Synthese
DOI 10.1007/s11229-009-9581-1

LOT, CTM, and the Elephant in the Room

Susan Schneider

Received: 2 June 2009 / Accepted: 2 June 2009

Abstract According to the language of thought (LOT) approach and the related computational theory of mind (CTM), thinking is the processing of symbols in an inner mental language that is distinct from any public language. Herein, I explore a deep problem at the heart of the LOT/CTM program—it has yet to provide a plausible conception of a mental symbol.

Keywords Language of thought · Computational theory of mind · Symbol · Connectionism · Fodor · Symbol processing

There is a major problem at the heart of the language of thought (or "LOT") program. Without a solution to the problem, LOT and the related Computational Theory of Mind (CTM) are ill conceived and must be rejected. Or so I shall contend. The problem is as easy to state as it is difficult to solve. According to LOT, thinking is symbolic. But LOT has not provided a plausible conception of a mental symbol. For the various candidate conceptions are highly problematic, given the philosophical work that the proponent of LOT requires mental symbols to do. And, of course, the core of language of thought picture is the claim that thinking is the causal sequencing of symbol tokens. Without this core, there is not much of a view left. Mutatis mutandis for CTM.[1]

My game plan shall be simple: Section 1 overviews a few background issues of import to our discussion. Here, I remind the reader of the significance of the symbol processing program and of the philosophical roles that mental symbols are summoned

[1] CTM is closely related to LOT, holding that thinking is a kind of classical computation involving the manipulation of semantically interpretable strings of symbols. Herein, my focus is mainly on the problem of symbols as it arises for LOT, for the symbols are said to be in *this* language.

S. Schneider (✉)
Department of Philosophy, Center for Cognitive Neuroscience and Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA 19104-6304, USA
e-mail: sls@sas.upenn.edu

Published online: 16 July 2009

to play for LOT. For as we shall see, one needs to bear in mind these roles in order to determine whether a candidate account of symbols is indeed suitable. Then, in Sect. 2, I consider each of the proposals for symbol types that I know of, arguing that each fails. Some fail for other reasons, but a recurring theme is that many of the theories do not deliver a notion of a symbol that can perform one or more non-negotiable functions that symbols are supposed to play for the LOT/CTM picture. In the final section I explore the implications of this result, indicating a direction that proponents of LOT and CTM could consider. My proposal evicts the elephant from the room; however, there is cost to the landlord.

1 The mind as a symbol processing engine?

Remember, LOT is primarily appealed to in the context of the following age old question: What is the fundamental nature of conceptual thought? And the proponent of LOT answers by claiming that the cognitive mind computes in an internal language-like representational code that is not equivalent to one's spoken language. Thinking is the processing of symbols by an algorithm, where this algorithm is supposed to be the program that the cognitive mind runs. These inner symbols are supposed to be neo-Fregean modes of presentation (called "MOPs") or what Jerry Fodor calls the inner "vehicle" of thought. They are narrow, or "in the head," being determined by the intrinsic properties of the individual. The language of thought position is absolutely central to contemporary cognitive science. In particular, it is at the heart of current information processing psychology and philosophy of mind, being one of the two leading positions on the computational nature of thought (the other being connectionism).[2] It is also of import to computer science and AI. This view has a distinguished group of proponents: Allan Newell, Herbert Simon, Jerry Fodor, Stephen Pinker, Gilbert Harman, Peter Carruthers, Gary Marcus, Lila Gleitman, Zenon Pylyshyn, Brian McLaughlin, and Georges Rey to name a few.

Why claim that there is a language of thought? The main motivation for the language of thought position is the *combinatorial* nature of thought. More specifically, thought is productive: in principle, one can entertain and produce an infinite number of distinct representations. This indicates that the mind has a combinatorial syntax, allowing for the construction of potentially infinitely many thoughts given a finite stock of primitive expressions (Fodor 1975; Fodor and Pylyshyn 1988). Further, conceptual thought is *systematic*. A representational system is systematic when the ability of the system to entertain/produce certain representations is intrinsically related to the ability to entertain/produce other representations (Fodor and Pylyshyn 1988). For example, one doesn't find normal adult speakers who entertain/produce "Jose loves Irene" without also being able to entertain/produce "Irene loves Jose." How can this fact be explained? Intuitively, "Jose loves Irene" is systematically related to "Irene loves Jose" because they have common constituents (Fodor and Pylyshyn 1988; Fodor and McLaughlin 1990). Explaining the combinatorial nature of thought should clearly be a central goal of any theory of the cognitive mind. Of course, the LOT/CTM

[2] Outside of philosophy it is more commonly called the "symbol processing view."

view is no longer the "only game in town," as Fodor used to boast (Fodor 1975). However, it is currently a source of intense controversy whether any connectionist models can explain these important features of thought (see, e.g., Fodor and Pylyshyn 1988; Fodor and McLaughlin 1990; Elman 1998; van Gelder 1990; Marcus 2001; Smolensky 1988, 1995).

