of syntax being assumed). So to the extent that mental computational systems are formal systems, and given the usual ways of explaining formality in this context, the syntactic/semantic distinction seems to have been considered a case of the intrinsic/extrinsic distinction for a while; semantic externalism did not change, but rather reinforced, this established practice. In these terms, internal processors would be sensitive to syntax, but not semantics, because syntax is part of, or within, the formal system and semantics is not.[11]

It should now be clear that at least some arguments against semantic causal efficacy within classical computationalism, including (BA), are in fact versions of

[11] Arguably, this explains some of the force of Searle's Chinese room argument, in which manipulation of symbols by their syntax does not suffice for understanding the meaning of the symbols. If it were *obvious* that manipulation of the symbols according to the rules sufficed to give the symbols meaning (setting aside the separate issue of understanding their meaning), the argument would lose a great deal of its bite.

(GA). If we combine the computational language of (BA) with the externalistic language of (GA), we can state a computational version of the argument against wide semantic causal powers, which we may call the Computational Argument:

(CA) (1) Semantics does not supervene on syntax$^{\mathbf{P}}$.

(2) Only properties to which a computational system's internal processors are sensitive (syntax$^{\mathbf{P}}$ or those that supervene on syntax$^{\mathbf{P}}$, such as syntax$^{\mathbf{L}}$) are causally efficacious.

(3) Semantic properties of mental representations are not causally efficacious.

The difficulty that the computational model of mind faces in accounting for semantic causal efficacy then amounts to this: The syntactic/semantic distinction in computationalism is a case of the intrinsic/extrinsic distinction; and only syntactic properties can be causally efficacious because only they supervene on those intrinsic physical properties (neural structure, in our case) to which internal processing mechanisms are sensitive.[12]

The distinction between types of syntax makes it possible to extend the scope of (CA) to at least some non-classical information-processing architectures.[13] For example, connectionist neural networks may or may not have syntax$^{\mathbf{L}}$, depending in part on whether representations in such networks have constituent structure (itself a hotly debated issue; see Macdonald and Macdonald 1995; Churchland 1998; Aydede 1997). But they certainly do rely on syntax$^{\mathbf{P}}$, since neural-structural properties are among those that such architectures specifically seek to model. Whether a version of (CA) arises within connectionist models depends on whether neural networks are individuated intrinsically or extrinsically. This appears to be another open question. But if it is assumed within at least some connectionist models that the neural structures being modeled are all intrinsic, then a version of (CA) would appear to apply to those connectionist networks as well.

Extrinsic Causal Powers

I propose to agree that the explanation of the causal efficacy of the syntactic$^{\mathbf{L}}$ and syntactic$^{\mathbf{P}}$ properties of symbols in computational models is not negotiable. Mental representations *are* manipulated, ultimately, in virtue of their neural-structural properties—whatever those are exactly. And there lies the rub. What I will argue is that we cannot determine a priori how syntax$^{\mathbf{P}}$ will be individuated in our best neuroscientific theories, and that therefore our best computational psychology cannot, and should not, presume anything about the nature of these implementation-level properties. For the assumption that all neural structural properties are or will be intrinsically individuated, which amounts to a suppressed premise in (CA) as

[12] Again, I set aside problems raised for supervenient causation in general, which affect "folk"-psychology-vindicating, narrow-content and purely syntactic computational theories alike.

[13] If dynamical systems models (Van Gelder 1995) do not involve a notion of representational states, such accounts of cognition would be immune to the problem. But it is doubtful that they eschew representation altogether (Bechtel 1998; Clark 1998; Chemero 2000).

well as (GA), is scientifically unjustified. If this is right, then the first premise of (CA) also will be unjustified, and (CA) will give us no reason to think that semantic properties are causally inert or that the computational model of mind cannot accommodate the possibility of semantic causal powers. For even if premise (2) is true, if premise (1) is false we have no reason to think that semantic properties are not among those to which internal processing mechanisms are sensitive. At the very least, my argument shifts the burden of proof to those who think that premise (1) already has support sufficient to make the argument sound.[14] It does not.

The first reason to doubt the intrinsicness of all syntactic$^{\mathbf{P}}$ properties involves a popular, if largely implicit, way of unpacking the metaphor of 'shape' or 'form' as these terms have been used to refer to the structure of mental representations. The unpacking goes by way of analogy to the concept of molecular structure, both in terms of its nature and causal-explanatory role. In 'folk' molecular theory, based in classical chemistry, you take the atoms and glue them together with electron bonds and get an entity that plays a robust causal role in virtue of the parts and their arrangement.[15] Similarly, at least in classical computationalism, you take the representation's parts and put them together (somehow) and get something that plays a robust causal role in virtue of those parts and their arrangement (Fodor 2000). These components—often, and not by accident, called atomic and molecular symbols—retain their identities within the larger unit, just as atoms retain their identities within a molecule according to the classical molecular model.

With this analogy in the background, it is not surprising that syntax$^{\mathbf{P}}$ would be assumed to be intrinsic, since molecular structure *is* intrinsic—*in the classical molecular model*. One might say that just as the intrinsicness of molecular structure is a dogma of the classical molecular sciences, the intrinsicness of syntax$^{\mathbf{P}}$ is a dogma of classical computationalism (and likely of alternative models of cognition as well). Moreover, chemistry's causal-explanatory taxonomy of chemical species is based on structural differences, which were posited to explain observed causal differences that could not be explained merely by reference to atomic composition (Le Poidevin 2000). It seems entirely reasonable to aim for a similar causal-explanatory taxonomy in cognitive science.

This is where the metaphorical explanation of causal powers in computationalism starts to fall apart. For contemporary physical chemistry seems to show that molecular structure may be extrinsic even while its causal role remains unquestioned. The classical molecular model holds that a molecule is a quasi-rigid

[14] We can also restate (CA) more explicitly as follows:

(1a) Syntax$^{\mathbf{P}}$ is intrinsic.

(1b) Semantics does not supervene on syntax$^{\mathbf{P}}$.

