ABSTRACT

Title of Dissertation: THE EXPLANATORY ROLE OF INTENTIONAL CONTENT IN COGNITIVE SCIENCE

Andrew Charles Knoll, Doctor of Philosophy, 2015

Dissertation directed by: Professor Georges Rey, Department of Philosophy

This work argues that intentional content plays at least two explanatory roles in cognitive science. First, it allows cognitive states to be type-individuated independently of their relations to one another and to mind external phenomena. Secondly, it allows for counterfactual preserving generalizations over states so individuated. Thus, intentional content does not play this explanatory role in highly encapsulated cognitive processes. By contrast, it is necessary to type individuate states that partake in isotropic cognitive processes.

This work thus cuts a middle path between those who would eliminate intentional content from cognition altogether, and those who take it to be the 'mark of the mental.' Chapter 1 argues that there is no good reason to eliminate intentional content from cognitive science. But, it also argues that there is a coherent notion of computation without representation on offer as well. So, many cognitive processes could be explained as computations over states without intentional content.

Chapter 2 argues that many extant accounts of the explanatory role of intentional content end up being otiose. Too often, such accounts are concerned with capturing our intuitions about the proper way to talk about cognitive processes. But, in many cases, this talk can be eliminated from our explanations without loss of explanatory power.

Chapter 3 lays out the main argument. Many encapsulated cognitive processes—including early perceptual processes—can be explained in terms of computation without intentional content. In contrast, processes that are open to isotropic revision require their states to be individuated in terms of intentional content.

Chapter 4 surveys some objections to this view. One worry is that if cognition is massively modular, then all cognition must be non-intentional. On the contrary, modular processes can also be open to isotropic revision, and thus be amenable to intentional explanation.

Chapter 5 provides an example of such a modular process: the phonological system. It argues that states of the phonological system must be individuated in terms of intentional content. Phonological processing thus provides a case study for intentional explanation more generally.

THE EXPLANATORY ROLE OF INTENTIONAL CONTENT IN COGNITIVE SCIENCE

by

Andrew Charles Knoll

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Georges Rey, Chair
Professor Peter Carruthers
Professor Paul Pietroski
Assistant Professor Alexander Williams
Professor William Idsardi

Dedication

To Lynn Hammer and Gary Knoll.

Acknowledgements

Georges Rey is responsible for reforming my initial interest in aesthetics and convincing me that problems of intentionality might actually be solvable in my lifetime. Since putting this bee in my bonnet, he has been unfailingly generous with his time, thoughts, good conversation and support. I'll be forever indebted for his influence on how I see the world and approach philosophy.

Carsten Hansen generously welcomed me to the Centre for the Study of Mind in Nature at the University of Oslo for several stays over the course of writing this dissertation. Conversations with him and his colleagues-- as well as excellent Norwegian espresso-- greatly stimulated my thoughts and writing.

Peter Carruthers has patiently and generously provided foils to my position throughout my time working on this dissertation. His work has expanded my own thinking about issues of mental representation and his objections to my position have greatly improved this dissertation.

John Collins has been a tremendously thoughtful reader of much of my work. In addition, his own work on intentional content has been a helpful model for my own.

Steven Gross provided some very useful conversations in the incipient stages of this thesis that helped get the ball rolling.

Many others have also provided very useful feedback on earlier drafts of the dissertation. Notable among these are Jessie Munton, David Pereplyotchik, J. Brendan Ritchie, and Sebastian Watzl.

In addition, I've been fortunate to have the good counsel of a number of philosophically inclined linguists and linguistically inclined philosophers, including David Adger, Nicholas Allott, Naomi Feldman, Bill Idsardi, Paul Pietroski, and Alexander

Williams. Their perspective has been invaluable to developing not only my ideas on language, but on cognition more generally.

Table of Contents

Chapter 1: The Question of Explanatory Role

1. *The Horizontal Project*

In 1975, Jerry Fodor first laid down the rules: No Computation without Representation. Thinking is computation. Computation is the transformation of semantically individuated representations. So, if you want a computational theory of mind, it had better be a representational theory of mind. And we wanted a computational theory of mind, as it made for productive explanation. That representations had to come along for the ride seemed all for the better. A Computational Representational Theory of Mind (CRT) both vindicated traditional predilections for mental representations, and gave us an account of how minds could get those representations to do things we expected them to do, like compose inferences. It was an elegant theory.

But, from the start, CRT labored under a latent insecurity. To think was to implement computations. But, to implement computations was just to token states that causally interacted without regard to their semantic properties. Perhaps we could not have computation without representation, but representation didn't seem to be doing any explanatory work for us. It was an inefficacious chaperone-- there mostly for appearances' sake. After all, there were rules: No Computation without Representation.

Not caring much for appearances or social convention, many philosophers suggested we could dispense with representations altogether. If we could make all the psychological generalizations we wanted in terms of the local syntactic properties of mental states, then there seemed no need to appeal to any semantic properties those states might have in propounding our psychology.

Sure, maybe computational explanation appealed nominally to "representations." But it would be a mistake to think that such "representations" had semantic properties-- that is, had intentional content, were about things such that they could be tokened correctly or incorrectly. Since computations operated solely upon syntactic properties, it seemed as though you could switch out one intentional content for any other without loss of generalization. As long as the syntactic properties of your mental states stayed the same, your explanation was unaffected. For the purposes of psychological explanation, any intentional properties so-called "representations" might have just didn't matter.

Meanwhile, others were convinced that psychology did need representations-- and not just of the nominal variety. Representations indeed had intentional content, were about things. Champions of this view were largely concerned with how we could have states with such intentional content. They spent most of the last 30 years trying to provide a respectable naturalistic reduction of intentional properties. So, debate raged both externally and internally. On the one hand, was the war in which eliminativists such as Stich, Dennett, Egan, and Chomsky arrayed themselves against realists like Dretske, Fodor, Block, Millikan, Neander, and Cummins. Meanwhile,

factions within the realist camp, scrapped over how best to reduce intentional properties to a non-intentional supervenience base.

This left debate over a central question somewhat neglected. If semantic properties were playing an explanatory role, just what was it? Only recently have philosophers begun to start tackling this question head on. Ramsey (2007) addresses this project as the "job description challenge." Rey (1996) calls it the "horizontal project"-- in contrast to the "vertical project" of reduction. The purpose of this thesis is to stake a claim in this horizontal project. The goal is to get clearer on just what explanatory role intentional content plays in cognitive science, independently of how it might get reduced to non-intentional terms.

Of course, just because the horizontal question has been somewhat neglected is not to say that parties to the debates over intentionality have had nothing to say about it. The eliminativist charge had it that intentional properties play no explanatory role. Appeal to intentional states was a mere *façon de parler*. After all, we often also use intentional idiom to good effect in other domains in which intentional properties don't seem to actually be playing an explanatory role. Water seeks its own level. Lightning wants to get to ground. Sunflowers follow the sun. Electrons try to fill their valence shells. Might not appeal to intentional properties in cognitive science be similarly rhetorical?

The first claim of this thesis is an affirmative answer to this question-- followed up quickly by the qualification that the attribution of intentional states is in other cases substantive and indispensable. Intentional idiom in cognitive science is in

some cases a mere *façon de parler*; but, in other cases, it's a substantive attribution of intentional content that does explanatory work.

This claim follows Burge (2010) in arguing that both the eliminativists and realists were wrong-- or correct, as the case may be. Some psychological processes do require intentional content for explanation, whereas others get along fine making generalizations over syntactically individuated computational states without appeal to intentional content. Tracing out the conditions under which intentional content is explanatorily efficacious, and characterizing what explanatory role it plays is the project the remainder of this thesis pursues.

As such, this project cuts across most of the philosophical debates over the use of intentionality in cognitive science throughout the last 30 years. It remains neutral on questions as to how best to reduce intentional properties. And, it stakes out a claim at odds with both eliminativists and also realists, who would have it that all mental states are intentional.

Thus, my position faces challenges from all sides. On the realist side, Fodor (1975; 1980) has argued that you just can't have computation without intentional content. Section 4 of this Chapter, following the lead of Piccinnini (2006; 2007), lays out a notion of computational explanation devoid of representation. Against Fodor, it argues that such a notion can in principle play a role in psychological explanation. On the other side, eliminativists have argued that there's in-principle no role for intentional content to play within a computational psychology. Section 3 of this

present Chapter surveys reasons why arguments for eliminating intentionality entirely from cognitive science fall short of the mark.

Since I argue that intentional content does play a role in some psychological explanations, I of course disagree with the eliminativists' ontology: in order to play such a role, intentional states must exist! Moreover, I am not committed to drawing the ontological inference that if attributing intentional content to a psychological state does no work in explaining the operation of that state within a particular psychological process, then the state has no intentional content. A state may have intentional content even if that content plays no explanatory role in some contexts[1].

There might, for example, be a mental state that has intentional content that is explanatorily efficacious in accounts of object recognition processes, but which does not factor into the explanation of how that same state is involved in upstream spatial frequency processing. Whether cognition traffics in such states that have intentional content that is efficacious in some cognitive processes but not in others is a supremely interesting question-- and a very hard one. It requires an account of, amongst other things, how precisely to individuate mental states across cognitive domains often studied in abstraction from one another. It also requires some account of the role that intentional content plays in psychological explanation. I take the present project,

---

[1] I'm enough of a Quinean to suppose that if we need *never* appeal to the intentional content of a state in our best explanations, then we ought conclude that the state has no such content. But, the present project remains neutral on how we ought best pursue metaphysics.

therefore, to be a precursor to tackling this harder question of how to integrate psychological explanations that don't make use of intentional content with those that do.

In any case, I'm not pursuing an ontological project of sorting mental states into those that possess intentional content from those that do not possess intentional content. I'm merely attempting to clarify, to the extent that mental states might have intentional content, under which conditions that content play an indispensable explanatory role, and under which conditions it does not. Thus, I prescind from the ontological aspects of the eliminativist-realist debates that were bent on sorting out just which states have intentional content if any.

Nonetheless, the project is firmly metaphysical rather than epistemic. It is emphatically *not* concerned with characterizing what the state of our *knowledge* must be such that we might be *justified* in attributing intentional content. Rather, the goal is to characterize a metaphysical matter of fact: namely, under what conditions intentional content plays a crucial explanatory role in cognitive science.

To be sure, there may be some notions of "explanation" that take facts about explanations to be facts about our epistemic capacities. I agree there may be *some* such interesting sense of "explanation" that may open up epistemic questions as to, for example, under what conditions we are justified in trading different epistemic virtues off one another in our explanations. I do not adjudicate such questions here. Instead, I assume, along with Burge, Pylyshyn, and others in the tradition, a general

answer to such questions, dubbed "Pylyshyn's Razor" (Devitt, 2006a): when we can eliminate intentional properties without loss of explanatory power, it's best to do so.

There are nonetheless interesting metaphysical questions about explanation that need not wait on fully fledged commitments on these epistemic questions. For example, I take it to be a straightforward metaphysical truism that quintessence plays no crucial role in our best physics. Once upon a time, of course, it did. And, changes in our epistemic position might well justify us taking it to do so again in the future (though this is doubtful). We can settle the metaphysical question of whether the explanatory power of our best theories survive elimination of quintessence without getting too bogged down in epistemological concerns about tricky conditions under which there might be explanatory tradeoffs between multiple theories, some of which posit quintessence and some of which do not. Given that *none* of our viable physical theories lose generalizations with elimination of quintessence, these epistemic concerns just aren't germane to the metaphysical question.

Since I do not favor eliminating intentionality from cognitive science altogether, there may well be some tricky cases in which positing or eliminating intentional content hinges on such subtle tradeoffs of epistemic virtues. All I wish to argue at present is that there is a significant subset of cognitive explanation that are *not* subject to such subtle concerns. There are many conditions in which it is clear that our theories' explanatory power unambiguously survives elimination of intentional properties.

The present project also largely prescinds from the internal dispute amongst the realists about how best to reduce intentional properties to non-intentional properties. Curiously, defenders of intentional content have rarely tackled this horizontal project head on. In pursuing the vertical, reductive project, theorists such as Cummins, Gallistel, Millikan, and Neander have made assumptions about the explanatory work played by intentional content, though these assumptions are often left implicit in their discussions. Chapter 2 makes these assumptions explicit and argues that most of them are untenable. In many cases, the explanatory work these theorists take intentional content to do could easily be done by non-intentional, computationally individuated states. Nonetheless, I allow that any of these theorists may be correct in their claims about what constitutes a state as intentional. My claim is merely that many such accounts of content constitution may attribute intentional content that does no explanatory work within psychology.

Burge (2010) is perhaps the first large scale attempt to undertake the horizontal project directly. His work has laid out a useful framework for adjudicating the explanatory efficacy of intentional states. Though I ultimately reject his answer to the question, I adopt his methodology throughout what follows.

2. *Intentionality*

Burge points out that a key feature of intentional states is that they have *correctness conditions*, such that an intentional state can be tokened correctly or incorrectly. I take correctness conditions to be a way of fleshing out more precisely

just what it means to claim that a certain state is *about* something else. As Collins (2009, p.269) puts it:

> For a state to have content [x], there must be conditions C and C* such that [x] is satisfied under C and not satisfied under C*

We can call the conditions under which [x] is satisfied the correct conditions and those in which it is not, incorrect conditions.

Notice that the correctness conditions I appeal to here may never either actually obtain-- or even possibly obtain in a metaphysical sense. We can cache out what it is for a state to have the intentional content UNICORN by noting that it is correctly satisfied by certain conditions in which there are unicorns, and incorrectly otherwise-- *even if it's metaphysically impossible for such conditions to ever obtain*. We can still characterize what it is for a state to be *about* unicorns in terms of the conditions in which the state would be correctly tokened.

Notice that this way of characterizing the aboutness of intentional content retains the notion of content originally championed by Brentano such that a mental state may have intentional content independently of whether that content is satisfied. As Rey (2003, p. 147) puts it:

> An (intentional) content is *however we are to understand x when we use the idiom "represent(ation of) x," but there is no x.*

This is the general notion of intentional content that I shall be addressing throughout the work. States have intentional contents insofar as they have correctness conditions that may or may not be satisfied. Moreover, such correctness conditions need not

even possibly obtain in order for a state to have the intentional content that they characterize.

Characterizing intentional content in these terms helps capture two features commonly associated with intentional states. First, we can be in error in our thinking. What I think to be a cow on a dark night may actually turn out to be a horse. We can make sense of error by pointing out that in this case the satisfaction conditions of the content of my thought failed to obtain. Furthermore, we can think about things that don't exist. I can hypothesize that fire is comprised of phlogiston even though it turns out that phlogiston doesn't exist. Or, I can evaluate theorems about triangles even though it is metaphysically impossible for a triangle to ever actually be instantiated[2].