Indeed, connectionism seems to be quite popular among the younger generation of philosophers of mind, who find the connectionist successes in certain low-level domains to be very promising. They seem to have discarded LOT altogether. But this could end up being a terrible mistake. Not only is it controversial whether connectionist models can fully explain the combinatorial nature of thought, but it unclear how very simple models of isolated neural circuits are supposed to "come together," giving rise to a larger picture of how the mind works (Anderson 2007). Existing "big picture" accounts are intriguing, yet patchy and speculative (see e.g., Hawkins 2005). Further, models of higher cognition seem to be precisely the terrain in which one would expect to see validation of the symbol processing approach, if validation is to come. As connectionists Randal O'Reilly and Yuko Munakata admit in their recent computational neuroscience textbook (MIT Press, 2000), the symbolic approach to higher-level cognition has a "long history of successful models." For, "in symbolic models, the relative ease of chaining together sequences of operations and performing arbitrary symbol binding makes it much more straightforward to simulate higher-level cognition than in a neural network."[3] In contrast, "neural network models of higher-level cognition are in their relative infancy" (p. 379). And although representation in the prefrontal cortex (PFC) is still poorly understood relative to many other brain regions, as they point out, representation in the PFC appears to be combinatorial and discrete. If correct, this would support an appeal to symbolic models to explain higher-level cognition (perhaps implemented by connectionist networks, perhaps not). The combinatorial and discrete representations of the PFC are distinct from the more distributed modality specific representation of the posterior cortex; prima facie, this latter representation seems more straightforwardly amenable to traditional connectionist explanation.[4] And all this comes from the latest computational neuroscience textbook, not just from symbolicists like Fodor and Marcus. So I would urge that both sides be modest—it is still very early in the game. LOT is still quite relevant, despite connectionist success stories. It may be wrong. But it *may* be right.

Add to this the fact that the precise relationship between LOT and connectionism is extremely subtle. The proponent of LOT has an important rejoinder to the aforementioned connectionist attempt to do without mental symbols: to the extent that the connectionist can explain the combinatorial nature of thought then, at best, connectionist systems would merely provide models in which symbols are implemented in the cognitive mind, and would not really represent genuine alternatives to the LOT picture. For the networks would ultimately be the lower-level implementations of symbolic processes (such a view is often called "implementational connectionism").[5]

[3] O'Reilly and Munakata (2000, p. 379). They offer an intriguing account of their own of higher cognition.

[4] For a discussion of the distinct processing in the PFC and posterior cortex see O'Reilly and Munakata (2000, pp. 214–219).

[5] Fodor and Pylyshyn (1988), Pinker and Prince (1988), Marcus (2001).

For connectionism and symbolicism to represent genuine alternatives radical connectionism must be correct. But not only are existing connectionist models of higher cognitive function few, but there are persuasive arguments that to the extent that they are not overtly implementationalist, putative radical connectionist models in fact make *covert* use of symbolic representations (Marcus 2001). So again, there is reason to look to the language of thought approach, even if one is sympathetic to connectionism. But it had better have a coherent notion of a symbol.

So now let us ask: does LOT have a plausible account of symbol natures? A requisite first step in answering this question involves considering the central and commonly agreed upon philosophical functions that symbols are supposed to play. For when we consider the candidate symbol conceptions in Sect. 3, we shall ask whether any conceptions do in fact fill these non-negotiable roles. And today's answer shall be: *none do*.[6]

2 Putting symbols to work

Jerry Fodor one voiced the worry that without a theory of symbol natures "… the whole [LOT] project collapses.[7] Indeed. For consider the dismal intellectual landscape that LOT/CTM faces should no theory of symbols be forthcoming. First, without a theory of symbols it would be unclear how patterns of neural activity could be, at some higher level of abstraction, accurately described as being the manipulation of symbols. For what is it that is being manipulated? It would thereby seem difficult for cognitive scientists working on symbolic processing to determine if the symbolic view is, in fact, correct. Further, a number of connectionists claim that their view is also symbolic. Maybe so. But how do we know that a given connectionist theory really is compatible with the LOT program if we do not even have a sense of what LOT symbols are? Perhaps the two views are merely talking past each other, employing different concepts of symbols to begin with. As cognitive scientist John Anderson, a Rumelhart Prize winner and leading figure in the symbol processing tradition, recently summarized the state of things: "Given this lack of agreement of what symbols are, it should come as no surprise that there is no consensus about what role symbols play in an explanation of the mind and how they should be coordinated with our knowledge of brain processing" (Anderson 2007, p. 31). Later, he writes, "Indeed, debates among these positions have the character of jousting with windmills. Because there is not even agreement about what symbols mean, these debates are a waste of time" (p. 33).

Second, consider the case of mental causation. Causation is a relation between events; and in the present context, the relevant events are supposed to be tokenings of symbol types. Now, imagine a microphysical theory that failed to deliver up an inventory of fundamental properties. Intuitively, this wouldn't be much of a theory, would it? After all, how would it make sense of physical causation? Here, I am not asking for a theory of causation, which is obviously not the job of microphysics. The general issue is this: how could a theory at a given scientific level go about testing its

[6] However, I must confess that today I am playing devil's advocate, hoping to convince the reader of the serious nature of the problem. For I ultimately claim that the computational role approach can surmount the objections offered herein (Schneider 2009b).