(2) Only properties to which a computational system's internal processors are sensitive (syntax$^{\mathbf{P}}$ or those that supervene on syntax$^{\mathbf{P}}$, such as syntax$^{\mathbf{L}}$) are causally efficacious.

(3) Semantic properties of mental representations are not causally efficacious.

The argument is valid, but unsound: (1a) may be false, and so (1b) may also be false.

[15] Ramsey J. (1997) coins the phrase "folk molecular theory" to describe the ball-and-stick conception of molecules in classical chemistry. This model has already been revised partly but significantly in the light of quantum mechanics; e.g. electrons no longer occupy orbits, but "occupy" orbitals, which are probability distributions for position measurements. Among others, Woolley (1978, 1985), discussed in the text, urges chemists to abandon the classical model entirely.

assembly of nuclei connected by electron bonds, whose structure exists even if the molecule were completely alone in an empty universe. Some theoretical chemists argue that this classical model is inconsistent with basic principles of quantum mechanics, and that in a consistent model a molecule's structure is extrinsic: they hold that if a molecule has a *classically describable* structure, then it exists in an otherwise non-empty world—one that, for example, contains other molecules or a vacuum electromagnetic field, with which the molecule is interacting.[16]

There are several distinct arguments to this conclusion in the chemistry literature; a seminal argument by Woolley (1978) can be summarized as follows. (1) In the classical molecular model, nuclei in a molecule have distinct fixed positions over time, whether or not the molecule is being observed or otherwise interacting with its environment. It follows that identical particles in a classical molecule can have distinct physical descriptions at any time. For example, a centrally located carbon nucleus will have smaller angular momentum than a peripheral carbon nucleus in a rotating classical molecule; they are dynamically inequivalent. Since these dynamical differences stem from their distinct fixed positions, if the classical model is correct then these physical differences should persist even if the rotating molecule is isolated.[17] (2) It is a basic principle of quantum mechanics that identical particles in a system cannot have distinct physical descriptions; they are completely indistinguishable and interchangeable. For example, if one measures the position of a carbon nucleus in a molecule which has at least two carbon atoms, it cannot matter which of the carbon nuclei is observed as long as one of them is detected within the expected range of positions. (3) Identical nuclei in observed molecules typically are distinguishable in their behavior, as the classical model predicts. But can the same be said of the nuclei of molecules when they are not being observed? If the classical model is correct, yes; but this claim has no quantum mechanical justification. (4) However, there is an alternative molecular model that is consistent with both observation and quantum mechanics. We can explain the observed behavioral differences among identical nuclei as arising from the molecule's interaction with the environment (including but not limited to interaction with our measuring instruments); without this interaction, however, identical nuclei will be indistinguishable, as quantum mechanics claims. This explanation requires us to give up the classical model, however, since it is essential to that model that nuclei occupy distinct fixed positions in the molecule over time. We must also give up the classical assumption that a molecule's structure is intrinsic. For if the alternative model is correct, if a molecule has a classically describable structure, then it must be interacting with

[16] Woolley (1978) is the *locus classicus* of the debate, reviewed in Weininger (1984), Ramsey (2000) provides a distinct argument to a similar conclusion. Claverie and Diner (1980), assessing and building on Woolley's paper, suggest that molecular structure requires interaction with other molecules or a vacuum electromagnetic field. A molecule of structure type S might then be partly individuated in terms of its relations to the requisite external entities or conditions.

[17] An isolated system is one which does not exchange either matter or energy with its environment (see Liboff 2003).

[18] An objection to this claim is to question the type of necessity involved. What is nomologically necessary need not be metaphysically necessary, so (the objection goes) these arguments do not show that

something in its world, which must therefore be non-empty. So its having a classically describable structure entails that its world has a certain contingent character.[18]

It is an open question in chemistry what the correct molecular model is, and thus whether molecular structure is extrinsic or intrinsic. But this openness is sufficient to show that there can be no a priori justification that any structural property is intrinsic. This would be the case even if we retained the analogy in computationalism between syntactic structure and molecular structure. We should not even assume a priori that a reduction of syntactic$^{\mathbf{P}}$ structures to even lower level *intrinsic* properties is in the offing.

A second reason to question the intrinsicness of all syntactic$^{\mathbf{P}}$ properties stems from cognitive neuroscience. Current cognitive neuroscientific research is dominated by localization efforts that involve trying to map particular cognitive capacities to particular brain areas or neural circuits (a "brain state" would be activity in a brain area). An adequate taxonomy of brain areas must be such that resulting generalizations can subsume activity (brain states) in brains with differences in neural connectivity and neuroanatomical and neurophysiological differences, even within species. As a result, brain areas are not individuated just by neurophysiological or neuroanatomical criteria, although Brodmann's 1909 standard brain map, which distinguishes 47 different areas in the human brain based on cytoarchitectonic differences, is often used as a starting point. Researchers also identify brain areas partly by the type of information they process or the role they appear to play in processing or in contributing to behavior (Bechtel and Mundale 1999; Kosslyn 1999). Since the functions being used to help individuate brain areas are drawn from ordinary psychological categories, and these categories are at least in part widely individuated, at least some brain areas, and hence brain states, are already individuated widely. There is of course nothing sacrosanct about these psychological categories, but those who think they are scientifically inadequate cannot look to cognitive neuroscience to ratify their view.

Familiar examples of this complex methodology include primary visual and auditory cortex, whose names explicitly indicate their normal cognitive function, as well as Broca's area in left posterior frontal cortex (approximately Brodmann's area 44), implicated in speech production, and Wernicke's area in the superior temporal gyrus posterior to primary auditory cortex, implicated in speech comprehension (but see Dronkers et al. 2000).[19] Because visual perception has been most amenable to neuroscientific research (much of which involves animal models), visual cortex (V1 and nearby regions in the occipital lobe) in particular has been extensively studied

Footnote 18 continued
molecular structure is extrinsic. For we can easily imagine a classically structured molecule in an otherwise featureless world. I reply that the intuitions on which the objection essentially relies are deeply theory-laden—classical-molecular-theory-laden, to be precise. Thus, the fact that we can imagine classically structured molecules in empty universes is question-begging and to no avail insofar as these intuitions alone are taken as evidence of the nature of molecular structure. Before we knew the nature of water, we could easily have imagined it being other than H_2O—and probably did.