Thus, the horizontal project I shall pursue will be to ask whether such a notion of intentionality has an explanatory role to play in cognitive science. Methodologically, I'll follow Burge's lead in asking whether attributing correctness conditions to mental states does any indispensable explanatory work. Insofar as intentional states just *are* states with correctness conditions in the above sense, discovering an explanatory role for correctness conditions is to discover an explanatory role for intentional content.

---

[2] Note also that this way of characterizing intentional content prescinds from debates concerning whether that content can be *reduced* either to relations with mind external properties or relations to other mental states. That a state has the content [unicorn] such that it is satisfied only by certain conditions in which there are unicorns leaves it open as to whether it *has* that content *in virtue* of its relation to unicorns or in virtue of its relation to other mental states. The correctness conditions may be external even though the state may have those correctness conditions solely in virtue of relations it bears to other mental states.

Chapter 3 will argue for a sufficient condition on the explanatory efficacy of intentional states. The claim is *not* that this is the *only* explanatory role played by intentional content; others may remain yet undiscovered. But, it is an important role. In short, the claim is that intentional content is explanatorily efficacious in generalizing over cognitive systems that admit of isotropic revision in Fodor's (2000) sense. That is, intentional content allows for generalizations over cognitive systems that are open to influence from an indefinite disjunction of cognitive states in a way that states that are individuated only in terms of their syntactic properties do not. By contrast, cognitive systems that are encapsulated such that their operations can be characterized in terms of the proximal stimuli that impinge upon them can get along with states individuated only in terms of their non-intentional, syntactic properties. More anon in Chapter 3 precisely why these distinctions hold.

Chapter 4 makes the case that there does indeed seem to be such a distinction between isotropic and encapsulated states at play within human cognition. It will further argue that the distinction does not map on to a distinction between finite-state versus read-write memory architectures, as suggested by Gallistel & King (2009), nor does it reduce to a distinction between personal and sub-personal processes. There may well be sub-personal modules encapsulated from more global belief fixation processes that are nonetheless amenable to intentional explanation.

Chapter 5 is a case study of one such process: the phonological system. The argument here mirrors the more general, abstract case that Chapter 4 lays out. Because the phonological system is subject to influence from an indefinite disjunction

of distal properties and other cognitive states, the best way to make generalizations over it is to individuate its states in terms of intentional content.

One more terminological issue to settle on before beginning in earnest. So as not to beg any questions, I'll use "mental state" throughout to refer to the computational states cognitive scientists theorize about, remaining agnostic as to whether such states have intentional content or not. Of course, Brentano (1874) and his followers have held that intentionality is the 'mark of the mental,' such that it is close to analytic that mental states are intentional. I don't have an opinion as to which of these linguistic conventions is *best* in some strong sense. But, obviously, adopting the Brentanian thesis from the start would be to beg the question. Sometimes one must just stipulate in order to get on with substantive argument.

There are, however, more substantive worries as to whether the project as I've characterized it here is even coherent. The remainder of this chapter will survey these worries and give reason to dismiss them. First, Daniel Dennett argues that distinguishing between metaphorical and substantive uses of intentional idiom, as I do, is just uninteresting and in itself insubstantive. Steven Stich, meanwhile, takes the distinction to be interesting, but argues that *in principle* there's just no room at all for intentional content to play an explanatory role within a computational theory of mind. Chomsky concurs with Stich, at least as far as linguistics goes, and adds some arguments of his own.

In all cases, I argue, the criticisms miss the mark. There is an interesting distinction to be made here between computational processes that operate over states

with intentional content and those that do not but can nonetheless be characterized with intentional idiom *as if* they have intentional content. There is no reason in principle why intentional content can't play an explanatory role in a computational theory of mind. And, to the extent that intentional content *doesn't* play such a role, there are better arguments for eliminating it from certain psychological explanations than those advanced by Chomsky and Stich.

3. Initial Objections

3.1 Dennett: An Empty Question

Dennett (e.g., 1987) holds that the present project is ill-founded. Asking whether intentional idiom is either a mere *façon de parler* or instead a substantive attribution that does explanatory work is an "empty question," in the terms of Parfit (1984). He takes the question I've posed here to be tantamount to asking whether the current Republican party is the very same as the party of Lincoln or not (cf. pp. 41-42). It might be useful to *think* of it as such for some purposes, but not for others. There is no fact of the matter as to whether the Grand Old Party bears some actual relation of identity to Lincoln's.

Dennett takes "intentional systems" to be systems that are fruitfully described using intentional idiom. As such, they comprise the examples concerning electrons seeking to fill their valence shells, as well as cognitive psychology. Of such systems, Dennett writes:

> ...distinguishing those intentional systems that *really* have beliefs and desires from those we may find it handy to treat *as if* they had beliefs and desires… would be a Sisyphean labor, or else would be terminated by fiat. (p. 22)

There may be a fact of the matter as to whether intentional *idiom* is useful in *describing* a given system, but there's just no fact of the matter whether intentional *properties* play an explanatory role.

Dennett comes to this conclusion largely because he takes all intentional content to be *derived* content. For Dennett, intentional systems derive any intentional content they might be said to have from us in one way or another. By contrast, many who take the idiom of cognitive science to be attributing intentional properties to mental states take the content so attributed to be *original content*. That is, mental states have intentional content independently of how *we* talk about, use, interpret, or otherwise interact with them. Of such original content, Dennett writes: "I have never believed in it and have often argued against it" (p. 288).

With this distinction in hand, we can precisify the question I'm posing here: what explanatory role does *original* intentional content play within cognitive science? Given that Dennett thinks original content does not exist, presumably he does not think that *that* question is "empty," as we had earlier worried! Instead, it would seem to admit of a straightforward answer: none. The question as to whether intentional idiom is playing a metaphorical or substantive attributive role is empty only if you take intentional content to just *be* the property of being fruitfully described with intentional idiom.

Dennett's own argument (pp.295-298) against original intentionality begs the question. He asks us to suppose that we build a robot capable of the same behavioral flexibility that allows us to survive from day to day. Such a robot, Dennett asserts, we intuitively take to be devoid of states with original intentional content (p.298). After all, he points out, the robot has been designed by humans to suit human purposes; any intentionality it might have must have derived from its designers.

But, this just begs the question against the champion of original intentionality. It's not clear that our intuitions-- or Dennett's, as they may be-- are correct. Indeed, it's not clear that you *could* build such a robot without somehow endowing it with states that have original intentionality. Fodor et al. need not be saddled with the claim that the mere fact a machine was made by humans for certain purposes *by definition* makes the device devoid of original intentionality. Instead, they could hold that certain states of the world have original intentionality, and you can build an artifact with such states by, well, putting them in it. One need only watch any a number of science fiction movies in which human designed machines put their minds to work on undermining the purposes of their creators to imagine this as at least a metaphysical *possibility*.

Of course, Dennett could argue that positing original intentionality, as a matter of fact, doesn't do any explanatory work. But, to take this position would be to stake out an answer to the very question we're now pursuing: just what explanatory role does original intentionality play in cognitive science? So, by Dennett's own

lights, that question is not *empty*. It's a substantive issue to which he answers that original intentionality plays *no* explanatory role.

3.2 Stich: Content is Explanatorily Otiose

Of course, it would be nice to have an argument for such eliminativist claims, and Stich (1983) obliges. His Syntactic Theory of Mind has it that psychological generalizations can range over *syntactically* individuated states without loss of generalizations.

The arguments he advances for this conclusion commit him to a very particular stance on how whatever putative content mental states have would be fixed. The argument is the following (cf. p. 165):

1) The Representational Theory of Mind individuates mental states in terms of their intentional content.
2) Intentional content is determined by the unique causal histories and sociolinguistic environs of a creature.
3) But, psychological explanation should respect a principle of *Autonomy*: it should explain the behavior of organisms in abstraction from their contingent causal histories or environmental contexts, solely in terms of states that supervene on the organism's internal constitution.
4) Individuating mental states in terms of intentional content thus violates Autonomy.

Of course, this argument only holds if it's in fact true that intentional content is fixed in terms of a creature's contingent causal history and environment. But, intentional content needn't be fixed in such a way. An externalist like Fodor (1987) would have it that the content of a state is fixed by its counterfactual causal

dispositions. An internalist, like Block (1986), has it that the content is fixed by particular computational roles that the state enters into. The argument might be troublesome for an "anti-individualist" like Burge, who does indeed individuate content in terms of creatures' contingent environmental situation. But, there is no reason on the face of it that a proper horizontal theory of content must commit itself to Burge's anti-individualism as the proper story about how content is vertically determined.

In fact, one of the nice features of pursuing this horizontal project is that it allows us to abstract away from debates concerning the vertical project-- how intentional content might be reduced to non-intentional properties. Indeed, such a reduction is interesting only insofar as it allows us to reduce whatever intentional states actually play an explanatory role in our best science! So, Stich's argument is not damaging to the overall project of establishing an explanatory role for intentional content, independent of whatever vertical, reductive gloss we might give on that content.

Stich does go on to press a point more trenchant to the horizontal project as a whole. If intentional content does play an indispensable explanatory role in cognitive science, it had better be able to capture generalizations that we cannot make solely in terms of mental states individuated syntactically. Stich argues that intentional content doesn't buy this explanatory power, and moreover, arguments to the contrary are insufficient.

Stich (pp. 171-178) considers an argument from Pylyshyn (1980) to the effect that intentional content is necessary to make certain psychological generalizations. Pylyshyn argues, by way of example, that a purely syntactic theory of mind could not explain why Mary ran out of a smoke filled building. More precisely, it could not explain why she would exhibit such behavior under a "strange collection of diverse circumstances." Her running might be eventuated by merely smelling smoke, or instead receiving a phone call telling her the building is on fire, hearing a fire alarm, etc. Pylyshyn argues:

> the reason the same [mental] symbols occur under such diverse circumstances is precisely that they represent a common feature of the circumstances-- a feature, moreover, that is not to be found solely by inspecting properties of the physical environments (E.g., what physical features do telephone calls warning of fire share with the smell of smoke?) What is common to all these situations is that a common interpretation of the events occurs-- an interpretation that depends on what beliefs Mary has about alarms, smoke, and so on… a semantic interpretation of the symbols as representations of something (p.161)

I quote Pylyshyn at length both because his argument isn't entirely perspicuous, but also because my own argument in Chapter 3 will be an expansion and elucidation of what I take to be his basic point. For now, I want only to counter Stich's immediate objections to Pylyshyn's approach.

Stich argues that mental states individuated purely in syntactic terms could easily explain why Mary would run from the burning building in all these diverse circumstances. He first (p. 174) asks us to presume that each token string of states that is semantically individuated under Pylyshyn's explanation has a corresponding

token string of states that are syntactically individuated. We should note that this is precisely what representational realists (e.g., Fodor 1987) standardly assume! Any *token* transformation of semantically individuated states is implemented by some *token* syntactic process.

Stich goes on, however, to suppose that these token syntactic states can be *type*-individuated in such a way such that each syntactic type corresponds with a semantic type. So, the belief type individuated by the content [the building is on fire] has a corresponding syntactic state type-individuated syntactically as, say, an 'F' state. The desire individuated by the content [if the building is on fire, then leave the building] corresponds to a syntactic state type-individuated as an 'F→ L' state.

But, it's altogether unclear *why* Stich can safely assume that we can type-identify syntactic states in a way that parallels the type-identity of intentionally individuated states. As we'll see in Section 4 below, syntactic states are usually type-individuated as such in terms of their relations to one another, or in terms of their physical shape.

As Fodor (1987, p. 140) points out, we could well have semantic processes that supervene on syntactic processes. So, yes, any token semantic process would be implemented by a token syntactic process. But, the very same semantic process could be implemented by a disjunction of type-distinct syntactic processes. So, it's an open possibility that each time Mary runs out of a burning building, she's implementing a type-identical semantic process. But, each time, that very same semantic process is implemented by type-*distinct* syntactic processes. So, sure, if all we want to do is

explain one-off why Mary fled the burning building in one particular instance, we can simply appeal to the token syntactic process that caused her behavior. But, if Mary has a history of running out of burning buildings, and we suspect she'll continue in this regard, it's unclear that we can make this generalization without appealing to an intentional characterization of Mary's behavior. After all, each instance of building flight--past and present-- might in principle be implemented by type-distinct syntactic processes.

Of course, whether intentional processes supervene on disjunctions of type-distinct syntactic processes *as a matter of fact* is a matter of empirical investigation. One way of construing the present project is to think of it as sketching out the circumstances under which such supervenience relations occur and those in which they don't. I take it Stich is correct that *if* we can type-identify intentional states with syntactic states, there is no reason to suppose that intentional properties are doing any explanatory work. It's precisely this situation that I argue in Chapter 3 obtains for some encapsulated processes in early perceptual processing. In those instances, we can adopt Stich's argument to conclude that intentional content plays no role in explaining these processes.

Fodor's point here is just that neither Stich nor I can simply help ourselves to the claim that intentionally individuated states can be type-identified with syntactically individuated states. Indeed, as I'll also argue in Chapter 3, there is good evidence that certain cognitive processes open to isotropic revision *can't* be so type-identified. Because the mental states in isotropic processes are open to influence from

an indefinite disjunction of other computational states, it may be difficult to type-individuate them in terms of their relations to one another. That is, it may not be possible to type-individuate such states syntactically in such a way that allows us to attribute type-identical states to the same person at different times, or across persons at the same time. Thus, we would lose generalizations concerning why a person engages in a behavior in a variety of circumstances, or why multiple people from different backgrounds all engage in the same behavior.

We can read Pylyshyn's argument above as gesturing at this claim. In each instance of Mary's building bolting, the mental states that precipitate it have different proximal causes. In one case, it's the smell of smoke, in another it's the sound of a voice, etc. Moreover, because the proximal causes are so diverse, the structure of the syntactic processes leading to Mary's behavior is similarly diverse. In one instance, she must access states related to sounds and linguistic interpretation, and the general trustworthiness of various individuals who might tell her things over the phone. In another, she must access states associated with smells and combustion processes etc. If mental states are to be individuated syntactically-- i.e., in terms of their relations to one another-- it's unclear why the states that eventuate in Mary's fleeing the building should be type-identical in each of these instances. After all, in each instance, the states that cause her fleeing seem to have very *different* relations to one another.

So, we have prima facie reason to suppose that we cannot always type-identify intentionally individuated states with syntactically individuated states as Stich seems to presume. However, Stich does open up the possibility that intentional

content may not play an explanatory role in cases in which we can type-identify any states the explanation picks out with intentional idiom with corresponding syntactically individuated states.