[7] Quoted in Pessin, op. cit., p. 33.

causal predictions if it can't even say what the kinds are which are supposed to figure in the predictions? Mutatis mutandis, how would a computational theory of mind do so, in absence of an account of the mental state types that figure in the algorithms?

The above two considerations go well beyond issues within philosophy of mind proper, being of clear import to researchers in cognitive science, for they concern LOT/CTM's value as a scientific account of how the mind works. Clearly, for the symbolic conception to be of value to cognitive research, symbols need to be well defined so that their relationship to other computational accounts can be clarified and so that predictions involving symbols can be made. But while philosophers who appeal to LOT would clearly recognize these functions as important, they conceive of symbols as doing significant additional work within philosophy of mind proper—namely, serving as neo-Fregean MOPs, being causally efficacious, and facilitating an answer to the traditional problem of intentionality. And as noted, these functions will be of special significance to today's project. So let us consider this philosophical work in a bit more detail.

Roles for symbols:

(i) *Figure as neo-Fregean MOPs*. The most crucial philosophical role that symbolic representations are supposed to play concerns their role as "modes of presentation" (often simply called "MOPs"). Consider: there are different ways of representing the same entity. And despite differences in ways of representing things, the diverse ways can pick out, or refer to, the same entity. Philosophers of language have called these ways of representing things "guises" leaving it open what their precise nature is. Within philosophy of mind, theories of narrow content take themselves to be offering accounts of MOPs. It is also reasonable to say that MOPs play an important role in contemporary theories of propositional attitude ascription, being appealed to account for psychologically different ways of grasping the same proposition. And cognitive scientists have taken computational entities like LOT symbols and activation patterns in connectionist networks to provide a notion of a MOP that is purely narrow, or "in the head," being determined by the intrinsic properties of the system. Such entities can be regarded as computational analogs of guises or MOPs; they provide computational structures that can be summoned to account for the distinct ways that we think of referents or states of affairs.

As MOPs, symbols are supposed to be at a suitable level of grain to capture one's way of conceiving the world, being computational versions of what the layperson would describe as one's inner "concepts" or "notions." In this vein, symbols must be finely grained enough to distinguish intuitively distinct thoughts that happen to corefer (e.g., Cicero/Tully, groundhog/woodchuck). For as Frege's Hesperus/Phosphorus case emphasized, these ways of conceiving of an entity are distinct from the entity referred to; for one may know of a referent via certain of its features and yet be ignorant of others.[8] Further, symbols, as MOPs, can be type

[8] For many theories this provides a manner in which coreferring thoughts differ in their meanings; for others, thoughts differ merely in (nonsemantic) cognitive significance. This latter position is what the proponent of LOT generally has in mind.

identical in cases in which individuals have the same inner mental states while the referent differs, as in Putnam's Twin Earth case or Tyler Burge's arthritis example.

(ii) *Causal efficacy*. Recall that computational theories of MOPs, like the LOT approach, look to theories of cognitive processing in cognitive science to supply the basis for an account of the causal production of thought and behavior. In this vein, symbol tokenings are supposed to be causally efficacious, in the sense that they function causally to generate other mental states and behaviors. Relatedly, symbols function in causal explanations of thought and behavior including, importantly, explanations of the aforementioned productive and systematic nature of thought (Fodor 1998, 2000).

(iii) *Facilitate naturalism*. Third, proponents of LOT and CTM regard symbols as being essential elements of their response to the problem of intentionality. It has long been suspected that thought is somehow outside the realm that science investigates, being categorically distinct from the physical realm. For how is it that a thought, which, as we know, arises from states of the brain, can be directed at, or about, something in the world? This is the classic problem of the nature of intentionality. In broad strokes, the proponent of LOT and CTM answers the problem of intentionality by claiming that the intentionality of a thought is a matter of a nomic, and ultimately, physical, relationship between symbolic computational states and entities in the world (Schneider 2009a). As Loewer and Rey put it, symbols are "locked onto" properties or individuals of a certain sort in virtue of standing in a certain nomic or causal relationship specified by a theory of meaning or mental content (Loewer and Rey 1993).

These roles are clearly recognizable as being of central import to the LOT/CTM tradition. This is largely due to the work of Jerry Fodor, who, in particular, has developed a well-known philosophical program in which mental symbols are said to play each of above philosophical roles. But is it even plausible to venture that each of these roles is *non-negotiable*, being required for the success of LOT and CTM? I believe so. Turning to (i) and (ii), these roles are absolutely key: symbols are inner mental states which are supposed to explain how and why an individual thinks and behaves in the way that she does. These roles are of enhanced import given that proponents of LOT generally do not appeal to mental content to do the work of neo-Fregean MOPs, for on their view, content is broad. Turing to (iii), LOT/CTM is a naturalistic program that regards locating an answer to the problem of intentionality as being of primary import. So I am inclined to say that each of these roles is non-negotiable.