[19] The neuroscientific results summarized in this section should not be considered definitive; moreover, localization of function in animal models (often, the macaque) may not map precisely to the same areas in human brains. Further refinements use the same psychological-functional methodology, however.

and subdivided into areas believed responsible for more specific psychological functions. (For higher cognitive functions, many different areas are implicated, and much remains to be discovered about their realization.) Subdivisions in early vision include V4 for color and V5 for motion (Felleman and Van Essen 1991). Subsequent visual processing has been divided into two pathways: the ventral 'what' pathway (from V1 towards the temporal lobes) for recognizing kinds of objects, and the dorsal 'where' pathway (from V1 towards parietal areas) for recognizing spatial location of objects (Ungerleider and Mishkin 1982; Haxby et al. 1991; Milner and Goodale 1995; Borowsky et al. 2005). Even more specifically, a region in the fusiform gyrus in occipitotemporal cortex (part of the ventral pathway, and somewhat lateralized to the right) has been identified that responds selectively to faces, called (appropriately enough) the fusiform face area (Kanwisher et al. 1997). The dual-pathway model has also been extended to working memory, a postulated complex psychological mechanism for actively maintaining information briefly for immediate use, whose neural substrate appears distributed in a network of regions, including visual and frontal cortex (Ungerleider et al. 1998). The reliance on cognitive function to help individuate brain areas does not disappear as the functions get more and more specific.

This method of individuating brain areas and states partly by cognitive function is unlikely to be merely heuristic—a stopgap measure until we know more about the brain—due to neural plasticity during an individual's development and, to an unknown extent, over an individual's lifetime. To a first approximation, neural plasticity refers to the brain's ability to change: physically, such as in the development of synaptic connections in (at least) early childhood, and functionally, such as when visual areas are (apparently) recruited for auditory processing in blind people. (Plasticity allows that the same mental state may be token identical to brain states of different kinds in different people or in the same person at different times.) Plasticity motivates continued use of cognitive function or behavior in the individuation of brain states because it forces us to pick out *significant* changes in neural structure; not just any neural structural change matters for taxonomic purposes. But significant neural-structural changes will be those which result in observable cognitive and behavioral differences, many of which are typed partly by reference to entities in or other features of the agent's environment.[20] So standard neuroscientific individuation methods, plus what we know about neural plasticity, give us no reason to assume that the syntactic$^{\mathbf{P}}$ structures that realize mental representations will all be intrinsic; quite the contrary.

Our current state of knowledge of the cognitive architecture of the brain, particularly for higher cognitive functions, is still relatively coarse-grained. In particular, we cannot yet identify the neural structures underlying types of mental representations paradigmatic of the Language of Thought, such as a belief about one's mother or a modus ponens inference. It follows that the concept of syntax$^{\mathbf{P}}$ is a

[20] The type of information a brain area processes may depend in part on the processing context in which it occurs—for example, which other brain areas are simultaneously active, or what kinds of processing immediately preceded it. Internal context-dependence does not result in extrinsic individuation in the sense that matters here; but internal processing contexts may be individuated extrinsically in the relevant sense.

placeholder for a type of property (or properties) that we expect will play a causal-explanatory role in neuroscience similar to that played by molecular structure in chemistry. At the very least, of course, computationalists should be genuinely neutral about how syntax$^{\mathbf{P}}$ will be individuated. But based on what we know from neuroscience, it is also reasonable to expect that at least some of the specialized cortical structures that realize representations in human cognitive systems will be individuated widely.

It follows that the background assumption that the syntactic$^{\mathbf{P}}$ structures of mental representations are all intrinsic is unjustified. We *can* assume that whatever syntactic$^{\mathbf{P}}$ properties turn out to be, they must be consistent with materialism and mechanism: they must be such that internal processors are sensitive to them. But this criterion does not rule out extrinsically individuated syntactic$^{\mathbf{P}}$ properties. It is consistent with computationalism that the structural properties of mental representations may be individuated such that they differ in type if the environment is relevantly different. *They would still be manipulable by the system*, and that is all that its mechanistic explanation of thought requires.[21]

An obvious, and *prima facie* strong, objection to the possibility of extrinsically typed syntactic$^{\mathbf{P}}$ properties is that such properties will not belong in a *causal* taxonomy. For an extrinsic taxonomy of this sort would indeed allow that two physically identical syntactic$^{\mathbf{P}}$ states might be type-distinct. This possibility clashes with the very deeply-seated intuition that these states must play the same causal roles: surely the system will manipulate these states in exactly the same way, even if their structural properties are individuated extrinsically and they therefore count as type-distinct. If it is possible to identify a single *intuitive* point around which this debate turns, this is it.

But this intuition too can be questioned, for there is no a priori justification for the claim that internally identical structural properties will play the same causal roles. This claim is undermined by the case of chiral molecules—mirror-image molecules that are the same in internal composition and structure, but differ in being right- or left-handed. Such molecules, which are pervasive in, although not limited to, organic chemistry, systematically differ in their causal powers in chiral contexts—that is, in interactions with other chiral molecules (Brown 1990; Crossley 1995; Hicks 2002; Jacques 1993). Some differ in potency in producing the same effect, some have qualitatively different effects, and some work together in a mixture to contribute to a single desired effect. For example, the left-handed form of ketamine (Ketalar) is an anesthetic, the right-handed form is an excitant that can cause psychic disorders. It is due to these striking causal differences and the

[21] 'But the external relations are in the wrong place! Nothing *outside* the system can matter to how it functions *inside*!' On the one hand, if the objection confuses the internal/external and intrinsic/extrinsic distinctions (see fn. 3), it is no objection at all: even if "nothing outside matters to how it functions inside", what's inside could still be extrinsically individuated. On the other hand, if the shape (syntax) of a representation is partly but essentially determined by, and individuated in terms of, features external to the system in which it is manipulated, then the objection would fail. I discuss this latter possibility below.