3.3 Chomsky: Gratuitous Externalism

The most recent-- and perhaps unexpected-- proponent of intentional eliminativism has been Noam Chomsky (2000). As Rey (2003) points out, Chomsky seems to be caught in the grip of the same overly externalist, "anti-individualist" notion of intentional content that bedevils Stich and Dennett. Both assume that intentional content is determined by the *de facto* causal relationships that obtain between mental states and the environment. For example, Chomsky writes:

> [S]tudies of determination of structure from motion used tachistoscopic presentations that caused the subject to see a rotating cube, though there was no such thing in the environment; "see," here is used in its normal sense, not an achievement verb. ...There is no meaningful question about the "content" of the internal representation of a person seeing a cube… [whether] the retina is stimulated by a rotating cube, or a video of a rotating cube…[or tachistoscopic presentations]" (2000, p. 159).

Here, the argument seems to be that it makes no sense to distinguish instances in which a mental state is in fact caused by a cube from those in which it is not. Calling the former "representations of a cube" and the latter something else is not a distinction that adds to the counterfactual generalizations of our theory. For, we can

identify the operations of the mind as type-identical despite changes in the distal stimuli that bring them about.

We can grant Chomsky this (quite reasonable) point. But, that mental states are not type-individuated in terms of their *de facto* causal antecedents is not sufficient to demonstrate that they lack intentional content. Remember, the notion of content we are working with is one in which a state can have intentional content [X] even when there is no X!

Indeed, the very fact that theories of vision type-identify mental states despite changes in their distal antecedents is *prima facie* reason to think that intentional content might indeed play a necessary role in such theories. After all, it seems to be an open question in virtue of what we can type-identify the mental states eventuated by videos and tachistoscopic presentations of cubes, but *not* type-identify these mental states with those caused by videos and tachistoscopic presentations of *spheres*, for example. A plausible suggestion is that the mental states caused by presentation of cube-like stimuli are type-distinct from those caused by spherical-like stimuli in virtue of the fact that the former are *representations of cubes*, and the latter are *representations of spheres*-- whether or not either is caused by actual cubes or spheres.

Connecting some dots in Chomsky's argument, we can reconstruct his likely reply to this worry. Chomsky might argue that we can type-individuate these mental states in terms of the *proximal* stimuli that in fact give rise to them. Take for instance this passage:

> The auditory system doesn't "solve problems" in any technical sense of this term and, if they knew how to do so, the researchers might choose to stimulate the receptors directly instead of using loudspeakers-- much as they did with the computer model which, in fact, provided the main evidence for their theory of sound localization, which would work as well as for a brain in a vat as for an owl turning its head to face a mouse in the bush (p. 158).

Again, the argument seems to be that the brain doesn't employ intentional contents to solve problems because we can identify its operations as type-identical whether it's actually in an environment in which there is a problem to be solved or not. In this case, though, the argument is supplemented with the claim that we could generalize about the operations of the auditory system in terms of stimulations of the "receptors." *If* we had a story about just which type of proximal receptor stimuli caused various computational consequences, we could describe the operations of audition in abstraction from any environmental stimuli, or representations thereof.

In Chapter 3, I'll argue that something very like this *conditional* claim is correct-- but nonetheless it's a big *if*! It's not at all clear in many cases that we can give an account of the type of proximal stimuli that give rise to certain computational structures. For some operations, such a story does seem plausible. We'll see in Chapters 3 that many accounts of early vision processes *do* seem to make generalizations over proximal stimuli. States of the early vision system can be type-individuated in terms of the proximal stimuli that give rise to them. But, Chapter 5 will demonstrate that generalizations over the phonological perception system are *not*

amenable to such proximal characterization. Sure, phonologists may be able to describe particular contingent proximal stimuli that will give rise to particular phonological states under particular conditions. But, given the diversity of proximal stimuli to which such states are sensitive, it seems unlikely that they can *type-individuate* phonological states in terms of such proximal stimuli.

It's simply not sufficient for Chomsky to point out that a brain in a vat could be stimulated by electrodes in just such a way as to carry out the same process as it would if so stimulated 'in the wild.' Of *course* we can in principle reproduce any particular proximal stimuli in the vat that would have had antecedent distal causes in the wild. But, our psychological explanations aspire to characterizing the counterfactual conditions under which *different* proximal stimuli would give rise to that same cognitive process. It's not obvious that we can characterize those cognitive processes solely in terms of the proximal stimuli that give rise to them. We might need to identify them in terms of their intentional content.

Compare how Chomsky's argument might apply to eliminating a *computational* level of cognitive explanation. We could reproduce any token mental process in a brain in a vat if only we induced certain *physiological* effects in certain receptors. It does not follow from this observation that cognitive explanation consists solely of describing the physiology of the brain. Instead, cognitive science is predicated on the notion that it can describe mental processes at a computational level that abstracts from particular physiological implementations of such processes. In like manner, just because we can describe any *token* mental process in terms of abstract

syntactic relations between computational states, it does not follow that there aren't further generalizations to be made at an *intentional* level that abstracts away from token syntactic implementations. This is just to recapitulate the point raised against Stich's syntactic theory of mind in Section 3.2 above.

Curiously, Chomsky seems to take the fact that mental processes can be described computationally in a manner that abstracts from physical implementation to be evidence in *favor* of the view that intentional content is explanatorily inefficacious. He writes:

> recent studies show that if the optic nerve of an animal is "rewired" to connect to the auditory pathway early in life, "the auditory cortex gradually takes on a representation that is normally found in the visual cortex" (Weng et al. 2001); the "representation" is some internal structure R, which is used when the "rewired" animal performs "vision tasks with the auditory cortex." In such performance, R enters into complicated relations with things in the outside world, but it does not "represent" them in anything like the sense in which a photograph of a landscape is said to represent the landscape (2003, p.276)

Here, Chomsky seems to be pressing the point that mental states can be individuated not in terms of their relations to either proximal or distal stimuli, but rather in terms of the structural relations they bear to one another. We can individuate a collection of mental states as a type R in virtue of the fact that they implement a particular abstract computational structure.

From here, Chomsky's reasoning is at least two ways ambiguous. On the one hand, he seems to be pointing to the fact that the very same structure could be implemented by two different physiological substrates: visual or auditory cortex. It's

not clear why this truism should have any bearing as to whether the states have intentional content.

Perhaps instead, we're meant to focus on the fact that we can explain cognition in terms of the same structure, R, in abstraction from any of R's "complicated relations" to either proximal or distal stimuli. Perhaps, for example, we see a rotating cube if and only if R becomes activated. But, there's just no general story to tell about the circumstances under which R becomes activated.

But, as we pointed out earlier, there *do* seem to be generalizations about the circumstances under which R becomes activated insofar as there are generalizations about the circumstances under which we see rotating cubes. We see cubes when presented with cubes, videos of cubes, stachistoscopic presentations of cubes, etc.--but generally *not* when we're presented with various presentations of *spheres*! The negative evidence as to *why* we see cubes under the former conditions but not the latter wants explanation. It's unclear we can provide one without appealing to the intentional content of the states of R.

Now, we could supplement this story about individuating mental states in terms of their abstract computational structure with an account of the type of proximal stimuli that activate such a structure. Thus, we might explain that various distal presentations of cubes tend to activate mental states with structure, R, *because* they all give rise to proximal retinal stimuli of a certain type. Generally, distal presentations of spheres don't give rise to R because they don't cause the same type of proximal stimulation. Such an explanation would capture the generalization that a

diverse disjunction of cube-like stimuli gives rise to perceptions of cubes, whereas a diverse disjunction of non-cube like stimuli does not. Such an explanation may not have to appeal to intentional content. We might argue that people are able to visually discriminate cubes from spheres not because they *represent them as such*, but because cubes generally cause proximal stimuli of a particular type, which in turn gives rise to a mental structure R, which is structurally distinct from the mental structure the kinds of proximal stimuli caused by spheres gives rise to. In fact, borrowing a suggestion from Burge (2010), Chapter 3 will argue that just such a non-intentional explanation suffices to characterize the dead reckoning capacities of desert ants.

But, it's far from clear that such a strategy will apply to cognition generally. After all, we've already noticed that there may be many cognitive processes that cannot be characterized in terms of the type of proximal stimuli that occasion them. Moreover, it looks as if there are many cognitive capacities in which we cannot identify structures amongst mental states like R that remain stable across and within subjects. Fodor's (1987) critique of meaning holism is premised on the observation that many psychological generalizations aspire to hold across *differences* in the computational relations amongst mental states within and across subjects. It's just not obvious that two subjects who perceive a rotating cube both instantiate the same relationship R amongst their mental states. In any event, the mere fact that a bit of computational structure can be realized in either visual or auditory cortex is not sufficient reason to suppose that processes involving the structure do not depend on the intentional content of the structure's states.

4. *Computation Without Representation*

The previous section argued that arguments to eliminate intentionality from psychological explanation more generally miss the mark. On the other side, Fodorians hold to his (1975) dictum that there just is "no computation without representation." If our best theory of mind is computational, then of necessity it is intentional as well. This section argues against this claim to the conclusion that it's at least *in principle* possible to give explanations in terms of computations without intentional content.

One reason for thinking that computations must operate over intentional states is that the states over which computations operate can only be *individuated* in terms of intentional properties. This way of thinking goes along with Fodor's (1987) rejection of inferential role theories of semantics. If we want to make sense of the idea that different people can have different thoughts about the same things, we must be able to individuate the states that represent the objects of their thoughts independently of the computations those thoughts enter into. So, for example, if we want to make sense of the fact that Aristotle believed stars were holes in the heavenly fabric and that you believe stars are huge orbs of plasma undergoing nuclear fusion, it seems that you and Aristotle must share some state that is individuated independently of its computational relations.

These considerations may be sufficient to establish that *some* aspects of human cognition can only be explained via mental states with intentional properties that are not constituted by their computational roles. In fact, I do side with Fodor on

this point. Nonetheless, they leave on the table the possibility that *other* aspects of cognition are explained by computations over states that are *not* individuated by their intentional properties.

Of course, if this is the case, we want an account of how such states would be individuated and type-identified. Establishing the coherence of computation without representation thus rests upon giving an account of how computational states can be individuated independently of intentional properties. It's not *prima facie* obvious how to do this. For Fodor (1998), computations just *are* "causal relations among symbols which reliably respect semantic properties of the relata" (10). Thus, computation can only be characterized in terms of the semantic properties of the states over which it ranges.

4.1 Computational Implementation

One attempt to characterize computations independently of semantic properties has been Chalmers' (1994; 1996; 2011) account of computational implementation in terms of physical systems' causal properties. Now, strictly speaking, Chalmers' account on its own is not sufficient to respond to Fodor's concerns. Chalmers' theory gives conditions on a physical system *implementing* computations. Thus, it is an account of how we can specify the computations implemented by a physical system independently of any intentional properties the system might possess. But, strictly speaking, the account *presupposes* that we are already able to specify computational processes in the *abstract*. We're currently in need of a theory that does just that.

Nonetheless, Chalmers' account is a helpful first step on the way to developing a theory that type-individuates computations in the abstract, independently of intentional properties. The account itself is quite simple. The idea is that a physical system implements a computational process if there is an isomorphism between computation, abstractly specified and the physical system. Such an isomorphism holds when there is a one-to-one mapping between, on the one hand, the abstract computational states and the state transition relations holding amongst them, and, on the other, the physical states of the system and the causal relations amongst them. Thus, we can specify when a system implements a computation without reference to the intentional properties of either the physical system or the computation it implements.

The prime objection to this theory of computational implementation is that it entails pan-computationalism. Since isomorphisms are cheap, under Chalmers' account, just about anything-- a rock, say--would count as a computation. Piccinini (2006; 2007) and others (Searle, 1990) worry that accounts like Chalmers' erase the distinction between processes that are computations, *tout court*, as it were, and processes that are not, *strictu dictu,* computations, but nonetheless helpfully *modelled by*, say, a silicon chip computer that itself implements computations. For example, the motions of the planets can be modeled by a computational account. But, you might think we'd like a way to individuate the causal relations between states of the machine modeling the planets as computational, without allowing that the causal relations between the planets themselves are computational.

Sometimes Piccinini writes as though this is an important distinction to maintain so that in giving an account of computational individuation, we hew to our *intuitions* concerning the extension of our folk concept or word, "computation." For those of us for whom such considerations are not compelling, Fodor has an alternative objection to the pan-computationalism entailed by Chalmers' account. Fodor worries that if it turns out that just about all physical systems implement computations, then there's nothing substantive to the claim that the *mind* does. If everything is a computation, Fodor argues, the claim that *thought* is becomes trivial (1998, p. 10). This is indeed the conclusion that Putnam (1988) and Searle (1992) derive from their commitment to pancomputationism. If there is to be any substance to the claim that thought is computation, argues Fodor, pancomputationism can't be true. And, we likely can preclude pancomputationism only if we characterize computation intentionally.

Chalmers responds to this objection by adding on the requirement that the causal relations between physical states be *reliable* (2011, p. 331). Thus, while there may be a one-off mapping between causal relations between atoms in a rock and the state transitions in a computational structure, the rock will not count as implementing the computation because those causal connections are not *reliably* instantiated.

We don't want to rely on this notion of computation in explicating the computational theory of mind, however. That theory generally has it that the computational properties of mind *supervene* on physical brain states. If we require that physical brain states *reliably* instantiate the same causal relations in order to

count as implementing computations, we'd be scotching this idea of supervenience. One of the nicest things about the supervenience thesis is that it allows us to theorize about the computations performed by minds in abstraction from the details of how these are physically implemented by the brain. Crucially, it allows us to suppose that the very same computations can be performed at different times by very different physical brain processes[3]. All it commits us to is the thesis that minds reliably carry out computations that are implemented in any given instance by something or other. It leaves open the possibility that the mind reliably performs particular computations even though they may *not* be reliably implemented by the same physical brain processes.

Perhaps, in the end, it will turn out that the computations performed on the mind *do* map roughly one-to-one onto reliably re-occurring physical processes. But, there's no reason to let our theory of computation dictate that conclusion, particularly since Chalmers' account can address this objection without appeal to reliable physical processes.

One possible alternative is advanced by Chalmers himself (pp. 332-333). He proposes something like a distinction between isomorphisms that are constitutive of phenomena and those that are not. The idea is that even though both digestive

---

[3] Chalmers' stipulation would, of course, allow us to maintain that mental processes supervene on the brain in the sense that different brains and physical systems all implement the same computation. But, it would require that each of those brains have reliably instantiated causal processes that would constitute those computations.

processes and brain processes implement computations, the claim that brain processes do is more substantive because those computations go toward *constituting* those processes as cognitive. In contrast, the computations implemented by bowel processes do not constitute those processes as *digestion.*

We don't want to commit to some processes being essentially cognitive or essentially digestive. To claim, as Chalmers does, that "it is in virtue of implementing some computation that a system is cognitive" (pp. 332-333) is to put the cart before the horse. At any rate, it doesn't answer Fodor's objection that we need a notion of computation that can *explain* cognition.