We are finally ready to answer the question: does any single conception of a symbol satisfy all these roles? Initially, one might suspect so. For the proponent of LOT finds these philosophical roles to be well conceived, and, given this, might assume that, based on the presence of well-conceived roles, that LOT has a clear conception of what a symbol is. However, identifying a philosophical role for a putative entity is not tantamount to giving its individuation conditions. For example, in his classic "New Work for a Theory of Universals" David Lewis identifies numerous roles for sparse properties yet shies away from firmly deciding between Armstrongian universals or a form of class nominalism (Lewis 1983). At least in principle, different conceptions—that is,

entities having different individuation conditions—can play the same roles. In a similar vein, to understand the nature of symbols, one needs a conception of the individuation conditions on that which is supposed to play the roles we just identified. To accomplish this task, in this context, means locating features according to which symbols should be taxonomized, or classified. For instance, should two symbol tokens be regarded as being of the same type when they have the same semantic content? Or should symbols perhaps, instead, be type individuated by their underlying neural properties?

But the aforementioned roles do provide some insight into symbol natures in the following sense: given competing conceptions of symbols, it is reasonable to select the one which is best able to play the important philosophical roles that symbols are supposed to play, if any does. This is only fair. While debates over certain metaphysical categories (e.g., laws) feature debates in which each side disagrees on what philosophical roles need to be satisfied by the category, this problem will not arise for us today. For while those interested in LOT may certainly debate whether symbols should play certain, more controversial roles, the roles that the different views fail to satisfy today are all one's which are clearly non-negotiable to all.[9] So let us now consider the various proposals.

3 Candidate conceptions

3.1 The semantic proposal

Symbols stand for things. But whether they do so essentially is another story. At one time, Fodor individuated symbols by their meanings (Fodor 1989, p. 167).[10] On the other hand, Searle (1980) speaks of meaningless symbol manipulations in the context of his classic Chinese Room thought experiment, and Stephen Harnad worries about the symbol grounding problem in which meaningless symbols are manipulated (1990). Clearly there is no consensus on whether symbols are to be individuated by meanings. But let us nonetheless determine if some sort of semantic proposal would work. In particular, since LOT is currently being developed against the backdrop of a referential semantics, let us see whether a referential proposal is plausible. According to the referential proposal symbols should simply be classified by their broad contents. So we have:

(CD1) Two primitive symbol tokens are of the same symbol type iff they have the same broad content.[11]

We can quickly see that (CD1) is problematic. (i) First, it fails to deliver a notion of a symbol which facilitates naturalism (Role Three) as a referential manner of typing

[9] Another key role for symbols is the following: Symbols are supposed to be the bearers of mental content, where the theory of content appealed to is generally externalist, e.g., see Fodor's asymmetric dependency theory. But syntactic eliminativists would dispute this role as they do not believe in content (Stich 1983).

[10] And also Crane (1990).

[11] Here I'm following Fodor's usage of "broad content" that is referentialist (Fodor 1994, p. 7). There is another usage in which "broad content" is taken as synonymous with "wide content."

LOT expressions ruins the prospects for naturalism (Pessin 1995). For the externalist hopes to naturalize intentionality by taking the intentionality of thought to be a matter of a symbol bearing some sort of external relationship (e.g., causal, informational) to a property or thing in the world. But if the intentionality of thought is supposed to reduce to a physical and non-intentional relation between the symbol and the world, the symbol itself cannot be typed semantically. For if symbols have semantic natures, the intentionality of thought wouldn't reduce to the physical. (ii) Second, bearing in mind that proponents of LOT generally envision symbols as being at a level of grain suitable to explain one's way of conceiving the world (Role One), referential individuation clearly will not suffice. For consider that coreferential symbol tokens (e.g., [Hesperus]/[Phosphorus]) will be regarded as being type identical while functioning very differently in one's cognitive economy.[12] (iii) Relatedly, treating coreferential but intuitively distinct symbols as type identical will lead to poor explanations of the causation of thought and behavior because the tokens can function very differently in bringing about subsequent thoughts and behaviors (Role Two) (Fodor 1994; Schneider 2005; Braun 2001; Richard 1990).

So a referential manner of typing symbols will fail to deliver a notion of a symbol which plays any of the aforementioned roles that symbols are supposed to play. Now, one might believe that the problems are all due to the fact that symbols are being typed referentially, rather than merely semantically. For instance, would the individuation of symbols by narrow content avoid some of these problems? Assuming that one has a plausible theory of narrow content in hand, and that the narrow contents track one's way of conceiving of the world, this would eliminate the second and third problems. However, the first problem extends to any kind of semantic proposal. And given the centrality of naturalism to the LOT program, this is a decisive reason to reject any semantic proposal. So let us turn to the next approach to symbol natures.

3.2 "Orthographic" or type identity proposals

Proposals concerning the nature of symbols have been said to be "orthographic" when they appeal, in some rough sense, to the "shape" of a symbol in the language of thought (Millikan 1993). Unfortunately, although expressions like "shape" and "orthography" are employed in discussions of LOT symbols, these notions, when applied to thoughts, are extremely vague. For the language of thought lacks an orthography, or even a phonology. After all, there is no such thing as "brain writing." But the view is not entirely baseless—beneath these vague metaphors is a suspicion that symbols are to be type individuated by certain of their neural properties. And this is not to be laughed at. For one thing, cognitive neuroscientists speak of concepts being located in particular assemblies of neurons, and, to an extreme, Christof Koch and others have defended the existence of "grandmother" cells: single cells which seem to fire only in the presence of a certain familiar individuals (e.g., grandmother, Clinton, etc.). Now, it is far from clear whether the presence of such cells motivates anything like a type identity theory in philosophy of mind, but it is at least prima facie plausible that some sort of type

[12] I shall employ the convention of using brackets to refer to MOPs. E.g., "[dog]" refers to the MOP, dog.

identity account will be borne out as neuroscience develops and refines accounts of higher cognitive functions. So let us consider the following:

(CD2) Two primitive symbol tokens are of the same symbol type if and only if they are tokens of the same brain state type.