[22] The most famous case of chirality is thalidomide, whose toxicity was traced initially to the left-handed enantiomer. Although further studies have conflicted with this result, the notoriety of the drug drove home the point that chirality could not be ignored in drug design. As Jacques (1993, pp. 106–107) notes, 'it is neither inconceivable nor uncommon for one and the same molecule (with right-and left-handed forms) to

pervasiveness of chirality in living organisms that chemists taxonomize chemical species by handedness and not just by internally-specified structure.[22] As things have turned out, the latter method is not fine-grained enough to track causal differences.

Now if molecular structure in general may be extrinsic, then ipso facto the structures of chiral molecules may also be extrinsic. But there are independent reasons to think that handedness is extrinsic (Nerlich 1995; Le Poidevin 1994, 2000). Chiral molecules are mirror images, or enantiomers, which cannot be superimposed, or rotated rigidly to coincide with each other, in a given space. By analogy, the letters 'b' and 'd' are mirror images that cannot be superimposed in a two-dimensional space. But they can be in a three-dimensional space. So these letters are enantiomorphic (handed) only in regions of certain spatial geometry. Thus, if they are enantiomorphic, they must be in a space with a certain contingent character: their being mirror images is not independent of the nature and contents of the space they are in. Similarly, being a left-handed molecule entails existing in a particular kind of space: one which contains nonsuperimposable right-handed molecules of the same internal configuration, even if it is conventional which counts as right and which as left, and even if some molecules can flip from one form to the other. This makes being left- or right-handed an extrinsic property of a molecule. And since handededness makes a causal difference, a taxonomy based on handedness is a causal-power taxonomy.[23]

One might respond that chirality at the molecular level would make no difference at the level of those objects whose behavior the computational model aims to explain. One would almost certainly be wrong. We cannot, at this point, clone a mirror-image mouse, because DNA is chiral; we would have to build right-handed DNA in order to create such a creature, and this we cannot (yet) do. But if we could, we would have every reason to expect, based on what we already know, that substances that might nourish or cure a normal mouse might be toxic to a mirror-image mouse. Two colonies of mice, one right-handed and one left-handed, very likely would respond in very different ways to the same environment; one mouse's thoughts of food would be another's thoughts of poison. (In other words, the standard Twin Earth thought experiments do not individuate molecular duplicates finely enough.) The fact that DNA is chiral and that the structure of DNA plays a

Footnote 22 continued
serve multiple purposes: as an antirheumatic, a pain-killer, an antipyretic, and so on.' In many cases, we don't yet know what different effects there are. For example, positron emission tomography (PET) brain scans show that the right-handed form of Ritalin concentrates in the striatum, whereas its mirror-image distributes non-specifically in the brain (Hicks 2002). Also, since the metabolic products of the two forms may differ, the long term effects may also differ. Chirality also runs deep: it is a feature of the rotational motion of subatomic particles, e.g. whether they spin to the left or the right (Latal 1991).

[23] In a world in which there existed just one of a chiral pair, classification by internally-specified structure alone would be coextensive with classification by handedness. This is, in effect, our world now at the level of organisms, or our world prior to the discovery of chirality. In such a world we have no reason to care or know about causal differences due to chirality. It doesn't follow that there would not *be* such differences were the missing form introduced; to the contrary, the discovery of chirality amounted to the "introduction" of some of the missing forms in our world.

causal role in biology makes the possibility of higher-level causal differences due to chirality extremely strong.

A second strong prima facie objection would grant the above points about molecules and syntax$^{\mathbf{P}}$, yet still find the argument lacking an essential element. Even if some syntactic$^{\mathbf{P}}$ properties were extrinsic, the objection would go, the complete computational explanation of mental causation requires higher-level properties—syntax$^{\mathbf{L}}$ and semantics alike—to supervene on syntax$^{\mathbf{P}}$. Yet no reason has been given to think that wide semantic properties will supervene on wide syntactic$^{\mathbf{P}}$ properties. The classic thought experiments from Burge and Putnam show that environments that ensure sameness of molecular types are sufficiently different to yield differences in semantic types—in other words, that the environmental features that fix semantic properties do not fix molecular properties. So even if some syntactic$^{\mathbf{P}}$ properties were individuated externalistically, why think these properties will be fixed by the same environmental features that fix semantic properties, such that we have reason to believe that the latter supervene on the former?

Why indeed? But notice, however, that the parallel question may be asked of syntax$^{\mathbf{L}}$ and that part of syntax$^{\mathbf{P}}$ that (by assumption) is individuated individualistically: why think the former will supervene on the latter? Being-inside-the-headness by itself no more guarantees supervenience between distinct properties than being-outside-the-headness rules it out. By analogy: even if having a mass of 2 g and being red are both intrinsic, there is no reason to think a difference in color necessitates a difference in mass, or vice versa. (In fact, differences in taxonomic grain sizes make it quite plausible that at least some syntactic$^{\mathbf{L}}$ type differences will make no syntactic$^{\mathbf{P}}$ type difference at all.) More generally, if P is extrinsic and Q is intrinsic, we can conclude a priori that P will not supervene on Q, as the classic externalist thought-experiments show; but if P and Q are both intrinsic, or both extrinsic, nothing follows a priori about their relation from this fact alone. In short, the objection demands an a priori answer to a question that leaves syntax$^{\mathbf{L}}$ just as badly off. There is no a priori guarantee of supervenience in either case.

Notice, moreover, that anyone who considers herself a naturalist about content—anyone who thinks it is possible to explain intentionality in non-intentional terms—is *already* committed to the possibility of a theory in which semantic properties and relations are ultimately explained in terms of physical properties and relations. Whether these sets of properties are individuated intrinsically or extrinsically is not essential to her position; supervenience is. So to the extent that semantic naturalism is defensible, so is the possibility of supervenience between semantics and syntax$^{\mathbf{P}}$. And to the extent that computationalism is a materialist theory of mind, it makes little sense to reject the possibility of a naturalistic explanation of semantics while adopting computationalism.