If the relevant type of explanation were conceptual, then I suppose Chalmers' proposal would suffice. It specifies a sufficient condition on a system being cognition. As Fodor puts it in his (1975), if all we want from explanation are definitions, we can explain that Wheaties is the breakfast of champions simply by noting that it is what champions eat. But, we might also want to explain what it is about Wheaties that *causes* people to be champions. In that case, we want an account of how the vitamins and minerals in Wheaties cause champions to display phenomena we discover them to have, such as endurance, strength, etc.

Similarly, Fodor wants from a theory of mind not just a conceptual explanation that the extension of thinking things and computing things overlap. Rather, he takes it as a strength of the computational theory of mind that it provides a *causal* explanation that accounts for various properties we empirically discover it to have. The nice thing about the computational theory of mind is that it can explain,

inter alia, how mental states can be truth preserving, and how mental states allow organisms to exhibit certain behaviors.

Thus, Chalmers' attempt to vindicate computational explanation as non-vacuous by appealing to its facility for *conceptual* explanation doesn't fully address Fodor's concern. If the computational theory of mind just amounts to the claim that computation is constitutive of thought, but not, say, digestion, it really just entails that a lot more things are thoughts than we initially thought there were. So, insofar as stomachs implement computations, they are thinking thoughts, even though those thoughts by not be relevant to an explanations of how they *digest*. But, it's just such a conclusion that Fodor wants to avoid. If it turns out that just about everything thinks, the thesis that thoughts are computations is just as vacuous as the thesis that only minds think, that thinking is computational, but just about every physical process is computational.

We can easily revise Chalmers' account to allow for computational explanations to be special varieties of *causal* explanation. The key is to recognize that while isomorphisms between physical systems and computational structures may be ubiquitous, only some of those isomorphisms prove to be explanatorily useful. For example, we might allow that, in Chalmers' sense, a rock implements a particular Turing machine because there is a particular instance at which causal relations between states of the rock are isomorphic to the state transitions of the Turing machine. However, this isomorphism does not do any explanatory work. There are no generalizations about the rock that the isomorphism helps us explain.

Alternatively, in the context of cognition, the fact that there is an isomorphism between mental processes and a particular Turing machine may well prove explanatorily fruitful. If the mental states end up being intentional, then the isomorphism could explain how the mind is able to process them in a truth-preserving manner, as Fodor argues. If the input states have particular causal connections to the environment, then the isomorphism can explain how an organism interacts with its environment, as Gallistel (1990) and Cummins (1989) press. Rather than appeal to computations that feature in conceptual explanations, as Chalmers suggests, we can just distinguish between computations that feature in causal explanations and those that do not.

Of course, it might turn out that computations also so feature in explanations of phenomena other than cognition. For example, recent studies have had great success at modeling ant colonies as implementing computations, employing algorithms that allow for efficient search procedures (Pinter-Wollman et al., 2013). We need not argue either that the colony is thereby *thinking* (as Chalmers does) or-- to bar that claim-- argue that it's not actually implementing computations in order for our explanations to remain interesting. Surely, it's quite interesting that both minds and ant colonies implement computations! We can allow that cognition is computation, but that lots of other things may be as well.

This response still adequately responds to Fodor's worry. If it had turned out that cognition was computation only insofar as just about *everything* is, then it would indeed be a vacuous claim. But, thinking of computational systems as those in which

certain computations are (causally) explanatorily efficacious makes it a substantive hypothesis that brains are such computational systems. There may be other non-cognitive systems that are also computational in the same sense. But, in each case, it's a substantive matter whether *certain* computations are explanatorily efficacious or not. Against this background, it's a quite substantive claim that brains implement explanatorily efficacious computations.

4.2 Type-Individuating Computations

So, we can exonerate Chalmers' theory of computational *implementation* from the principal charge against it. It still remains, however, to demonstrate how this account of implementation can be parlayed into an account of computational *individuation* that is independent of intentional characterization but nonetheless does not fall prey to Fodor's objection that any such account will end up being trivial. First, we'll take a look at the attempts Piccinini and Rescorla have made at this project, only to reject them as being needlessly baroque. Rather, Chalmers' account of computational implementation can be modified to give an account of how to individuate computations independently of intentional states.

Piccinini himself proposes that within the purview of computer science, computation is the manipulation of "symbols [that] are typically marks on paper individuated by their geometric shape" (211). Piccinini seems to embrace this view of computational individuation in order to counter arguments to the effect that computations must be semantically individuated because the functions they compute

must be specified in intentional terms. Such an argument would have it that computations are individuated by the functions they compute, where those functions are individuated by the ordered pairs of objects that the inputs and outputs of the computation are *about.* According to this view, a computation is individuated semantically as an addition process because the input/output pairs it generates are something like <(1,2), 3>, <(2,3),5>, etc., where "1," "2," "3" are *about* the numbers, one, two, three, etc.

In order to circumvent this way of thinking about computation, Piccinini recommends that we define functions over the geometric marks rather than any entities those marks might be about. Thus, we individuate the above computation as operating over the geometric figures, '1,' '2,' '3,' etc. This is one way to circumvent a commitment to intentional properties being essential to computational individuation.

But, it's a needlessly baroque maneuver with unhappy consequences. For one, it entails that processes operating over symbols with different geometric properties are *de facto* different computations. Thus, a process that generates input strings such as: <(1,10), 11>, <(10, 11), 101> would be a different computation than that described above, despite the fact that "1," "10," "11," and "101," are just the binary analogues to "1," "2," "3," and "5". But, on the face of it, it seems like we'd want to treat these two processes as instantiating the *same* computation.

Of course, one way to do this is to identify the inputs and outputs of the computation as the objects *represented* by the symbols over which the two processes operate. In this case, we can talk about the two systems described above as

implementing the same computation despite making use of differently shaped symbols by claiming that these symbols are *about* the same set of entities (in this case, numbers). This is, of course, just the strategy that Piccinini wants to reject.

Fortunately, Chalmers' theory of computational implementation suggests another way of capturing what the two systems have in common without individuating their states in terms of intentional properties. In the first place, there is a one-to-one mapping between the numerals 1, 2, 3, etc. and the numerals 1, 10, 11, etc. Furthermore, the input-output pairs generated by the first process map one-to-one with those generated by the second process. So, for each input-output pair, were you to switch each symbol of the first process with its analogue symbol in the second, you'd end up with the same two sets of input-output pairs as before. That is, there's a one-to-one mapping between the input-output pairs of the two systems that preserves the relations amongst them. So, the two sets of input-output pairs are isomorphic.

The consequence of this is that we can abstract away from the two different systems described above in a way that doesn't require appeal to intentional properties of their states. We can say that they both instantiate the same computation in that they both generate a set of input-output pairs that are isomorphic to one another. Thus, the two systems are just two instances of a variety of systems that all fall into a class of systems individuated by this property of generating input-output pairs that are isomorphic to one another.

We can more finely individuate computational processes in terms of their algorithmic implementation. To distinguish between two different algorithms

implementing the same input-output transitions, we can stipulate that two systems implement the same algorithm if and only if the states that mediate the input and output are isomorphic as well[4]. Thus, we can avoid individuating computations in terms of intentional properties, without individuating them to such a fine grain as the geometric properties of their inputs.

Of course, Rescorla (2013) argues that the above considerations argue in *favor* of individuating at least some computational states in terms of intentional properties. He points out that the same input-output pairs can be thought of as different functions depending on how we interpret their meanings. For example, if we interpret the pairs given above as representing numbers in arabic base-two notation, then they do represent the addition function. However, if we interpret them as representing numbers according to standard arabic base-ten notation, they no longer do. If "10," "11," and "101" represent the numbers two, three, and five, respectively, as they do in

---

[4] *Strictu dictu*, the correct account needs to be slightly more subtle, given that there may be varying degrees of abstraction in characterizing algorithms. For instance, you might have two systems that generate the same input-output pairs and that have internal states that are *largely* isomorphic to one another. However, suppose the second system has a state with no analogue to a state in the first, but which makes no difference to the final output. For example, two systems could have homomorphic states (that is, there is an *injective*--- but not one-one-- function from one to the other) and generate the arabic numeral corresponding to the product of the base-10 arabic numerals that are input. But, the second system might also go into a state analogous to adding zero. In one sense, the two systems are algorithmically equivalent in that their states are homomorphic and their input-output pairs are the same. In another, they are not, since the second system goes through some extra machinations on its way to generating the output. This second sense is just a more fine-grained specification of the algorithm involved, and the former is a more abstract one. Some contexts might lead us to care that the two systems are algorithmically different in the less abstract sense, and other contexts might compel us to acknowledge them as equivalent at the more abstract level. In any case, we have the resources to characterize them as the same algorithm at either level of abstraction either by showing that they are isomorphic, or that they are homomorphic at whatever level of abstraction we find appropriate.

base-2 notation, then the function that outputs "101" when input "(10,11)" indeed instantiates addition. However, if the numerals are interpreted on the standard base-10 model, then the process that generates the input-output pairs no longer implements addition. For, the sum of the numbers ten and eleven is not one hundred and one.

We should grant Rescorla's point that it's sometimes helpful to individuate computational processes as *interpreted* computational processes, in which case, we may treat isomorphic structures of input and output as distinct given that they have different semantic properties. But granting this truth does not preclude the proposal that there is a level of abstraction at which we can individuate computational processes independently of their semantic properties.

Rescorla's argument indeed seems to assume that this is the case. His claim just is that, though two computational processes may fall into an equivalence class of formally isomorphic relations amongst primitives, if the intentional properties of those primitives differ, there is reason to distinguish them at another level of abstraction. So, Rescorla's argument doesn't establish that some computations cannot be *individuated* non-semantically-- it just establishes that sometimes it is helpful to further differentiate computational systems by way of their semantic properties. That claim is almost certainly correct. I'm sure there are many computations that operate over intentional states, just as there are many computations that do not.

Rescorla also pushes the idea that semantic individuation "facilitates homogeneous description of systems that are heterogeneous under syntactic and physical description" (p. 692). But, as we have seen, the isomorphism account also

allows for this level of abstraction. Systems individuated in terms of isomorphisms between their primitive components and the relations holding amongst those also abstract away from any particular account of the physical or geometric properties of those primitives.

But, now we have come back to where we started. Chalmers' isomorphism account of computation falls prey to the objection we raised at the beginning of this section. Namely, Chalmers' account seems to entail pan-computationalism. It leaves no room for a distinction between computational systems and non-computational systems that can nonetheless be modeled computationally. Since the modeled system and the computation used to model it would both be isomorphic with each other, both of them would be computations on Chalmers' view.

Of course, these considerations still leave on the table the possibility that there is a way to characterize computation non-intentionally that nonetheless does not entail a commitment to pan-computationalism. Piccinini's work is an attempt to defend this position. Ultimately, Piccinini's proposal is unsuccessful, but it will be instructive to see why.

We found unsatisfactory Piccinini's initial attempt to rescue the notion of non-intentional computation from the implication of pan-computationalism by way of individuating computational states in terms of their geometric properties. However, Piccinini (2007) does not seem to fully endorse the view that geometric shape is crucial for individuating computational symbols. He also relies on a mechanistic construal of what constitutes computational systems.

By a mechanistic account of some system, X, Piccinini means:

> a description of X in terms of spatiotemporal components of X; their function and organization, to the effect that X possesses its capacities because of how X's components and their functions are organized (p. 506).

In the abstract, computations consist of strings of letters from a finite alphabet and a list of instructions for generating new strings from old strings. For Piccinini:

> A letter is simply a type of entity that (i) is distinct from other letters and (ii) may be concatenated to other letters to form lists, called 'strings.' A string is an ordered sequence of letters (pp. 508-509).

Since mechanisms are to be specified in terms of spatiotemporal components rather than abstracta, Piccinini also individuates "digits" as "concrete counterparts" to abstract letters: "a digit may be a component or state of a component of the mechanism that processes it" (p. 510).

Digits are individuated as such by way of the *functional role* they play in a mechanism. However, it's unclear how this notion of functional role individuates the elements of computation in a way that's fundamentally different from the way Chalmers' proposal individuates computational states via their isomorphic relations to one another.

For instance, by way of explicating this functional individuation, Piccinini writes: "in an AND gate, all of input types '0,0', '0,1', and '1,0' give rise to outputs of type '0'" (p. 511). The idea seems to be that a digit is individuated as a type '1' or

type ‘0’ just in case it behaves in a way that comports with this AND gate structure. That is, two particular voltages can be individuated as digits of type ‘0’ and ‘1’ just in case they behave in a way that can be mapped on to the abstract description of an AND gate! But, the property of being able to be mapped on to such a structure just is the property of being homomorphic with that structure! It’s thus hard to see how Piccinini’s description of computation is significantly different from theories that individuate computational processes via homomorphisms.

Piccinini acknowledges that his account shares much with Chalmers’ when he writes: “digits and strings thereof are equivalence classes of physical entities or states” (p. 514). Presumably a digit or string falls into an equivalence class if it is homomorphic with the other members of that class, just as Chalmers suggests.

So, on the face of it, Piccinini’s account does not have much that distinguishes it from Chalmers’. Crucially, it does not yet have the resources to surmount the objection that entails that just about everything is a computer since there are homomorphisms to be had between computational structures and just about anything else. He can overcome the objection by retreating to his stance that computational states are individuated geometrically. But, as we saw, this stance individuates computations too finely for the purposes of empirical psychology.

Fortunately, there’s just no need to provide an account of computation that does not entail pan-computationalism. We can buy into the pan-computationalism entailed by Chalmers’ account as long as we can distinguish between computations that are explanatorily useful, and those that are not. So, it may be that just about

everything implements computations. But, some of those computations are useful to explanatory psychology because they allow us to give a causal explanation of the generalizations discovered by its enquiry.

Shagrir (2001) seems to take something like this view. He starts off by noting, as we have, that any physical system could be given multiple syntactic descriptions in something like Chalmers' sense. He proposes that we determine which of these possible syntactic descriptions constitute what he calls *the* computational structure of a physical system. Shagrir concludes, as have I, that the relevant syntactic descriptions are just those that feature in our best empirical generalizations about a system.

However, Shagrir further concludes that the intentional content of intentional states is essential to determining what those syntactic descriptions are. By contrast, I'm pressing the view that we can (in principle, in some instances) individuate computational systems independently of intentional content. So, we should see why it's unnecessary for Shagrir to appeal to intentional properties in his characterization of computational generalizations.

Shagrir's argument comes down to noting:

> We are therefore left with two sorts of features that could explain the choice of computational structure: environmental features correlated with the intrinsic physical/neural properties of the cognitive system, and phenomenal features (conscious experiences) correlated with these neural properties. But both kinds of features are precisely the ones we associate with the content of the system's states. Thus content impacts computational individuation. (p. 384)

We might allow that the computational processes that are explanatorily adequate are those that help explain how a creature interacts with its environment. Thus, in the case of the desert ant discussed by Gallistel (1990) and Burge (2010), it is the fact that there are computational processes implemented by the ant that are homomorphic to the spatial relations of its environment that helps explain how the ant is able to navigate the space of its normal environments. It may also well be true that connections between cognitive systems and the environment are involved in constituting the intentional content of cognitive states.