Of course, this condition is just a version of the well-known type identity theory of mental states, fashioned to symbol types.

As with the earlier proposal, numerous problems emerge for (CD2). As Fodor has noted, (CD2) pushes much of the problem of symbol individuation to the problem of how brain states are individuated, and it is by no means clear how to individuate brain states. He further suggests that there are presently two main ways to individuate brain states: neuroanatomically or by their neurocomputational roles.[13] So let us consider each of these suggestions in turn. First, an appeal to the individuation of brain states in terms of their neurocomputational roles is problematic, for proponents of LOT—Jerry Fodor and Gary Marcus, for example—doubt that computational neuroscience, being connectionist, will ultimately deliver even an implementationalist account of how the cognitive mind operates. So it would be odd for the notion of a symbol to depend on the success of a neurocomputational approach to higher thought. For if Fodor and others are right, symbols would be unable to satisfy the important philosophical role of serving in explanations of the causal features of thought (see Role Two); in particular, symbols would fail to explain the systematicity and productivity of thought.

Now let us consider Fodor's suggestion that brain states can be typed by certain of their neuroanatomical features. Here, two complications arise: first, important concepts (e.g., MOTHER, DOG) seem to survive hemispherectomy, commisurotomy and various types of injuries (Gazzaniga et al. 2002). This indicates that familiar concepts may be duplicated throughout one's cortex. Prima facie, if the same concept is multiply realized in different parts of the cortex then it is unclear how a neuroanatomical proposal would pan out. Secondly, (CD2) fails in the across person case for a related reason. For it is now a well-known fact in neuroscience that there are great interpersonal variations in the data structures of the cortex. So this would not yield a theory of symbols in which different individuals' token symbols of the same type. In sum, these issues make the development of a neuroanatomical version of (CD2) very challenging indeed. Another proposal that may come to mind involves an appeal to the semantic features of the brain states, but this sort of position will not help the symbolicist, for as noted, a semantic conception of a symbol will only muddle LOT's ambition to deliver a naturalistic theory of intentionality. As noted, facilitating naturalism is an important role that symbols are supposed to play.

In sum, both the semantic and type identity approach fail. Still, there is one more approach to consider (or rather, family of approaches). This popular kind of approach types symbols by the role the symbol plays in computation (Aydede 1999; Stich 1983; Schneider 2009b; Fodor 1994).

[13] Jerry Fodor (in conversation).

3.3 Computational role approaches

Proponents of this approach include Fodor (1994) and Stich (1983), although neither have developed their views of symbols in sufficient detail. There are two basic types of computational role accounts. The division concerns how the notion of computational role is to be understood by the theory. According to one view, the computational role of a symbol includes the entire role the symbol is capable of playing in one's cognitive economy. The symbol is individuated by *all* of the generalizations that involve the symbol, where such generalizations are not "folk" generalizations, but, in the spirit of a posteriori functionalism, are generalizations detailed by a completed cognitive science. Let us call such accounts "total computational role" accounts. The second type of proposal has a more restrictive understanding of computational role in the sense that it singles out only some elements of the total role the state is capable of playing as being type individuative.[14]

Consider the former construal of computational role. It is widely agreed that the individuation of symbols in terms of total computational role will taxonomize symbols finely enough to play the first role of being neo-Fregean MOPs, capturing one's way of conceiving things. And being non-semantic, it should suffice for the purposes of naturalism. Yet it is also commonly agreed that this manner of individuation will not yield a sense in which different individuals share symbols, and this is taken to be highly problematic for explaining the causation of thought and behavior (Role Two). In particular, while it allows for explanations of the thoughts and behaviors of single individuals, it does not seem to allow for explanations that cover groups of individuals. Of course, we have the antecedent sense that two people can, and perhaps frequently do, represent entities in the same way. For example, different individuals may both know of Venus as the morning star, rather than the evening star, and behave accordingly. Yet critics charge that this conception of a symbol delivers no sense in which different individuals can share the same MOP, where MOPs are taken as being LOT expressions. For if symbols are individuated by the role they play in one's entire cognitive economy, then given that we commonly differ with respect to memories, personality traits and so on, symbols will not be shared. This situation gives rise to the following 'Publicity Argument' (Aydede 1999; Prinz 2002; Schneider 2009a,b). If thought is indeed symbolic, then explanations of why different individuals think or behave in the same ways would presumably appeal to their tokening type identical symbols, yet, as we've just noted, on this conception of symbols different individuals apparently do not have symbols of the same type.[15] As a result, different individuals will not be subsumed under the same psychological generalizations. So psychological

[14] Both of these conceptions generally take the roles to specify the causal powers of the symbol, rather than merely what the symbol happens to cause (i.e., its actual causal profile given the circumstances the system encounters).