In sum, considerations from chemistry and neuroscience show that we have no good scientific justification for the a priori assumption that the neural structures that underlie the mental processes and representations posited in a computational model must be intrinsic. This means that (CA) gives us no good reason to think that semantics is causally inert in a classical computational model of mind. Even if we accept the second premise—that supervenience on or identity with syntactic$^{\mathbf{P}}$

properties is a necessary condition for causal efficacy in a computational model—we have no reason to think that semantics cannot meet this condition, since the first premise may well be false. The essential dependence of semantic properties on the way things are in an agent's environment is fully compatible with the claim that semantic properties may be identified with or supervene on some syntactic$^{\mathbf{P}}$ properties of mental representations, since the latter may also be extrinsic. The problem of externalism is an artifact of the assumption that all syntactic properties—syntax$^{\mathbf{L}}$ and syntax$^{\mathbf{P}}$ alike—are intrinsic. This assumption is neither justified nor essential to the computational model of mind.

It follows that within the context of computationalism the burden of proof regarding semantic causal efficacy now falls on those who deny it. For as Block's Paradox illustrates, prima facie semantic properties *are* causally efficacious; the problem has always been to reconcile this datum with the computational model. If, as I have argued, the computational model is indeed compatible with this claim, then those who would *still* deny semantic causal efficacy must find some other reason to back their claim—and they must do so in a way that does not undermine the accepted explanation of the causal role of syntax.

Concluding Remarks

It is worth recalling at this juncture a plea once made by Haugeland regarding the explanatory possibilities of classical AI:

> None of [his introductory discussion] proves that computer systems *can* be truly creative, free, or artistic. All it shows is that our initial intuitions to the contrary are not trustworthy, no matter how compelling they seem at first. If you're sitting there muttering: "Yes, yes, but I *know* they can't; they just couldn't," then you've missed the point. Nobody knows. Like all fundamental questions in cognitive science, this one awaits the outcome of a great deal more hard research. (1985, p. 12)

If we substitute 'semantic properties' for 'computer systems', and 'causally efficacious' for 'creative, free, or artistic', Haugeland has neatly summarized the moral of my argument. In this case, the broad claims that are typically made about the causal role of mental symbols in computationalism do *not* require us to assume anything about them other than that they must be accessible to mechanistic manipulation. *Structural* properties may be essential to satisfy this constraint, but the assumption that such properties are all *intrinsic* is not. The latter assumption

[24] Haugeland's plea is itself reminiscent of Woolley's concluding remarks (1978, p. 1078): 'Naturally I recognize that since much of the quantum theory appropriate here has yet to be worked out, experimentalists will regard this critique of molecular theory as being of little direct help to them; equally one hopes that these experiments are still performed to investigate primarily the properties of matter, which need not be the same as "the determination of very precise molecular structures", and without some agreement about the theoretical principles that will be useful in the last quarter of the twentieth-century, this area of physical chemistry may well find itself up a blind alley as its traditional molecular models become less and less relevant to contemporary experiments.'

may be no more fruitful for cognitive scientific research than its molecular analogue in chemistry.[24] We should consider giving it up.

Nevertheless, it may seem far from clear how classical computationalism can accommodate the possibility of mental representations manipulated in virtue of their semantics and still remain a computational model. Isn't the genius of computationalism the idea that thought can be explained mechanistically by means of a merely syntax-manipulating machine whose input-output relations nevertheless make semantic sense? Isn't Fodor's Formality Condition is a *sine qua non* of any model that claims to be computational? But there is no trick here. I accepted this condition, understood (when disambiguated) as the claim that mental representations are manipulated in virtue of their structural properties, whatever these may often be *called*.[25] My solution merely involved pointing out its consistency with widely individuated structural properties, and arguing that the wide individuation of structural properties that figure in causal explanation is an empirical possibility. If "structure" is *defined* as "syntax" (and everyone knows syntax isn't semantics!), then in this sense any "violation" of the Formality Condition on my part amounts to a verbal faux-pas.

A deeper discomfort, however, stems from fundamental disagreement over the nature of computation, which in turn drives debates over the proper individuation of the states in a computational model of mind and over what role, if any, content can play in such a model.[26] For some (e.g., Egan 1995; Piccinini 2006), the term "computational" is defined by the theory of computability, and many of those who do not tie it directly to the mathematical conception nevertheless interpret the Formality Condition in its light. The states of a computational system, as mathematically defined, are essentially not semantic, and therefore not essentially semantic, because the mathematical theory individuates them that way. Others explicitly deny that computability theory determines what it is to be "computational". For example, Churchland and Sejnowski (1992) characterize a computational system as one in which "the physical system's states can be seen as representing states of some other systems, where transitions between states can be explained as operations on the representations".[27] I have no intention of settling this debate here, although I will note that the causal history of a theory does not determine its content: although the classical computational model of mind was inspired by computers, theories often outgrow the analogies that motivate and guide their initial development.[28] However, from the perspective of this debate over fundamentals, discomfort with my solution can be understood as a request for some indication of the sorts of modifications I have in mind

[25] See, e.g., Egan 1995, p. 181 and fn. 1.

[26] Arguments for narrow individuation (semantic or non-semantic) include Stich (op. cit.), Segal (1989, 1991), Egan (1992, 1999), Piccinini (2006); for externalistic semantic individuation, see Burge (1986), Davies (1991), Peacocke (1994). I hasten to add that a significant portion of the debate on individuation in computationalism involves perceptual states in Marr's computational theory of vision, and so is not directly relevant to this paper. But it is probably safe to assume that essential features of that debate could be transposed into the case of the individuation of concepts and sentences in the Language of Thought.

[27] These two options are by no means exhaustive. For example, Smith (op. cit.) denies that the term "computational" picks out a natural kind, while Copeland (2002) argues that many of the central claims attributed to the founders of computability theory are misinterpretations.

[28] Thagard (2002) provides independent reasons against drawing the computer/mind analogy too closely.

that do not amount to simply abandoning any recognizable form of computationalism. The following remarks are intended to show how this is possible; I leave the development of these suggestions to another occasion.