However, it's just not obvious that because physical states of a creature's neural system are correlated with certain environmental properties that those physical states therefore *represent* those environmental properties. In fact, in the following chapter, I'll argue that we can explain many phenomena by appeal to homomorphisms between computational states and environmental properties without appealing to intentional states.

Regardless, we need not establish this claim now to see that Shagrir's argument is misguided. We could concede, along with Gallistel and Cummins, that homomorphisms between computational processes and environmental properties *constitute* the intentional content of those computational states. But, it does not follow from this that the states cannot be *individuated* independently that intentional content. In fact, establishing that a functional isomorphism obtains between a computational system and the environment would seem to require individuating the computational

states and environmental states independently before going on to show that these two independently individuated systems are nonetheless homomorphic!

Thus, contra Shagrir, we can pick out some computational processes as explanatorily efficacious independently of their intentional content. If we adopt the thesis that computations are individuated as equivalence classes of isomorphic structures, along Chalmers' lines, we can satisfy Fodor's requirement that the thesis that the mind is computational is a substantive claim. While just about every physical process may be computational in Chalmers' sense, only some computational processes play an explanatory role. Thus, it's a substantive thesis that the operation of minds can be explained by the postulation that they are computers whereas the operations of rocks generally cannot be.

It would be further interesting if computations-- individuated independently of intentional properties were sufficient for capturing *all* the generalizations of our best psychology. Such a result would constitute the elimination of intentional properties from the posits of psychology. However, the present work argues that such an elimination is *not* available to us. It does argue that *some* psychological generalizations may be captured in terms of computation independent of intentional properties, whereas others *do* require the postulation of intentional states. Caching out the distinction between these two modes of explanation is the chief goal of this study.

Of course, we haven't yet demonstrated that computation either with or without intentional states is explanatorily efficacious. The present chapter merely establishes, *contra* Fodor and others, that it's *coherent* to postulate a notion of

computation without representation that could possibly do explanatory work in psychology. Chapter 3 will argue that it does in fact do such work and explicate just what that work is. This will allow us to develop an account of the explanatory work that intentional properties do over and above non-intentional computational processes. We'll then be in a position to see how this difference between computational explanations and intentional explanations influences how we should understand explanations in the cognitive sciences in Chapters 4 and 5

Before we move on to this positive account of the explanatory role of intentional content, however, Chapter 2 will look at existing proposals about the explanatory role of intentional states and point up what is unsatisfactory about them. We'll then be in a position to see why a new answer to the horizontal question of intentionality is necessary.

5. Cognition With and Without Representation: An Itinerary

So, the eliminativists don't give us reason to rule out an explanatory role for intentional content within cognitive science in principle. And, in principle, there is a notion of computation without intentional content on offer. But, if we are to admit intentional content as an explanatorily efficacious posit within cognitive science, then we owe an account of just what sort of explanatory role such content may play. The remainder of this work is devoted to sketching out such an explanatory role.

Chapter 2 argues that the explanatory roles for content assumed in the reductive projects of many intentional realists end up being in many cases otiose. Along lines that Chomsky gestures at, it argues that some overly externalist accounts

of intentional explanation do not allow for counterfactually robust generalizations. In particular, too often, attempts at vertical reduction took approaches seem too often concerned with capturing our intuitions about the proper way to *talk* about mental processes.

Chapter 3 continues with this analysis, presenting a series of example cognitive processes. It argues that a number of cognitive processes that occur early in both phylogeny and diachronic processing can be explained in terms of computation *without* intentional content. This is not *because* they are early systems. Rather, it's because such processes can be generalized over solely in terms of the proximal input to which they are sensitive and the computational transformations they perform on that input.

By contrast, processes that cannot be generalized over their proximal inputs are amenable to intentional explanation. That is, cognitive processes that are isotropic, in that they're open to revision from an indefinite disjunction of inputs, require their states to be individuated in terms of intentional content. We'll see that such processes also occur in early phylogenetic systems, such as the navigational capacities of honey bees.

Chapter 4 surveys some possible objections and alternatives to the view set out in Chapter 3. It addresses Gallistel & King's (2009) suggestion that intentional content crops up in systems involving read-write memory systems. It argues that this condition is neither a necessary nor sufficient condition for the efficacy of intentional content.

The chapter also addresses worries that depending on how the facts about cognition pan out, my proposal will either entail that *all* cognition is intentional, or that *none* of it is. On the one hand, recent psychological literature suggests that *all* cognition is cognitively penetrable. The worry is that there just aren't any bottom-up encapsulated processes that would be amenable to explanation via computation without intentional content. The chapter argues, along lines laid out by Firestone & Scholl (2015), that the evidence against encapsulated processes isn't conclusive.

On the other hand, you might worry that if cognition is massively modular, then my proposal entails that all cognition is non-intentional. But, the chapter points out that modular processes can also be open to isotropic revision, and thus be amenable to intentional explanation.

Chapter 5 provides an example of such a modular process in the form of the phonological system. It argues that states of the phonological system cannot be individuated independently of their intentional content precisely because they are open to revision from an indefinite disjunction of possible stimuli. Phonological processing thus provides an object lesson for intentional explanation more generally.

Chapter 2: Horizontal Positions in Vertical Accounts

1. *Introduction*

As we noted in Chapter 1, most of the literature on intentionality over the past century has been explicitly concerned with what Rey (1996) has called the "vertical problem" of intentionality. In its most abstract form, the vertical problem is how to reduce appeals to intentionality in cognitive science to non-intentional terms. So, at one level, it involves a debate pitting intentional reductionists against the behaviorists, various connectionists, and others who hold that intentionality can be eliminated from cognitive science altogether. The vertical problem also encompasses the debates within the reductionist camp itself concerning how intentionality ought to be reduced to non-intentional properties.

The present project is not concerned with this vertical problem, but rather what Rey terms the "horizontal problem," or Ramsey (2007) calls the "job description challenge": the problem of articulating just what sort of explanatory work intentionality does in cognitive science. While few philosophers have explicitly grappled with this problem, those working on the vertical problem have taken stances on the issue. After all, in order to make arguments as to whether their favored reductions will allow intentional states to do the explanatory work they're supposed to do, reductionists have to have some assumptions-- if not a fully fledged theory--

about just what work intentional states do.

The following chapter will examine these assumptions-- sometimes implicitly assumed, other times explicitly defended-- and find them all lacking. Though most theorists think of themselves as characterizing a notion of intentional content that does explanatory work, they sometimes overly constrain their accounts so as to accurately capture either our folk intuitions, or the *talk* of cognitive scientists, about when organisms are in intentional states. In so doing, many reductionists have failed to propose reductions that assign intentional states in such a way that they do any *explanatory work in psychology*. One notable exception to this general failing is the work of Jerry Fodor, which has always been concerned with capturing not our folk intuitions or scientific jargon, but rather a notion of intentionality that will do real explanatory work.

In what follows, I leave aside whether any of the theories canvassed are adequate answers to the vertical problem of naturalistic reduction. Perhaps some of them are. My purpose is only to note that insofar as these theories address the horizontal problem, their answers are wanting.

2.Teleosemantics

There are innumerable ways of assigning contents to mental states. Whether we ought adopt one or the other must be decided by whether one assignation facilitates explanation better than another. Teleological accounts generally posit that content is constituted by the biological functions that certain entities perform.

There are three main variations on this theme. Consumer semanticists, exemplified by Ruth Millikan, hold that states' content is determined by how they facilitate the functions other systems in an organism have been selected to perform. Tyler Burge has recently been advocating what I shall call an Aristotelian semantics, such that the contents of states are fixed by the functions they in fact play in an organism's current environment-- not necessarily the functions they had in the past been selected for. Finally, informational teleosemantics, defended by Karen Neander, holds that contents are fixed by functions the states themselves have to extract information from the environment-- not by functions played by systems that make use of the states.

Underlying these different reductive accounts are different conceptions of just what explanatory role intentional contents are needed to play. The following will explicate just what these different conceptions are and argue that none of them provide an account of intentional states that adequately captures a role for intentional content that is required for scientific explanation.

2.1 Millikan's Teleological Consumer Semantics

Millikan (1984) argues that the content of intentional states is fixed by functions certain parts of the organism were evolutionarily selected to fulfill. The idea is that it's because a mechanism performed a certain function that it has survived the selection process. For example, it's because hearts function to pump blood (and not because they function to make pumping sounds) that they have been selected. The

functions that determined a mechanism's selection are what Millikan calls "Proper functions."

A bit more neology: "Normal" behavior is behavior in which Proper functions are carried out. "Normal" conditions are conditions that must obtain for Proper functions to be carried out. Thus, it is biologically Normal that blood have a particular viscosity so that a heart can behave Normally-- that is, fulfill its Proper function to pump blood.

Behaviors are not "normal" in the sense that they are the functions that the mechanisms *commonly* or on average perform. Indeed, Millikan is careful to note that Normal behaviors and Proper functions are sometimes rarely implemented.

The content of an intentional state is constituted by the state the world would have to be in so that that the mechanism making use of the intentional state could perform its Proper function. This is why Millikan's teleosemantics is a consumer semantics: content is determined by the Proper functions of the mechanisms that *consume*, or make use of, intentional states.

So, for example, consider the states of magnetosomes in magnetotaxic bacteria. These bacteria have tiny magnets-- magnetosomes-- that orient the bacteria toward magnetic north. Since these particular bacteria are anaerobic, it's a happy circumstance for them that travelling toward magnetic north in their normal environment pulls them under water away from toxic oxygen rich surface water. So, states of the bacteria's magnetosomes are consumed by the bacteria's locomotor system. Under usual circumstances, the states of the magnetosome cause the motor

system to propel the bacteria to anaerobic environments.

The Proper function of the motor system is to move bacteria toward anaerobic environments. This is because the reason current bacteria have the motor systems they have is that ancestral bacteria who had motor systems that took them toward oxygen rich environments died off: the motor system has been selected to fill the function of moving the bacteria to anaerobic environments. Insofar as the motor system relies on the states output by the magnetic detection system to fulfill this Proper function, the states of the magnetic detection system are about anaerobic environments.

2.1.2 Vindicating Intuitions

Millikan views it as a virtue of her theory that it assigns content to states of the bacteria in line with our intuitions about what content we ought to assign to those states. She approvingly notices that "intuition tells us that what the pull of the magnetosome represents is the whereabouts of oxygen free water" (1989, p. 290).

Most informational theories of content will not accommodate this intuition. For, all we need do is introduce a bar magnet in toxic oxygen rich waters, and the bacteria will flock there, poisoning themselves. Thus states of the bacteria are not causally responding to the oxygenation of water, but rather magnetic field orientation. The informational theorist will conclude that the magnetosome states are about magnetic field orientation rather than anaerobic properties of the environment, because they carry more information about the former. Thus, when the bacteria suicidally move toward the bar magnetic in oxygen rich water, it is not because they have a false representation. They are correctly representing that magnetic north is

strongest at that location. Millikan concludes that "[n]one of this makes any sense on a causal or informational approach" (1989, p.290).

It's not immediately obvious why an information theory of content could not make sense of these data. The bacteria's magnetosome states counterfactually causally co-vary with magnetic field orientation, but not oxygenation. Thus, the causal theorist will likely conclude that if the bacteria represent anything, it's magnetic field orientation, and not oxygen levels. This account seems to make perfect sense of the data. It does, of course, fly in the face of Millikan's intuitions of what the content of the states ought to be. So, it seems that for a theory of content to "make sense" for Millikan, it must comport with our intuitions about what the content of states ought to be.

On this measure, her adaptationist consumer semantics may well make sense. To determine the content of an intentional state, a causal role semanticist asks which correspondence between the state and the environment is causally robust across counterfactual conditions. The consumer teleosemanticist, on the other hand, asks which correspondence between intentional states and the environment would disrupt the functioning of downstream processing were it to be broken.

So, by Millikan's lights, states of the magnetosomes represent oxygenation because if the correlation between states of the magnetosome and anaerobic environments were to be broken, the function of systems downstream in processing would be disrupted. The bacteria's other systems can best carry out their proper functions if the bacteria are in anaerobic environments. Thus, presumably the

magnetosome was selected because more often than not it got the bacteria into anaerobic environments. So, the Millikan's teleo-consumer semantics assigns representations of oxygen levels to the bacteria, in consort with our putative intuitions that the bacteria represent such things.

So, one point in favor of Millikan's teleo-consumer semantics is that it accords with Millikan's intuitions about what the content of mental states ought to be. There are several reasons for thinking this is not an adequate criterion for judging the plausibility of vertical theories of content.

First, while Millikan's theory may accommodate this particular intuition, it has other rather unintuitive consequences. The most well known is the swamp man (Davidson, 1987): a physical duplicate of me spontaneously materializes from a swamp. Because it does not have my evolutionary history, none of its psychological states have the content that mine do, even though they are physically and *a fortiori* functionally identical to mine. For, Millikan's theory assigns contents relative to evolutionary history, not physical or functional make up. That swamp man's psychological states don't share my content strikes many as counter-intuitive.

Unintuitive consequences are of course not unique to Millikan's theory. Most theories of content reach some unintuitive consequence or another. It thus seems misguided to use intuitive plausibility as a criterion on the adequacy of vertical theories of content: each one will fail the test to some extent! Of course, intuitions may constitute excellent explananda for a theory-- think of the success generative linguistics has had systematically explaining people's linguistic intuitions. In this

case, though, it takes as explanandum the fact that people *have* certain patterns of intuitions. It does not suppose that linguistic theory must be constrained so as to make those intuitions *true*.

A deeper worry is that, by using intuitive plausibility as a guide to evaluating vertical theories of content, Millikan has taken an untenable position on the horizontal theory. As Pietroski (1992) notes, in attempting to give an assignation of correctness conditions to mental states in a way that accords with our intuitions, Millikan seems to have failed to assign them in such a way that they can underwrite intentional explanations.

If the only thing that content attributions are doing is vindicating the truth of our intuitions about how to *talk* about content, then they're not doing any interesting explanatory psychological work. Therefore, there's no good reason to posit their existence. Doing so is tantamount to characterizing God as the mass-energy of the universe to vindicate our intuitions that He exists and takes many forms. Positing his existence does not allow us to make any generalizations we'd otherwise be unable to make. Vindication is always nice, but I take it that what we want from science is a description of the world independent of our interests-- not one that merely confirms our convictions.

2.1.3 Normalizing Explanations

Millikan (1995) does claim that her theory does more than just vindicate our intuitions. It provides what she calls, following Philip Pettit (1986), "normalizing explanations." Such explanations describe phenomena in terms of norms, where:

> To explain a phenomenon by subsuming it under norms is to exhibit it as an instance of conformity to or departure from proper operation of some teleological system (pp. 187-188)

Such explanations explain such phenomena as why a bear is asleep, why the washing machine door is closed, why the motor stalled, and why the car went through a red light (p. 188).