[15] This objection was also offered by Fodor in conversation with the author and is akin to his earlier position on functionalist theories of narrow content (Fodor and LePore 1992).

I respond to the objection concerning symbols in Schneider (2009b), however. But as noted, today I am playing advocate in order to persuade the reader of the import of the problem of symbol individuation and of the serious difficulties facing the different views of symbol natures. I briefly outline my positive view in the conclusion.

explanations will fail to be "public"; different individuals, or even the same individual at different times, will not satisfy the same generalizations.

Let us turn to the other approach to computational role then. This other approach is a sort of "molecularism" about LOT symbol types, designed in the spirit of molecularism about narrow content individuation. Molecularism about symbols singles out from the class of causal relations (and more specifically, computational relations) that the expression has a certain privileged few as being type-individuative. The causal connections appealed to are designed to get a common equivalence class of systems which, when tokening a given symbol and in common conditions, will, ceteris paribus, think or behave in the similar ways. For instance, there might be a population of "novices" who all know the same small amount of information about a kind. Here, let us employ talk of "mental files." In the mental file for a natural or artificial kind concept, a novice may only have the most very basic facts. For example, someone may know only that "brane" names a fundamental entity in M-theory. Other novices only have this very sparse grasp of this putative natural kind as well. So the proponent of molecularism about symbols may ask: why can't a mental word [brane] be shared between the novices; i.e., those who also have this very skeletal [brane] concept? A similar scenario can apply to names; consider the contrast between an expert's knowledge of Einstein and a novice who may only know that he's the person who devised relativity theory.

The strongest case for molecularism concerns mental words for logical and mathematical expressions: just as it is relatively easy to locate definitions for logical and mathematical expressions is also easy to isolate a computational role that all who have these kinds of MOPs share. In the context of discussions of the plausibility of conceptual role semantics, logico-mathematical concepts are generally the parade cases. However, the problem is that the case for MOP type constituting computational relations is significantly weaker for other expression types. Critics charge that molecularist theories of narrow content have failed in a related attempt to identify certain conceptual or inferential roles as being constitutive of narrow contents. In broad strokes, the problem is that there seems to be no principled way to distinguish between those elements of conceptual or inferential role that are meaning constitutive from those which are not (Fodor and LePore 1992). Similar issues emerge for molecularism about symbol types, although the issues do not concern content individuation but symbol individuation. Indeed, Murat Aydede has posed the following dilemma for the molecularist: insofar as one singles out a select few symbol-constitutive computational relations, the analysis will fail to distinguish intuitively distinct symbol types. But if one builds more relations into the requirements for having a symbol, different individuals do not have symbols of the same type. For instance, suppose that the symbol, [dog], is individuated by computational relations involving [canine], and [standardly four legged]. But alas, what if someone thinks dogs fly, or that they do not bark, or that they are not house pets? If they think any of these things, do they really have the same notion of a dog as you and I? If not, are we really prepared to say that they nonetheless share the same MOP or mental symbol? A natural reaction is to strengthen the requirement on having [dog] in various dimensions. But now, because [dog] is sensitive to many elements of the symbol's computational role, different individuals will not generally token it. Instead, they will each have their own idiosyncratic mental symbols for dogs. This

is a similar dilemma that faced conceptual role accounts of narrow content (Aydede 1999).

The proponent of molecularism may suspect that one merely needs to develop a detailed argument for including certain computational relations and excluding others. But I will not attempt to devise such an argument, for I suspect that it would go nowhere. For as I shall now argue, insofar as there is *any* difference in computational role between tokens of identical symbol types, LOT/CTM will either be incomplete or there will be counterexamples to computational generalizations. The strategy behind the argument is the following: Assume that P is some principle of individuation of LOT symbols and that P is *not* equivalent to individuation by total computational role. This is, for instance, the case with molecularist individuation conditions. Then, wouldn't there be cases in which two LOT expressions are type identical according to P while differing in their total computational roles? This must be so, for if there are not, then P is just equivalent to a principle that types LOT primitives in terms of sameness and difference of total computational role. But as I will argue below, if there is a case in which two LOT tokens of the same type differ in any element of their total computational roles, as molecularist ones do, then either there will be missed predictions or there will be counterxamples to computational generalizations.

Here's the argument: Let "CR" denote the causal role of a given LOT token, *a*. And let "CR*" denote an individuation condition for the type that *a* is a token of. CR* is a condition that employs individuation by computational role, where the computational role includes every computational-level causal relation that the token enters into with other primitives except for one relation, R*. So R* is not individuative of the type that *a* is a token of. But *a* has R*. I take it that the causal relations that specify the computational role of a given token are detailed by the computational laws. So there is a given computational law, L, which specifies R*. Now, let *b* be a token that only has the causal role specified by CR*, and not CR, because *b* lacks R*. And let us suppose that like *a*, *b* is typed by CR*. Then, either: (i), both *a* and *b* will not be subsumable in L. Or, (ii), they will both be subsumable in L. In the case of (i) the theory will have a missed prediction: it will miss that *a* has a causal relation that is specified by L. (ii) Now consider the second scenario, in which they will both be subsumable in L. In this case *b* does not have the causal relation detailed by L. So we wouldn't expect it to behave in accordance with L. Hence, *b* will be a counterexample to L. Hence, insofar as even one element of the total computational role is ignored by an individuation condition, as with molecularist conditions, there will be counterexamples or mixed predictions.