I will assume that a symbol-manipulating system must obey the Formality Condition (as a minimum) in order to count as computational. I suggest that a system satisfies this condition if it is a formal symbol-system; it need not also be a formal-symbol system. A symbol-system is formal if its symbols are manipulated by rules defined purely in terms of properties and relations internal to the system (i.e., its syntax$^{\mathbf{L}}$).[29] A symbol is formal if it has its structure essentially but not its content, and informal if it has its content essentially but not its structure.[30] For example, a formal language is both a formal symbol-system and a formal-symbol system. This distinction between formal and non-formal symbols can be exploited to resolve the issue of semantic causal efficacy within a computational model of mind.[31] A version of the classical computational model of mind in which the mental representations are formal symbols might be called Formal Language of Thought (F-LOT) computationalism, while a version of the model in which mental representations are informal symbols might be called Natural Language of Thought (N-LOT) computationalism. In these terms, to now we have assumed, in effect, that classical computationalism *just is* F-LOT computationalism. I suggest that this assumption, not the Formality Condition, is the source of the model's problem with semantic causal efficacy, and that we can abandon the formal-symbol assumption without impugning the model's computational status.

N-LOT computationalism can allow for semantic causal efficacy if the structures of its informal symbols (not the roles that *already existing* internal structures are recruited to play, as in Dretske 1988) are individuated partly but essentially in terms of the system's external relations, and if the semantic properties of these symbols supervene on these extrinsically individuated structures. In this model, there may be no justified distinction between the external relations which partly determine a mental representation's structure and those upon which its semantics supervenes in an adequate naturalized semantics. To borrow Rey's (1997) terminology,

[29] I follow Haugeland's (1997) characterization of formal systems as (a) token-manipulation systems that are (b) digital and (c) medium-independent. A token-manipulation system is self-contained in that the rules specify only the next allowable arrangements of the tokens only in terms of the current arrangement (which may be the initial state, or a subsequent state determined by the rules).

[30] Anything that is a symbol must represent something even if it is not essentially characterized by its content; otherwise every object is actually a symbol, rather than merely potentially one. Just manipulating a set of objects by formal rules does not entail that the objects are symbols: Beckett's (1955/1965, pp. 69–74) character Molloy provides a vivid example. The fact that we often call the objects in many mere-object-manipulating systems "uninterpreted symbols" just reflects our habit of assuming, based on past practices of providing interpretations, that they are or soon will be symbols (e.g., numerals, the "predicates" given when constructing a formal language). Suppose I set up a formal system with fruit and spices; using a salt shaker as a pivot, I decree that if one side of the salt shaker is flanked with two peaches separated by a stick of cinnamon, then an apple may be placed on the opposite side of the salt shaker, but not a banana. Just so, I can manipulate the mere-objects '3', '3', '+' and so on, but the mere-object '3', for example, is no more a symbol than my peaches or Molloy's sucking stones. See Crane (op. cit.) for a similar view.

[31] Wilson (2004, pp. 149–150) draws a similar distinction between "conventional" and "natural" symbols, without further elaboration.

Brentano's problem (explaining how mental representations represent anything) and Descartes' problem (giving a mechanistic explanation of rational thought) may not have discrete solutions in informal-symbol systems. An informal symbol's being of a particular syntactic$^{\mathbf{P}}$ type may be no more independent of its semantic type than the individuation of a brain area is independent of the cognitive function it subserves. For perhaps no informal symbol is available for processing in accordance with syntactic$^{\mathbf{L}}$ rules until it has a determinate syntax$^{\mathbf{P}}$, and that it has a determinate syntax$^{\mathbf{P}}$ if and only if it has certain semantic properties.[32] If so, then the brain could be a semantic engine after all.

These cursory remarks only serve to illustrate my point that there is room within computationalism to solve the causal efficacy problem with minimal modification of the model. In particular, we can relax two assumptions that (to my knowledge) have never been explicitly defended in print: that mental representations are formal symbols and that the Language of Thought is a formal-symbol system. These assumptions are not essential to computationalism, for with or without them the model's characteristic mechanistic explanation of thought is preserved: internal processors are constrained to operate on the structures of symbols in accordance with (by assumption) intrinsically defined rules of the system. If this is right, then it would be false to say that a computational system does not have access to the meanings of its symbols.[33] This claim is true of all formal-symbol systems, but not of all formal symbol-systems. Some computational systems manipulate informal or natural symbols; and minds, if not computers, may be among them.

Acknowledgements I am grateful to the following for helpful comments and suggestions on various versions of this paper: Murat Aydede, Martin Davies, Michael Devitt, John Greenwood, Amy Hurshman, Amy Kind, Peter Kung, Onyoung Oh, Jeff Ramsey, David Rosenthal, Andrew Smith and Mas Yamada, and audiences at the New Jersey Regional Philosophical Association Fall 2004, the Society for Philosophy and Psychology June 2005, and the Claremont Colleges Philosophy Work-in-Progress group.

References

Aydede, M. (1997). Language of Thought: The connectionist contribution. *Minds and Machines, 7*, 57–101.

[32] Others who argue that there is no syntax without semantics (Crane 1990; Shagrir 2001) do not make this claim for syntax$^{\mathbf{P}}$. However, Peacocke (1995: 259) comes closest to the suggested view when responding to a Dennett-inspired objection to "content-involving computation"—that an organism only has access to an indicator of its internal states, not to whatever those internal states represent: "[S]uppose we fix on this detected internal state and specify the complex of internal and environmental relations in virtue of which detecting it is pretty much as good as detecting that the organism has ingested food. It seems to me that we are then specifying the relations in virtue of which the detector of that state is able to have the content that food has been ingested … If, as I have argued, the semantic and relational properties of later states can be explained by the presence of a state with this content about ingestion, there is a sense in which the mind is a semantic engine after all, and without any implication of executing the impossible." I do not claim that Peacocke would endorse my account, however.

[33] This does require denying a direct-reference semantics for the mind, whereby a mental representation's meaning *just is* its referent. If this were the case, however, there could be no *problem* of semantic causal efficacy, since it is indisputable that the brain does not manipulate, e.g., cats (as opposed to 'cat's).