Now, I have been operating under the Quinean assumption that there is an interesting distinction to be made between ways of cutting up the world that prove necessary to do our best science and ways of cutting up the world that do not. I take our "best science," to simply be the best description of the world independent of our interests in it. That is, a description of the world, which may ascribe interests to us and describe what they are, but which is otherwise insensitive to them. Thus, to ask whether intentional states have an explanatory role in our best science is to ask whether they are necessary to make generalizations about how the world is independent of the interests of those doing the generalization.

One worry I have for Millikan's account, then, is that, while intentional states may well feature in "normalizing explanations," such normalizing explanations only explain how the world is relative to our pre-theoretical folk psychological conceptions. If this is the case, intentional states do not play an explanatory role in our best science because they do not help explain how the world is independent of our interests in it.

For, on the face of it, any norms we might use to describe a given phenomenon would seem to derive from our interest in that phenomenon. That is, the norms seem hypothetical rather than categorical. And they are hypothetical relative to our interest

in the phenomenon at hand. The washing machine staying locked after the spin cycle is over violates a *norm* only insofar as we want it to open up in order to get our clothes out! Similarly, the car's motor stalling violates a norm insofar as we would rather it keep running and get us to our destination. A car that runs a red light violates a norm only insofar as we would rather cars (other than our own, at least) not run red lights.

Now, of course, it is an interesting fact about our environment that cars, as a matter of fact, do largely tend to stop at red lights. So, why a particular car bucks this trend may well serve as an interesting explanandum independently of whether it matters to our interests *qua* drivers as to whether cars run red lights. In that case, the phenomenon such that *most* cars do stop at red lights is not to be explained by referencing anything that's special about *cars*-- let alone anything having to do with teleofunctional norms that govern them. Rather, we should appeal to the people driving the cars and their desires to stop cars at red lights. Cars stop at red lights because the people driving them want them to.

But, Millikan's suggestion is explicitly that we *not* take such deviations from *de facto* regularities as an explanandum. Rather, the normalizing explanation is to characterize the explanandum in terms of the car deviating from some teleofunctional *norm* that abstracts away from usual, common or even counterfactually characterized behavior.

The worry is that given Millikan's examples of normalizing explanations, any norms we ascribe will merely record our interests in a particular phenomenon. Thus, any explanations couched in terms of those norms will be explaining phenomena only

relative to our interests in it. Instead of explaining why a car stops at a red light by referencing the causal influence of the desire its driver has to stop it, we are to explain it in terms of the violation of the norm that cars *ought* to stop at red lights. But, it would seem, to say that cars ought to stop at red lights is only to note that we generally have an interest in them doing so. To say that the washing machine door ought to unlock after the spin cycle is merely to note that we'd rather it would. So, if the normalizing explanations that invoke intentional states are of a similar kind, then Millikan will have failed to carve out a notion of intentional state that does real scientifically explanatory work.

2.1.4 Explicating Proper Functions

Of course, there may well be instances in which characterizing phenomena in terms of conformity with some teleofunctional process *does* indeed allow us to make generalizations about how the world is independent of our interests. Presumably, Millikan believes that many biological explanations have this characteristic, as do intentional explanations. She writes:

> Now to study how an entity as falling within a biological category 'works' involves (1) understanding what functions are proper to it and to its constitutive systems, parts and states and (2) understanding how these functions are Normally performed (1986, p. 55)

So, extrapolating, to explain the operation of the heart is to explain that it's because it circulated blood that it was evolutionarily selected, and that, historically, it did so via a certain configuration of atria, ventricles, and valves. Atria, ventricles, and valves are thus natural kinds because they figure in a biological explanation of the

Normal Proper functioning of hearts.

By Millikan's own lights, such an explanation does not facilitate any predictions: after all, just because this is how hearts Normally function entails nothing about how they actually function in most cases. For recall, Normal function is just the function evinced under the conditions in which it was selected. Neither does it, according to Millikan, subsume a phenomena under laws. Again, to note that, historically, some hearts have functioned a certain way in particular environmental contexts says nothing about whether the operation of any particular heart, or the operations of most hearts, obey any law governed process.

So, how does the above explain how my particular heart works? *Prima facie*, it seems only to note the fact that my heart is the descendent of previous hearts, which themselves worked in such a way so as to circulate blood. But, it says nothing about whether my heart works in this way or in some much different fashion. I'll have to speculate about the sense in which the above account is an explanation because Millikan leaves it rather unclear precisely how such explanations do their explaining.

One thing the above account might do is provide criteria that licenses categorizing what I've been calling my "heart" *as* a heart. Millikan certainly thinks that this is one use for Proper functions:

> "the categories "heart," "kidney" and "eye," as naming parts of both crayfish and people, are carved out by reference to their most obvious proper functions." (64)

It may well be that what makes my heart, a lizard's heart, and a fish's heart all the same thing-- namely, hearts-- is that they are all structures that have a common

evolutionary etiology such that they were selected to fulfill the same Proper function. It does not follow from this fact alone that the Proper function of hearts can be invoked to make any interesting generalization about how *particular* hearts work.

Perhaps Millikan is correct that an appropriate analysis of the concept, 'heart', makes use of Proper function as a necessary condition. But, we should take care to distinguish this *constitutive* use of Proper function from the *explanatory* use we are currently searching for. The natural language question, "why is John a bachelor?" is multiply ambiguous: one thing we may be asking is what properties of John qualify him to fall under the concept, 'bachelor' (he's a marriageable single man, of course!). Another thing we may be asking is what is the process by which John came to possess those properties, whatever they may be (he dropped out of seminary and never married)[5].

So too, we should be careful to distinguish two explanations we might be after when we set out to explain "how my heart works." We might, I suppose, be asking why my heart qualifies as a heart in the first place. If that's the case, then Millikan's account of the constitutive conditions on being a heart counts as an explanation

[5] This is essentially the objection Fodor (1975, pp. 6-7) levels against Ryle's conception of psychology. Fodor insists that psychological science, like breakfast science, give explanations in terms of causal etiology. Just as we explain why Wheaties is the breakfast of Champions by describing the campion-making effects of its vitamins, we explain psychological capacities by describing the causal properties of mental states. We could instead answer that what makes Wheaties the breakfast of champions is the fact that many people we consider to be Champions eat it. Such an explanation, however, does not justify the claim that these Wheaties eaters are in fact champions. Likewise, we could explain how it is that organisms have intentional states by pointing out that they have many states that we consider to be intentional. But, such an explanation is not sufficient to justify our claim that those states are in fact intentional.

(correct or not). But, I take it Millikan would agree with me that biologists are after more than just conceptual analyses of concepts like heart and kidneys. They also want to explain how my heart works in some other sense. They want to explain what it's doing and how. If that's what we're after, it's hard to see how a normalizing explanation is going to help.

Of course, it could be that the biologist is not concerned with how hearts are currently working, but with how they have worked at particular points in the past. The explanations they are searching for just are accounts of which functions caused hearts to be evolutionarily selected, and which did not. Delineating the heart's Proper functions and Normal conditions for those functions would certainly constitute an answer to this question. So, it seems the sorts of explanations Millikan takes biologists to be looking for are simply accounts of which functions are Proper and which are not. As such, these explanations are not so much explanations of why something has happened, as explications of particular historical facts.

No doubt, these are quite interesting historical facts that we'd want our science to capture. The question we must raise, though, is whether assigning content to mental states is necessary to understand which are the Proper functions of any system.

Crucially, for our purposes, it seems that intentional content is superfluous to any account of Proper function. Consider the case that Pietroski (1992) brings against Millikan. Back in the mists of time, a creature, a kimu, develops a mental state that fires at the sight of red things and causes him to move toward them. Consequently, he

moves toward the red flowers growing on the top of a hill. Doing so moves himself away from the predation the snorfs perpetrate on his fellows in the valley below. He thus survives and propagates his red detecting genes such that now all kimu detect and move toward red things.

By Millikan's lights, the Proper function of the kimu's red detector is to get the kimu away from predating snorfs. Thus, the content of the mental state it produces is something like 'snorf free zone'. Pietroski's observation is that whatever content a cognitive scientist would assign to this state, it certainly would not be that. Ergo, Millikan's account of content is not consonant with current practice in cognitive science and therefore flawed.

I press the deeper point that assigning content using Millikan's method cannot help further *any* explanation-- be it one a typical cognitive scientist might offer or the sort of explanation Millikan says biologists are after. Suppose that a biologist wants to explicate which of the kimu's operations are Proper functions and which are not. She'll want to note that the Proper function of the kimu's detector is to get the kimu away from snorfs under Normal conditions in which there are red flowers atop hills and snorfs only hunt in the valleys.

Now, we could of course go ahead and say that the content of the kimu's red detector state is 'snorf free zone.' But, this certainly doesn't tell us anything more about what the Proper functions of the kimu are. It merely records information that we have antecedently arrived at. So, the attribution of content to the state is superfluous.

You might object that any vertical reduction of content is going to result in contentful states merely recording antecedently available information. For example, if content is constituted by asymmetric dependency relations á la Fodor, then saying that a mental state is about trees just is to record the information that there is an asymmetric dependency between that state and trees. The whole point, after all, of a vertical reduction is to demonstrate how intentional content can be redescribed in non-intentional terms.

The difference is that Fodor assumes that the postulation of contentful states will buy him generalizations and predictions about how the mind works that can't be captured merely by listing out asymmetric dependencies. Millikan explicitly denies that psychology is in the business of making generalizations and predictions over and above recording which functions are Proper and which conditions are Normal:

> The happenings that either folk psychology or a developed scientific psychology could be expected to explain are strictly limited to those that occur in accordance with proper functions of the body's systems (or that occur as common aberrations of these functions-- abnormal psychology).
> ...
> Hence it is not likely to be the business either of folk psychology or of scientific psychology to attempt to explain why I move my right hand 20 inches northwest or why I hand you a picture of Washington. (1986, pp. 60-61)

So, positing that mental states have contents corresponding with the Normal conditions under which consumers of the states perform their proper functions does not help buy her any additional generalizations. It merely provides a convenient *nomenclature* for recording generalizations that we have antecedently made without

the use of intentional properties.

2.1.5 Explaining Malfunction

Now, perhaps I'm wrong and Millikan does have an account of generalizations intentional contents can buy us over and above generalizations about the Proper functions they get reduced to. It may be that biopsychology admits of explanations other than simple explications of Proper functions.

Indeed, sometimes Millikan writes as though some of the chief explananda of biopsychology are *malfunctions*. She writes that to explain a phenomenon is to "exhibit it as an instance of conformity or departure from proper operation of some teleological system" (1993, pp. 187-188). Thus, to explain how the heart works is not simply to record what the Proper function of the heart is, but also to give an account of why any particular heart is failing to fulfill its Proper function. So, perhaps the blood is too viscous or the valves are stenotic relative to their Normal conditions, and that's why the heart is not functioning Properly.

Again, however, it's hard to see how attributing content to mental states facilitates explanations of this type. Suppose we want to explain why the Kimu's red detector is not functioning Properly. Perhaps there's a situation in which a red weed begins growing in the valleys. The kimu flock to the valleys, where they are decimated by snorfs. The red detector is no longer playing its proper function of getting kimu away from snorfs. In explaining why it's malfunctioning, we might *say* something like: "the kimu falsely believe that there are no snorfs in the valley."

Now, by Millikan's lights, it is true that the kimu have a mental state about the

valley being snorf free. Unfortunately, the valley is crawling with snorts. So, it does make sense to say that the kimu falsely believe the valley is snorf free. But the question is whether attribution of this false belief helps explain anything that couldn't be explained otherwise.

In this case, there is an explanation that explains perfectly well the malfunctioning of the red detector without attributing false beliefs to the kimu. The red detector causes the kimu motor systems to move them toward red things. There are red things in the valley. Therefore, the kimu move there. Unfortunately, there are also snorfs in the valley. So, the kimu sensory motor system fails to perform its Proper functions of getting the kimu to snorf free zones. Attributing intentional states to the kimu just is not necessary to explain their deviation from their proper functioning.

So, we're left back where we started. If intentional attributions are supposed to explain deviations from proper functioning, then at most all they do is regiment our way of using intentional language to *describe* malfunctions. They do not buy us any explanations we'd otherwise be unable to get. Moreover, Proper functions seem like the only teleological processes around that are not characterized in terms of our particular interests in the world. So, if intentional states were going to explain anything about how the world is independent of our interests, it would be Proper functions. But, while intentional idiom may provide a convenient way of *talking* about Proper functions, attributing intentional properties just doesn't buy us any explanatory power.

Perhaps Millikan does have a correct account of content constitution: how it is that states come to possess intentional content. I leave aside whether she has the correct answer to this vertical question. What the foregoing does make clear, however, is that the explanatory work she takes such content to do is redundant. Her explanations work whether or not we assign intentional content to states of the creatures whose functions we are explaining. Thus, Millikan's answer to the *horizontal* problem of what explanatory work intentional content does within science is unsatisfactory.

2.2 Neander: Teleo-Informational Production Semantics

Neander disagrees both with Millikan's answer to the Vertical and the Horizontal questions. Her alternative Vertical theory is that some systems function to carry information about the world, and it is these functions that determine the content of mental states. Her approach is therefore often characterized as a *production* semantics. Whereas Millikan believes that the functions executed by systems that *consume* intentional states fix the contents of those states, Neander argues that intentional states are *produced* so as to function to carry information about things in the world-- and it is these distal stimuli that fix the content of the states.

While this way of fixing content is supposed to be able to do more work within cognitive science, Neander's strategy ends up falling prey to the same sorts of objections as Millikan's. In particular, the intentional ascriptions her theory prescribes merely record generalizations made independently of intentional properties. They,

themselves, do no explanatory work.

One reason Neander thinks a production teleosemantics is a better fit with cognitive science is that she takes it to obey what she calls the "coherence constraint":

> if the contents of mental representations are to play a role in explaining cognitive capacities, they must cohere with the relevant information processing (2006, 180).

Insofar as cognitive science couches its explanations in terms of information processing, intentional contents that play a role in such explanations must "cohere" with them.

Thus, commitment to the coherence constraint is not so much a substantive difference, but rather a methodological difference between Neander and Millikan. It's meant to assert that it's at least *prima facie* misguided to abstract away from the information processing going on in cognition and look only at the adaptive functions being performed by an organism when we go about fixing its contents, as Millikan does. We can't read "coherence" as a gloss on some substantive way in which intentional content must be fixed by information processing. Rather, it's just an injunction that we look closely at the information processing explanations we give of organisms so that we are sure that the contents we assign to their states at least do not conflict with these explanations.