Further, the sort of counterexample will be a kind of "Frege case." By "Frege case" I mean a certain sort of counterexample that arises for psychological laws that subsume states that are individuated in a manner that is too coarsely grained for the purpose of capturing important behavioral similarities. Frege cases are well-known in the literature on broad content and intentional explanation.[16] But Frege cases can arise at the computational level as well. In general, Frege cases are situations in which an agent satisfies the antecedent of a psychological generalization, but fails to satisfy the

[16] For more on intentional-level Frege cases see: Schneider (2005), Fodor (1994), Aryo (1996), Aydede and Robbins (2001).

consequent because the theory treats mental representations as being type identical that are actually causally distinct in the way the mental representations function in the system's cognitive economy.

It is crucial to note that Frege cases at the computational level cannot be solved in the way that some have tried to solve intentional level cases. Both Jerry Fodor and myself have tried to solve intentional-level Frege cases by saying that they are included in the *ceteris paribus* clauses of intentional laws, because they are tolerable exceptions (Fodor 1994; Schneider 2005). In making such an argument one explains that there was a difference in underlying LOT states, and that the individual didn't realize that the states referred to the same individual. Without going into detail concerning the intentional-level cases, let me simply observe that even if this strategy works in the intentional case, the crucial thing to note is that unlike Frege cases arising at the intentional level, there is no relatively lower psychological level to appeal to distinguish the relevant states. That was the job of LOT. And we've just supposed that the symbol tokens are type identical.

So molecularism falters as an explanation of the causal ancestry of thought and behavior (Role Two), for on this conception of symbols, there will either be missed predictions or counterxamples to computational generalizations. Clearly, if molecularism is to be effective it must respond to this problem. In addition to this, as discussed, it requires a plausible argument for regarding certain computational roles as essential to the symbol, while disregarding others.

4 Conclusion

The elephant sits in the room, dead center. Enormous and ugly, its grey flesh presses onto the walls.

Surely we should reject LOT if no conception of a symbol is forthcoming. Without symbols the roles go unsatisfied, and LOT will be too impoverished to provide its best wares. For instance, recall that MOPs characteristically distinguish coreferring thoughts. If the identity conditions on symbolic MOPs are not specified one clearly doesn't know whether differences in MOP will be finely grained enough to differentiate corefering thoughts. Further, without an individuation condition how can the proponent of LOT be confident that symbols are ultimately part of the domain that science investigates? Indeed, why would one want to claim that symbols exist at all if no individuation condition is forthcoming? *Ceteris paribus*, is better to opt for a theory of mind with a well-defined notion of a mental state. For the proponent of LOT to uphold naturalism without making progress on symbol natures looks like a hidden appeal to the sort of mysterious mental entities that one wants to avoid in the vicinity of Rutgers.

Today's project is negative—*very negative*—but I would urge that the proponent of LOT consider the following possibility. For I confess that I have been playing devil's advocate. Elsewhere I've provided a theory of the nature of symbols, arguing for the individuation of symbols in terms of total computational role. Herein, my aim was to convince you of a neglected, yet quite severe problem that LOT faces in lacking a theory of symbols. I sought to reveal the elephant in the room. That's enough for today.

But if you care to, we could try to evict it. But I do not want to advertise falsely—some would say that doing so has significant costs. Earlier, I briefly discussed the troubles that total computational role theories face. Yet elsewhere, I've argued that this manner of individuating symbols is actually required by the LOT theory, like it or not (Schneider 2009b). And I've provided three arguments for this position. The first of these arguments claims that Classicism requires that primitive symbols be typed in this manner; no other theory of typing will suffice. The second argument contends that without this manner of symbol individuation, there will be computational processes that fail to supervene on syntax, together with the rules of composition and the computational algorithms. The third argument says that cognitive science needs a natural kind that is typed by total computational role. Otherwise, either cognitive science will be incomplete, or its laws will have counterexamples (Schneider 2009b).

Both Fodor and Prinz have responded to these three arguments in the same way, providing the aforementioned Publicity Argument.[17] For them, this notion of a symbol is simply too costly, trading away LOT's capacity to explain how groups of individuals think and behave in similar ways (Role Three). I have offered four replies to this. (i) The first reply disentangles psychological explanation, which mainly proceeds by functional decomposition, from the too strong requirement that psychological explanation requires that systems literally have symbols of the same type. Functional decomposition does not require that two systems have any of the same LOT symbols in their database. (ii) Second, according to Fodor and other proponents of LOT/CTM, explanation that is sensitive to the broad content of the mental state plays a crucial role in cognitive science. But as with functional decomposition, intentional explanation covering different systems can occur without the systems having the same internal symbolic states. (iii) Third, I've argued that generalizations involving LOT states do not, by and large, quantify over particular symbol types; rather, they only quantify over symbols in general. So different individuals frequently do fall under the same generalizations in virtue of their LOT states. (iv) And finally, the only situations in which LOT symbols need to be subsumed under laws with respect to their particular symbol types (as opposed to being subsumed in virtue of simply having some LOT symbol or other) involves explanation which, by its very nature, involves the detailed workings of a particular system. And in such situations, it is inappropriate to call for symbol types that are shared across distinct sorts of systems (Schneider 2009b). Thus, as far as I can tell, Fodor and Prinz's charge of "publicity violation" is entirely benign: the publicity failure pertains to the states, but it doesn't extend to the actual cognitive explanations themselves. Symbols can still function in causal explanations.