Aydede, M. (2004). The language of Thought. *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/language-thought.

Baker, L. R. (1995). Metaphysics and mental causation. In J. Heil & A. Mele (Eds.), *Mental causation.* New York: Oxford University Press.

Bechtel, W. (1998). Representations and cognitive explanations: assessing the dynamicist's challenge in cognitive science. *Cognitive Science, 22*, 218–295.

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science, 66*, 175–207.

Beckett, S. (1955/1965). *Molloy*. In S. Beckett, M. D. Molloy, & The Unnameable (Eds.), *Three novels.* New York: Grove Press, Inc.

Bhushan, N., & Rosenfeld, S. (Eds.). (2000). *Of minds and molecules: New philosophical perspectives on chemistry*. Oxford: Oxford University Press.

Block, N. (1990). Can the mind change the world? In G. Boolos (Ed.), *Meaning and method: Essays in honor of Hilary Putnam*. Cambridge: Cambridge University Press.

Bontly, T. (1998). Individualism and the nature of syntactic states. *The British Journal for the Philosophy of Science, 49*, 557–574.

Borowsky, R., Loehr, J., Friesen, C. K., Kraushaar, G., Kingstone, A., & Sarty, G. (2005). Modularity and Intersection of "What", "Where" and "How" processing of Visual Stimuli: A new method of fMRI localization. *Brain Topography, 18*(2), 67–75.

Brown, C. (Ed.). (1990). *Chirality in drug design and synthesis.* San Diego: Academic Press, Inc.

Burge, T. (1979). Individualism and the mental. In P. A. French, T. E. Uehling, & Wettstein (Eds.), *Midwest studies in philosophy IV: Studies in metaphysics.* Minneapolis: University of Minnesota Press.

Burge, T. (1986). Individualism and psychology. *The Philosophical Review, 95*, 3–45.

Burge, T. (1989). Individuation and causation in psychology. *Pacific Philosophical Quarterly, 70*, 303–322.

Burge, T. (1995). Mind-body causation and explanatory practice. In J. Heil & A. Mele (Eds.), *Mental causation*. New York: Oxford University Press.

Chemero, A. (2000). Anti-representationalism and the dynamical stance. *Philosophy of Science, 67*(4), 625–647.

Churchland, P. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *The Journal of Philosophy, 95*, 5–32.

Churchland, P., & Sejnowski, T. (1992). *The computational brain*. Cambridge: The MIT Press.

Clark, A. (1998). Time and mind. *The Journal of Philosophy, 95*, 354–376.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 10–23.

Claverie, P., & Diner, S. (1980). The concept of molecular structure in quantum theory: Interpretation problems. *Israel Journal of Chemistry, 19*, 54–81.

Copeland, B. J. (2002). Narrow versus wide mechanism. In M. Scheutz (Ed.), *Computationalism: New directions*. Cambridge: The MIT Press.

Crane, T. (1990). The Language of Thought: No syntax without semantics. *Mind & Language, 5*, 187–212.

Crossley, R. (1995). *Chirality and the biological activity of drugs*. Boca Raton: CRC Press, Inc.

Davies, M. (1991). Individualism and perceptual content. *Mind, 100*, 461–484.

Devitt, M. (1991). Why Fodor can't have it both ways. In B. Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and his critics* (pp. 95–118). Cambridge: Basil Blackwell Ltd.

Devitt, M. (1996). *Coming to our senses*. New York: Cambridge University Press.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge: The MIT Press.

Dronkers, N. F., Redfern, B. B., & Knight, R. T. (2000). The neural architecture of language disorders. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 949–958). Cambridge: The MIT Press.

Egan, F. (1992). Individualism, computation and perceptual content. *Mind, 101*(403), 443–459.

Egan, F. (1995). Computation and content. *The Philosophical Review, 104*(2), 181–203.

Egan, F. (1999). In defence of narrow mindedness. *Mind & Language, 14*(2), 177–194.

Felleman, D. J., & van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*, 1–47.

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

Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Brain and Behavioral Sciences,* 3, 63–110.

Fodor, J. (1987). *Psychosemantics*. Cambridge: The MIT Press.

Fodor, J. (1985/1990). Fodor's guide to mental representation. In J. Fodor (Ed.), *A theory of content and other essays.* Cambridge: The MIT Press. Reprinted from *Mind.*

Fodor, J. (1994). *The Elm and the expert*. Cambridge: The MIT Press.

Fodor, J. (2000). *The mind doesn't work that way*. Cambridge: The MIT Press.

Fodor, J., & McLaughlin, B. (1990/1995). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition, 35*. Reprinted in Macdonald and Macdonald.

Fodor, J., & Pylyshyn, Z. (1988/1995). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71. Reprinted in Macdonald and Macdonald.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. Cambridge: The MIT Press.

Haugeland, J. (Ed.). (1997). *Mind design II: Philosophy, psychology and artificial intelligence*. Cambridge: The MIT Press.

Haxby, J. V., Grady, C. I., Horwitz, B., Underleider, L. G., Mishkin, M., Carson, R. E., Herscovitch, P., Schapiro, M. B., & Rapoport, S. I. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences of the United States of America, 88*(5), March 1, 1621–1625.

Heil, J., & Mele, A. (Eds.). (1995). *Mental causation*. New York: Oxford University Press.

Hicks, J. (Ed.) (2002). *Chirality: Physical chemistry*. Washington, D. C. : American Chemical Society (distributed by Oxford University Press).

Horst, S. (2005). The computational theory of mind. *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/computational-mind/.

Humberstone, I. L. (1996). Intrinsic/extrinsic. *Synthese, 108*, 205–267.

Jackson, F., & Pettit, P. (1988). Functionalism and broad content. *Mind, 97*, 381–400.

Jacques, J. (1993). *The molecule and its double*. New York: McGraw-Hill Inc (trans. by Lee Scanlon, The Language Service, Inc., Poughkeepsie, New York).

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17*(11), 4302–4311.

Kim, J. (1982). Psychophysical supervenience. *Philosophical Studies, 41*, 51–70. Reprinted in Kim (1993), 175–193.