Neander of course argues the further substantive point that once we conduct such a survey, we'll find that the contents assigned by her production semantics comport better with the cognitive science than those assigned by Millikan's consumer semantics. Neander's arguments are compelling in this regard-- but it's unclear the

contents her story assigns to mental states play any more an explanatory role than do those assigned by Millikan's.

Neander (2006; forthcoming, Ch. 7, pp. 20 ff.) identifies excitation of T5(2) cells in a toad's optic tectum as a representation. These cells tend to fire when there is a stimulus in the cells' distal receptive field. Therefore, T5(2) excitation carries information about the presence of distal stimuli in this receptive field. But, of course, it also carries information about distal stimuli in larger areas encompassing that receptive field. And, in a token instance in which a particular distal stimulus-- a worm, say-- is stimulating the T5(2) cells, they are carrying information about the specific location of the distal stimulus-- a location smaller, but still within, the cells' receptive field. So, Neander asks: does the excitation of T5(2) cells represent a location consisting of their distal receptive field, an area larger than that field, or the smaller portion of that field where a worm is in fact located?

Neander answers that the representation has as part of its content the location of the cells' distal receptive field. Otherwise, she claims, we'll be unable to explain aspects of the toad's behavior. Toads tend to orient themselves toward the distal receptive fields of their T5(2) cells when they receive certain stimuli. Our explanation of this behavior would be inadequate if we claim that the T5(2) cells represented an area larger than their receptive field: it would leave it a mystery as to why the toad tends to orient only toward the receptive field of the cells, and not as frequently to locations outside of those fields. Were the T5(2) cells to represent the precise location of a distal stimulus within their receptive field, it would leave it a mystery why the

toad frequently orients to some other location within the distal receptive field rather than the precise location of the stimulus.

By process of elimination, it must be that T5(2) cell activation represents the distal receptive field of the cells. Assigning content to T5(2) excitation as of a smaller region of the of the distal field, or a larger region encompassing the distal field not only would be explanatorily impotent, but would seem to interfere with explanations of why the toad orients itself as it does.

However, it's not clear to me that assigning representational content as of the distal field of the T5(2) cells helps with any explanations either. Neander seems to think that it may explain why the frog orients itself toward the distal field of T5(2) cells when they are stimulated. I suppose the explanation is supposed to be something like the following.

The explanandum is that *ceteris paribus,* the frog orients toward the distal field of its T5(2) cells when there is something in that field. The explanans is that It is because the frog represents that something is located in the receptive field of the T5(2) cells that it orients itself toward that space. So, T5(2) representations with contents as of the distal field *explain* the frog's orientation behavior.

2.2.1 Motivating Intentional Attribution

There's a worry that this argument relies on a false trichotomy. Just because attributing representation of extra-receptive-field locations or precise locations within the receptive field proves explanatorily fruitless, it does not follow that attributing representations of the distal receptive field will be explanatorily useful. That is, we

want a reason as to why we should attribute *any* intentional content at all to the T5(2) excitation.

To this worry, Neander writes:

> if...brain states have the function of carrying information, then they have intensional states insofar as they are not extensional.“ (2006, p.169).

It’s not immediately clear why informational states must be intensional (with an “s”!). After all, the sort of “natural information” Neander has in mind here, coming from Dretske (1981) often doesn’t individuate information intensionally. In this sense, smoke carries information about whatever has caused it in extension-- under whatever description that may be.

To understand why Neander takes such informational states to be individuated intensionally (with an “s”!), we need to take a further look at her account of toad visual systems. Neander (2006; forthcoming, Ch. 7) recounts that toads orient toward “worm-like” stimuli with elongated, rectangular shape moving parallel to their long axis, but not to such shapes moving perpendicular to that axis, or to other shapes, such as squares.

Since elongated shapes moving parallel to their long axis are usually worms, centipedes, and other foodstuffs in the toad’s usual environment, responding to such shapes functions to allow toads to capture their prey. Presumably, this capacity to orient toward such shapes was selected because of its ability to get nutritious things in the toad’s digestive tract. Orienting the frog toward food is the Proper function of such T5(2) stimulation. Thus, by Millikan’s lights, it seems we should take the toad to

have a representation with the content 'food' or 'nutrient package' or something similar.

Neander however, argues that we should instead take the toad to represent something like 'oblong shape moving parallel to its long axis.' After all, the toad can fail to identify food when relying on this strategy of orienting and snapping at such shapes. Experimenters get the toad to orient toward *non-nutritious* cardboard shapes, and nutritious millipedes would fail to evince the response were they stunned and dragged perpendicular to their long axis. Thus, Neander writes:

> the capacity that is possessed by the toad, as opposed to the capacity that is merely approximated, is what must be explained by a functional analysis of the relevant capacity. What the toad possesses is a capacity to recognize a certain configuration of visible features. What the toad does not possess is a capacity to recognize nutritive potential. (forthcoming, Ch.7, p. 31)

The virtue of Neander's account seems to be that she is able to capture counterfactual generalizations about the toad. Counterfactually, the toad responds differentially to a particular configuration of visual features, but not to properties of being nutritious. The ability to respond systematically to nutritious stimuli is only a "capacity that is merely approximated" in virtue of nutritious things contingently co-varying with certain configurations of visual features.

It must be this ability to track counter-factual generalizations that Neander is pointing to when she argues that states with the function of carrying information are intensional (with an "s"!). Sure, any given instance in which a worm causes T5-2 firings, that firing carries information about that worm. But, in general,

counterfactually, T5-2 firings carry information about a particular combination of visual features, and information about worms only incidentally. Thus, we can specify the information the firings *function* to carry *intensionally* (with an "s"!) as the property of oblong shapes moving parallel to their long axes.

Specifying the information carrying functions of the toad's states in these terms allows Neander to give a description of the toad that coheres more with the *counterfactual* generalizations of cognitive science than our folk psychological intuitions. She writes:

> Price [2001] appeals to the Davidsonian claim that rationalizing behavior is the purpose of intentional ascriptions in folk psychology, and claims that a frog (and by extrapolation a toad) would be irrational if it were to eat a dummy stimulus if it did not think that it was food. Perhaps rationalizing behavior is the point of content ascriptions in folk psychology, but the point of content ascriptions in cognitive neuropsychology is not to rationalize behavior. The point is to explain cognitive capacities. And, whatever the folk might think, the toad does not need to know about food as such, and nor does it know about food as such. As Jerry Fodor puts it, '[t]he mathematics of survival comes out precisely the same either way.' (forthcoming, Ch. 7, p. 32)

Thus, she avoids simply assigning intentional content in a manner that simply comports with our folk psychological intuitions, or the interest-relative norms we might ascribe to the toad, as Millikan's strategy seems to.

2.2.2 Intentional Attribution

Nonetheless, it's unclear what explanatory benefit there is in attributing *intentional* (with a "t"!) content to T5-2 states over and above the intensional (with an

"s"!) characterization of the information that they carry counterfactually. To see why, we should look to Neander's key claim that "normative aboutness is born [by] inheriting its norms from… functions and the aboutness from… natural, factive information" (forthcoming, Ch.4, p. 36).

Here, functions are attributed to organisms as a solution to what Neander calls the "generalization problem": the "need to find useful general descriptions of the operation of kinds or types of living systems" (Ch.3, p.23). For Neander, an ideally functioning organism is

> a system in which all of the traits (parts, entities or structural elements) that were selected for doing something can (i.e., possess the disposition to) do what they were selected for doing (Ch. 3, p.24)

Thus, organisms with a diverse range of physiological variety can be described in terms of their deviation from this ideally functioning exemplar. For example, the function of the T5-2 cells in toads is to co-vary with certain stimuli. But, they may fail to fulfill that functions in an abnormal toad that, for example, has an ablated thalamus. So, functional characterizations of creatures do have a normative component. A toad with an oblated thalamus would be failing to function correctly relative to an ideally functioning toad.

Insofar as we can specify that states of the toad carry information *about* worm-like shapes, intensionally specified (with an "s"!), the information carrying properties of mental states also have *aboutness* in some sense. So, combining these features of normativity and aboutness, we can reconstruct an argument as to why T5-

2 states have intentional content (with a "t"!).

As we've seen, T5-2 cells have the capacity to carry a diverse variety of *natural* information. In different circumstances, they may carry information about worms, about cardboard cutouts, about millipedes, etc. If the intentional content of a T5-2 state is going to inherit its aboutness from the natural information it carries, we need some way of privileging some of this information over others. As we've seen, Neander points out that, in general, the T5-2 states co-vary counterfactually with particular visual features more robustly than any of the other things they contingently carry information about. Thus, we can take the *aboutness* of the information T5-2 states carry to be something like 'worm-like shapes moving parallel to their long axes.'

In an ideally functioning toad, T5-2 states *function* to carry this information. So, if a toad's T5-2 states fail to carry information about those visual properties, they are falling short of the normative strictures on their function.

Putting the aboutness of information and normativity of function together, then, we can say that T5-2 states have intentional content as of 'worm-like shapes moving parallel to their long axes.' Attributing such an intentional content to the toad allows allows us to capture the generalizations that (1) counterfactually, T5-2 states tend to co-vary with particular worm-like visual properties and (2) T5-2 states that fail to covary with such properties are *in error* insofar as they don't comport with their ideal function.

But, it looks like we can do this explanatory work *without* appeal to

intentional content. We can generalize that T5-2 states respond differentially to particular visual stimuli simply by, as we did earlier, noting that fact! We can explain why sometimes this co-variance breaks down by noting that sometimes a particular toad falls short of its functional ideal for various reasons (e.g., an ablated thalamus).

Indeed, if attributing intentional content amounts to attributing a function to carry information across counterfactual circumstances, it's unclear what explanatory role it might have that can't be filled simply by the *conjunction* of functional explanations and natural information explanations. Toads capture prey by having states that *function* to process the retinal *information* that tends to co-vary with the shape their prey. That is, *ceteris paribus*, the states of the frog co-vary with worm-like stimuli impinging on the retina. If saying that toads capture their prey because they *represent* the oblong shapes of their prey is just a way of glossing the previous sentence, it's unclear how the intentional idiom is any more than just that: a manner of speaking.

If the intentional attribution is doing any explanatory work, the error states it attributes to the toad should enable generalizations that we couldn't otherwise get. But, again, it's hard to see what these would be. A toad that has its T5-2 cells stimulated in just the right way, by whatever means, will orient in a certain direction. Whether that stimulation was the result of some distal worm, some worm-like shape projected on the retina, or direct electrical stimulation of the T5-2 cells themselves makes no difference to this generalization. Yet, Neander's theory would have it that the first two T5-2 excitations were "correct" whereas the third was "incorrect"--

because in that case (arguably) the toad was not functioning ideally. It's not obvious what this nomenclature does for us beside being a useful notation for instances in which the toad is functioning ideally and those in which it's not.

Actually, the situation is more complicated. I've until now been glossing over an ambiguity in what Neander means when she says the toad represents "a certain configuration of visible features". This gloss on the content is ambiguous between a distal and a proximal reading. On the distal reading, the "visible features" are features of the distal stimulus itself. Under this reading, T5-2 states carry information that worms, cardboard cutouts, etc. *themselves* have the property of being rectangular and moving parallel to their long axes. Under the proximal reading, "visible features" are properties of the proximal retinal stimuli. Under this reading, T5-2 states carry information about such proximal properties. I've been assuming the proximal reading.

But, Neander (2012, pp.33-35) argues for a "distality principle" such that a given state represents distal worm-like motion rather than proximal worm-like stimulation of the retina. This is because, she argues, T5-2 states are sensitive to such proximal stimuli only *in order to thereby* be sensitive to distal worm-like motion[6]. Therefore, only situations in which T5-2 firings are occasioned by actual distal worm-like motion are "correct" tokens: states caused by direct stimulation of the retina or

---

[6] This prescription, Neander argues, does not require the Millikan reading that the toad represents nutritious foodstuffs as such because it is sensitive to worm-like motion only insofar as to thereby be sensitive to nutritious foodstuffs. This "distality principle" is only to be applied to disambiguate distal from proximal stimuli-- not to disambiguate two potential distal contents.

direct stimulation of T5-2 cells themselves are "incorrect."

But, again, it's unclear how this change in attributing correctness and incorrectness to T5-2 states buys us any explanatory power. We know that worm-like movements projected on the retina will have the same effect as distal worm-like motions presented to the retina. Yet, we call one instance correct and the other incorrect. This asymmetry in intentional idiom doesn't seem to track any asymmetry in the actual operations of the toad. Whether the toad's system tokens T5-2 states correctly or incorrectly doesn't seem to make any difference as to how the toad operates. We can give a description of the toad's behavior in complete abstraction from whether its states are correct or incorrect-- merely applying these as labels according to Neander's prescriptions after the fact.

Now, one thing that's problematic with Neander's account of toad vision is that the state to which she's assigning intentional content has already been type-individuated *independently* of its intentional content-- in neurophysiological terms as the firing of T5-2 cells. Thus, since our description of the toad in this instance seems to be done entirely at the neurophysiological level, we might expect that of course whatever content we assign to the T5-2 cell firings will be inconsequential for the purposes of the explanation at hand. But, that in itself is not reason to reject Neander's theory of content. It may be that her theory of content *constitution* is correct-- even if the content assigned to T5-2 states just happens to be inefficacious in the context of the current explanation.

Suppose, then, that we have some mental state, M, that we *cannot* type-

individuate in neurophysiological terms. Perhaps it's implemented by different neural structures within and across subjects. Does Neander's theory of content constitution help explain the workings of M?

We may once again assume that M has a counterfactually robust, though not perfect, disposition to co-vary with, as Neander puts it, "a certain configuration of visible features." Thus, we may take this as the content Neander would have us attribute to M. Again assume that the distality principle holds, so M represents distal instances of these features, even though it could be tokened by the projection of such features on the retina. Moreover, M was selected for its ability to covary with such distal stimuli.

So, any instance in which M activates in the absence of such stimuli will be an incorrect tokening of M. So, instances in which M fires in the presence of a square cardboard cutout will be incorrect, but instances in which it fires in the presence of a worm-like cardboard cutout will be correct. Nonetheless, we know by stipulation that M will usually fire in the presence of worm-like-motion-- but sometimes will not.

Whether those firings are correct or incorrect has no bearing on explaining any other aspects of the toad's behavior. M firings will evince the same behavior in the toad whether they are correct or incorrect. We can even type-individuate M states as such without appeal to their intentional content. We can just point to the ability of M states to robustly counterfactually co-vary with worm-like stimuli (whether proximal or distal). A toad has M states iff it has such states.

All this is just to note, *contra* Neander, that characterizing intentional content

in terms of the information that a state functions to counterfactually carry does not in seem like the most explanatorily useful characterization. Just as they do for Millikan, the ascription of intentional states in Neander's semantics seems merely to note, *post hoc*, generalizations that have been made without recourse to intentional properties.