My position is admittedly controversial, however, and clearly, an elaboration and defense would extend well beyond today's discussion. For today, I am satisfied if the reader is convinced that a very serious problem is at the heart of the LOT program. This is worth doing in its own right.

Acknowledgment Thanks very much to Jerry Fodor for many intriguing conversations on these topics.

[17] Fodor, personal correspondence and discussion. Prinz, personal correspondence. This issue is also discussed in Aydede (1999).

References

Anderson, J. (2007). *How can the human mind occur in the physical universe*? Oxford: Oxford University Press.

Aryo, D. (1996). Sticking up for oedipus: Fodor on intentional generalizations and broad content. *Mind and Language, 11*(3).

Aydede, M. (1999). "On the type/token relation of mental representations", *Facta Philosophica: International Journal for Contemporary Philosophy, 7*(12).

Aydede, M., & Robbins, P. (2001). Are Frege cases exceptions to intentional generalizations? *The Canadian Journal of Philosophy, 31*(1).

Braun, D. (2001). Russellianism and prediction. *Philosophical Studies, 105*, 59–105.

Crane, T. (1990). The language of thought: No syntax without semantics. *Mind & Language, 5*(3), 187–212.

Elman, J. (1998). Generalization, simple recurrent networks, and the emergence of structure. In M. A. Gernsbacher & S. Derry (Eds.), *Proceedings of the 20th annual conference of the cognitive science society*. Mahway, NJ: Lawrence Erlbaum Associates.

Fodor, J. A. (1975). *The language of thought*. NY: Thomas Crowell.

Fodor, J. A. (Ed.). (1989). Substitution arguments and the individuation of belief. In *A theory of content and other essays*. Cambridge, MA: MIT Press, 1990. (Originally appeared in Boolos, G. (Ed.), *Method, reason and language*. Cambridge, UK: The Cambridge University Press, 1989.)

Fodor, J. A. (1994). *The elm and the expert: Mentalese and its semantics*. Boston: MIT Press.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.

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

Fodor, J. A., & LePore, E. (1992). *Holism: A shoppers' guide*. Oxford: Blackwell.

Fodor, J. A., & McLaughlin, B. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition, 35*, 183–204.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis (also in *Connectionism: Debates on psychological explanation* (Vol. 2). Eds. by C. Macdonald & G. Macdonald. Oxford: Basil Blackwell, 1995).

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2002). *Cognitive neuroscience*, 2nd edn. New York: W. W. Norton & Company.

Harnad, S. (1990). The symbol grounding problem. *Physica D, 42*, 335–346.

Hawkins, J. (2005). *On intelligence*. New York: MacMillan.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy, 61*, 343–377.

Loewer, B., & Rey, G. (Eds.). (1993). *Meaning in mind: Fodor and his critics*. Oxford, UK; Cambridge: Blackwell.

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

Marcus, G. (2001). *The algebraic mind*. Boston: MIT Press.

Millikan, R. G. (1993). On mentalese orthography. In B. Dahlbom (Ed.), *Dennett and his critics: Demystifying mind* (pp. 97–123). Cambridge, MA: Blackwell.

O'Reilly, R., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. MIT Press.

Pessin, A. (1995). Mentalese syntax: Between a rock and two hard places. *Philosophical Studies, 78*, 33–53.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 23*, 73–193.

Prinz, J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, MA: MIT Press.

Richard, M. (1990). *Propositional attitudes: An essay on thoughts and how we ascribe them*. Cambridge: Cambridge University Press.

Schneider, S. (2005). Direct reference, psychological explanation, and Frege cases. *Mind and Language, 20*(4), 223–447.

Schneider, S. (2009a). The language of thought. In P. Calvo & J. Symons (Eds.), *Routledge companion to philosophy of psychology*. NY: Routledge.

Schneider, S. (2009b). The nature of primitive symbols in the language of thought: A theory. *Mind and Language, 24*(5).

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*(3), 417–457.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*.

Smolensky, P. (1995). Reply: Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In C. Macdonald & G. Macdonald (Eds.), *Connectionism: Debates on psychological explanation* (Vol. 2). Basil Blackwell, Oxford.

Stich, S. (1983). *From Folk psychology to cognitive science: The case against belief*. Boston: MIT Press.

van Gelder, T. (1990). Why distributed representation is inherently non-symbolic. In G. Dorffner (Ed.), *Konnektionismus in Artificial Intelligence und Kognitionsforschung* (pp. 58–66). Berlin: Springer-Verlag.