Kim, J. (1984). Self-understanding and rationalizing explanations. *Philosophia Naturalis, 21*, 309–320.

Kim, J. (1993). *Supervenience and mind: Selected philosophical essays*. New York: Cambridge University Press.

Kim, J. (1998). *Mind in a physical world*. Cambridge: The MIT Press.

Kosslyn, S. M. (1999). If neuroimaging is the answer, what is the question? *Philosophical Transactions of the Royal Society: Biological Sciences, 354*, 1283–1294.

Langton, R., & Lewis, D. (1998). Defining 'intrinsic'. *Philosophy and Phenomenological Research, 58*, 333–345.

Latal, H. (1991). Parity violation in atomic physics. In R. Janoschek (Ed.), *Chirality: From weak Bosons to the α-helix*. Berlin and Heidelberg: Springer-Verlag.

Lewis, D. (1983). Extrinsic properties. *Philosophical Studies, 44*, 197–200.

Le Poidevin, R. (1994). The chemistry of space. *Australasian Journal of Philosophy, 72*, 77–87.

Le Poidevin, R. (2000). Space and the chiral molecule. In N. Bhushan & S. Rosenfeld (Eds.), *Of minds and molecules: New philosophical perspectives on chemistry*. Oxford: Oxford University Press.

Liboff, R. (2003). *Introductory quantum mechanics* (4th Ed ed.). San Francisco: Pearson Education Inc., publishing as Addison Wesley.

Macdonald, C., & Macdonald, G. (Eds.). (1995). *Connectionism: Debates on psychological explanation, Vol. 2*. Oxford and Cambridge: Blackwell Publishers Ltd.

Marr, D. (1982). *Vision*. San Francisco: Freeman Press.

Milner, A. D., & Goodale, M. G. (1995). *The visual brain in action*. Oxford: Oxford University Press.

Nerlich, G. (1995). On the one hand: Reflections on enantiomorphy. *Australian Journal of Philosophy, 73*, 432–443.

Peacocke, C. (1993). Externalist explanation. *Proceedings of the Aristotelian Society New Series, 93*, 203–230.

Peacocke, C. (1994). Content, computation and externalism. *Mind & Language, 9*, 303–335.

Peacocke, C. (1999). Computation as involving content: A response to Egan. *Mind & Language, 14*(2), 195–202.

Pettit, P. (1986). Broad-minded explanation and psychology. In P. Petit & J. McDowell (Eds.), *Subject, thought and context* (pp. 17–58). New York: Oxford University Press.

Piccinini, G. (forthcoming). Computation without representation. *Philosophical Studies*. Online publication date 22 Sept. 2006.

Putnam, H. (1975). The meaning of 'meaning'. In H. Putnam (Ed.), *Mind, language and reality: Philosophical papers, vol. II*. New York: Cambridge University Press.

Ramsey, J. (1997). Molecular shape, reduction, explanation and approximate concepts. *Synthese, 111*, 233–251.

Ramsey, J. (2000). Realism, essentialism, and intrinsic properties: The case of molecular shape. In N. Bhushan & S. Rosenfeld (Eds.), *Of minds and molecules: New philosophical perspectives on chemistry*. Oxford: Oxford University Press.

Ramsey, W. (1997). Do connectionist representations earn their explanatory keep? *Mind & Language, 12*, 34–66.

Ramsey, W., Stich, S., & Garon, J. (1990/1995). Connectionism, eliminativism and the future of folk psychology. *Philosophical Perspectives, 4*. Reprinted in Macdonald and Macdonald.

Rey, G. (1997). *Contemporary philosophy of mind: A contentiously classical approach*. Oxford: Basil Blackwell.

Rupert, R. (2004). Challenge to the hypotheses of extended cognition. *The Journal of Philosophy, 101*(8), 389–428.

Segal, G. (1989). Seeing what is not there. *The Philosophical Review, 98*(2), 189–214.

Segal, G. (1991). Defense of a reasonable individualism. *Mind, 100*, 485–493.

Shagrir, O. (2001). Content, computation and externalism. *Mind, 110*, 369–400.

Smith, B. C. (2002). The foundations of computing. In M. Scheutz (Ed.), *Computationalism: New directions*. Cambridge: The MIT Press.

Smolensky, P. (1991/1995). Connectionism, constituency and the Language of Thought. In B. Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and his critics*. Oxford: Basil Blackwell Ltd. Reprinted in Macdonald and Macdonald.

Stich, S. (1983). *From folk psychology to cognitive science*. Cambridge: The MIT Press.

Stich, S. (1991). Narrow content meets fat syntax. In B. Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and his critics*. Oxford: Basil Blackwell.

Strawson, P. F. (1985). Causation and explanation. In Vermazen & Hintikka (Eds.), *Essays on Davidson: Action and events*. New York: Oxford University Press.

Thagard, P. (2002). How molecules matter to mental computation. *Philosophy of Science, 69*, 429–446.

Ungerleider, L., Courtney, S., & Haxby, J. V. (1998). A neural system for human visual working memory. *Proceedings of the National Academy of Science USA, 95*, 883–890.

Ungerleider, L., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge: MIT Press.

Vallentyne, P. (1997). Intrinsic properties defined. *Philosophical Studies, 88*, 209–219.

Van Gelder, Tim. (1995). What might cognition be, if not computation? *The Journal of Philosophy, 92*, 345–381.

Weininger, S. (1984). The molecular structure conundrum: Can classical chemistry be reduced to quantum chemistry? *Journal of Chemical Education, 61*, 939–944.

Wilson, R. A. (1994). Wide computationalism. *Mind, 103*, 351–372.

Wilson, R. A. (1995). *Cartesian psychology and physical minds*. Cambridge: Cambridge University Press.

Wilson, R. A. (2004). *Boundaries of the mind: The individual in the fragile sciences*. New York: Cambridge University Press.

Woolley, R. G. (1978). Must a molecule have a shape? *Journal of the American Chemical Society, 100*, 1073–1078.

Woolley, R. G. (1985). The molecular structure conundrum. *Journal of Chemical Education, 62*, 1082–1084.