As such, this notion of content does not seem suited to explaining many phenomena that on their face seem amenable to an intentional explanation. For example, Georges Rey (2012) has noted that if explanations of the early visual system employ representations of geometric objects, or explanations of language processing employ representations of words and phonemes, then these states *never* carry natural information about the things they represent-- to do so is impossible, as cylinders and phonemes just don't exist! It's in just these cases in which we can't appeal to the natural information that mental states carry that it seems like attributing intentional content would be helpful. Neander's account can't explain how we can have representations of such inexistent objects.[7]

So, while Millikan's consumer teleosemantics focuses on the functions for which certain mental states were selected, and Neander's production teleosemantics focuses on the counterfactual causal antecedents of mental states, both give an account of intentional states that robs them of explanatory power. Both accounts

---

[7] If Rey's examples are too outré for your taste consider representations of souls and God. Only highly abnormal, educated humans represent these things as extant iff they actually are. Normal humans represent these things as being all over the place-- despite their absence.

ascribe intentional content to states in such a way that the ascription merely takes note of generalizations that can be made independently of any actual intentional properties.

2.3 Burge: Neo-Aristotelian Teleo-Ethological Semantics

Tyler Burge enunciates a variation of the teleosemantic view in his latest (2010) tome. Unlike Millikan and Neander, Burge is explicitly concerned with ensuring that the intentional properties of representations play an ineliminable role in psychological explanation. Because he is so sensitive to this issue, it's remarkable that his account of intentional states seems on the face of it not to allow for such an explanatory role.

Burge also is unique in not giving a reductive account of intentional content. Moreover, he thinks that the vertical problem is not a problem at all, and consequently disparages attempts at solving it. But, while he does not have a fully fledged theory laying out the constitutive conditions of having a state with a particular content, he does lay down some curious strictures on how the contents of perceptual states are to be individuated.

Burge writes: "Psychological explanation is framed to fit with explanation in zoology and ethology" (322). Insofar as his claim is that psychological explanation aims to be consistent with other levels of scientific explanation, the claim is uncontroversial. But, Burge means something much stronger. His claim is that organisms can only represent in perception properties that feature in zoological and ethological explanations. Indeed, we can be certain frogs do not represent undetached

fly parts, bee-bees, holograms, or shapes moving on a screen as such because these things:

> are clearly outliers in the environments in which the perceptual systems evolved. They are not among the typical causal agents interacting with individuals' perceptual systems in fulfillment of individual biological functions. (p. 465)

Thus, psychology must mesh with biology by only attributing representations of entities that play a role in fulfilling biological functions.[8] Why Burge supposes this, and indeed, just what sort of "biological functions" count for these purposes is unclear.

His motivation seems to be an attempt to dissolve the "gavagai" problem raised by Quine and the disjunction problem raised by Fodor and other theorists of intentionality that Burge denigrates as working in the "deflationary tradition."[9] The disjunction problem is the seeming difficulty of discerning the appropriate correctness conditions from amongst a disjunction of possible conditions. Thus, given that frogs snap at both flies and bee-bees, we might wonder whether the frog incorrectly tokens "fly" representations in the presence of bee-bees, correctly tokens "fly or bee-bee" representations in the presence of both, or whether it has representations with some

---

[8]. Burge does not deny that humans, or frogs, for that matter, can represent things at a conceptual level that do not help them fulfill biological functions. His claim is only that such representations never feature in early perceptual states. If they do pop up, it is in in downstream, global, conceptual processes. See p. 101.

[9]. It's odd that Burge applies this deflationist label to Fodor, who has, far from deflating problems of intentionality, taken them more seriously than most other philosophers.

other correctness conditions.

Burge wants to circumvent the problem by claiming that the frog represents flies because only they are "among the typical causal agents interacting with individuals' perceptual systems in fulfillment of individual biological functions" (465). In the normal environments in which frogs evolved, flies serve as prey and provide nutrients for frogs, whereas bee-bees do not. Therefore, it's appropriate to ascribe "fly" representations, but not "fly-or-bee-bee" representations to the frogs. Therefore, we should explain why frogs snap at bee-bees by claiming that they are incorrectly representing the bee-bees as flies.

But, it just doesn't seem to be the case that states that have correctness conditions that are true of flies but not bee-bees allows us to make any generalizations or predictions we would not otherwise be able to make. The frog in fact snaps at both flies and bee-bees. Counterfactually, it would seem, the frog will snap whenever it receives proximal stimuli of the kind provided by flies and bee-bees. To call some of these snappings correct and others incorrect does not allow us to capture any generalizations beyond these.

To be sure, any explanation of an organism's cognition should explain how it is that it behaves as it does in its normal environment. It's not *prima facie* obvious, however, that doing so requires that organisms only represent things that tend to crop up in their normal environments.

Moreover, Burge seems worried that fixing content via counterfactual conditionals á la Fodor will make it impossible to attribute any content at all to

perceptual states:

> The range of objects of perception, and the environmental grounds for explaining constitutive conditions for a state's having the perceptual content it has, are constrained by factors beyond the animal's discriminative capacity (which leaves open too wide a range to determine what an animal perceives and how). The range is also constrained by the needs and activities of whole individuals—eating, predating, mating, navigating, fleeing, parenting, nesting, and so on. The science of perception explicitly leans on such a constraint. (319-320)

It's unclear again why we should suppose perceptual science is committed to such a constraint, or indeed why relying on general counterfactual discriminative capacities won't fill the bill.

3. Cummins, Gallistel, & Ramsey: Functional Homomorphisms

A second family of answers to the vertical question has it that the content of intentional states are constituted by functional homomorphisms between those states and objects in the environment. Proponents of this view have done an admirable job pointing out the ways in which structural correspondence between mental and worldly states can explain how creatures get around in the world. However, they fail to establish that *intentional* properties add any predictive or generative power to our explanations over and above what mere correspondence itself provides.

Cummins (1989) begins with the assumption that "mental representation *explains* how systems manage to get into states that covary with states of the world" (87). This assumption immediately faces two challenges. First, on the face of it, it certainly isn't clear that this is the explanandum theories of intentionality are aimed

at. If people like Rey (2012) and Mendelovici (2013) are correct, there may be many phenomena representations explain even though the representations are not about anything that actually exists! If the objects of representations don't exist in the first place, it's hard to see how they would be useful explaining how we come to covary with non-existent states of the world! But, perhaps theorists of intentional inexistents are just wrong: perhaps representational states without existent objects are not required by our best psychology. In any case, Cummins is burdened to argue that they are not.

Secondly, Given that there are on the face of it lots of ways for two systems to co-vary with one another, Cummins also owes us an account of the particular sort of covariance that requires the positing of intentional states for its explanation. As we've noted, the mercury in a thermometer comes to covary with the ambient temperature--but this sort of covariation doesn't seem to require explanation in terms of intentional states.

In any event, it's unclear that Cummins ever actually sets about defending the above claim that intentional states explain the covariation of mental states with states of the world. Most of his work goes into showing that thinking of the intentional *talk* used in cognitive science as talk about functional homomorphisms can account for how that talk is deployed. He does not demonstrate how thinking about functional homomorphisms as instantiating intentional properties can add to the explanatory power of cognitive science.

Cummins basic idea is that we have a representation whenever the following

conditions obtain. There is a mental system of operations and an interpretation, *I*, that maps each component of that system onto some object in the world. There's further a function, *f,* that ranges over these worldly objects. If there is a function, *g*, which the mental system satisfies, and which is isomorphic to the function, *f,* then the components of *g* represent the components of *f* to which the interpretation maps them. So, for example, the buttons and display states of a calculator represent numbers because there is a function, *g*, that maps button presses onto display states, and that function is isomorphic to the addition function that obtains between numbers.

Gallistel (1990) picks up largely the same notion of representation in his explanation of animal cognition. An animal represents a domain when there is an isomorphic mapping between entities and relations in the animal's brain and entities and relations in its environment. So, for example, if the firing rate of a neuron covaries with an animal's velocity along the north-south axis, and that of another neuron covaries with velocity along the east-west axis, the neurons represent the animal's velocity in cardinal space (31).

Thus, the Cummins/Gallistel view is a version of inferential role semantics. It's only when the relations amongst mental states isomorphically mirror the relations amongst things in the world that those states represent anything. However, it's unclear how ascribing intentional properties to such states increases the explanatory power of our accounts of cognition.

Cummins writes that interpreting a computational system so as to ascribe intentional contents to it "from an explanatory point of view...provides a link between

mere state crunching (button-pressing-to-display-transitions) and *addition*" (94). He gives the example of a mechanical artefact discovered by an archaeologist. Discovering that states of the artefact can be isomorphically mapped onto the function for addition explains what the machine does: it's an adding machine (95).

Of course, in this example, the discovery of an interpretation under which states of the machine map isomorphically onto elements of the addition function does not explain how the machine itself works. Presumably, in order to discover the isomorphism in the first place, the archaeologist would need a complete description of the machine's operations independent of the interpretation mapping its operations onto the addition function. What the interpretation allows the archaeologist to do is explain why a person *designed* the machine to work as it does. The existence of the isomorphism is a clue that the designer of the machine *intended* that the machine be used as an adding machine.

So, there's a crucial disanalogy between the explanatory role of intentional states as ascribed to the archaeological artefact and the intentional states ascribed to psychological systems. The archaeologist ascribes intentional states to the artefact by way of explaining the behavior of the machine's creator. The intentional properties do not add any generalizations or predictions about the behavior of the machine itself.

When psychologists assign intentional states to mental processes, if they are to do any explanatory work at all, they are to make just these generalizations and predictions about the behavior of the mental processes themselves. Psychologists do not ascribe intentional properties to mental states to explain the behavior of any

putative creator! Cummins' archaeology example does not illuminate how intentional states *qua* states of isomorphic functions are to add to the explanatory power of cognitive psychology.

If there is no account to be had of how intentional states *qua* states of isomorphic functions are to increase the explanatory power of psychology, then Cummins has deflated the horizontal question: intentional states do no explanatory work in psychology. Sometimes Cummins himself seems to allow as much: "Representation, in this context, is simply a convenient way of talking about an aspect of successful simulation" (95). That is, attribution of intentional states is merely nominal. What's doing the real explanatory work is the correspondence between structures of the mental process and structures in the world such that one can be said to "simulate" the other.

It's merely a way of noting that a correspondence obtains between states of the world and a computational system. It does not explain *how* the computational system comes to covary with the system to which it corresponds. It's merely a convenient way of thinking about the correspondence. There may well be an interpretation under which states of magnetotaxic bacteria correspond with states of water oxygenation. By Cummins' lights, this is sufficient to constitute states of the bacteria as representations of the oxygenation of the water. But it's unclear how ascribing intentional properties to them in this way does any explanatory work. It does not allow any generalizations or predictions that the mere correspondence itself allows.

3.1 Ramsey's 'Job Description'

William M. Ramsey (2007) advances two additional arguments for thinking that isomorphic accounts of representation play an explanatory role in cognitive science. First, they make sense of how complex computational processes decompose into more simple computational processes. Secondly, they explain how creatures with such isomorphic states successfully navigate their environments.

Unlike the figures examined above, Ramsey is explicitly committed not to the vertical project of reducing intentionality to non-intentional terms, but to the horizontal project of describing the explanatory role representational states play in cognitive science. He calls this project the "job description challenge," viz.:

> some account of just how the structure's possession of intentional content is (in some way) relevant to what it does in the cognitive system (p. 27)

Nonetheless, his considerations of the role of intentional content fail to meet his own challenge.

3.1.1 Explaining Compositionality

The first way in which homomorphic representations are supposed to meet this challenge is to account for the decompositional structure of computational processes. Consider a system that takes pairs of inputs and outputs single states such that these states homomorphically map onto the process of multiplication. That is, there is a homomorphic mapping from states to numbers such that the output state always maps to the number corresponding to the product of the numbers the two input states map to.

Ramsey notes that such a machine would have to employ some particular algorithm in order to instantiate this function. For example, suppose it receives the states mapped to 7 and 3 as input. It might, for example, then instantiate the state mapped to 7 twice, output a state mapped to 14, and then instantiate the 7 state a third time before outputting a state mapped to 21. We can best make sense of how the output of the machine corresponds to the product of the input by understanding that intermediate states of the machine correspond to the process of repeatedly adding the inputs.

Ramsey claims that unless we take states of the above system as being *about* numbers, "we couldn't understand how the system succeeds by breaking a large computational operation down into related sub-operations" (pp.75-76). Attributing intentional content to computational states in some way makes sense of how certain computations can be interpreted as carrying out processes (like multiplication) in virtue of carrying out other computations (like adding). Thus,

> It is both explanatorily useful and informative to *see* a sub-system of a multiplier as an adder. It is not so useful or informative to see it as a mere syntactic shape shifter (p. 76, my emphasis)

This formulation of Ramsey's conclusion is ambiguous. On the one hand, it may be that without intentional content, we would not be able to explain how computational processes that are homomorphic to the multiplication function over integers in virtue of being homomorphic over the addition function. Alternatively, the claim could just be that it's pragmatically useful for scientists to "see" the computational process *as if* it has states with intentional content.

Ramsey recognizes this ambivalence in the formulation of his conclusion. He puts the dilemma this way:

> It seems these notions of representation serve as representations not for the system, but for the psychologists attempting to understand the system in a certain way (p.98).

It may be that psychologists stand in the same relation to psychological systems as we stand toward, say, compasses. States of the compass have intentional content insofar as we *interpret* them as such-- but the states don't have any *original* content independent of our activity. If that's the case, Ramsey doesn't have a solution to our problem-- namely, what is the role of original content in cognitive science?

Ramsey maintains, nonetheless, that the intentional content he attributes in virtue of these homomorphisms *is* original in this sense. He points out that we could not explain how intentional accounts of cognition are intentional unless they make use of intentional states:

> There is no discernible way that something could serve as, say, an adder, without it also being the case that it converts representations of numbers into representations of sums. Without such a conversion of representations, it simply wouldn't be doing addition (p. 100).

Ramsey's claim is indubitably correct, such as it is. But, it begs the question. Ramsey seems to be arguing that the explanandum attributing intentional content to computational states is supposed to explain is the fact that computational processes operate over intentional states. But, precisely what is under question is whether computational processes must be characterized in intentional terms. We saw in Chapter 1 that they need not be.

Ramsey makes similar claims multiple times. For example:

> computational processes treat input and output symbolic structures a certain way, and that treatment amounts to a kind of job assignment-- the job of standing for something else (pp. 76-77)

This job description for intentional content again begs the question. Being about, or standing in for, something else just *is* what it is to be an intentional state. If we characterize the job of computational states as being go stand in for other things, we have merely assumed that they have intentional content. But, we haven't given an account of what explanatory work-- if any-- that content itself performs.

3.1.2 Explaining Simulations

So, Ramsey's above account of the explanatory role of representations begs the question. But, he does have another account of how intentional content plays an explanatory role in cognitive science. Here, he again points to the ability of computational structures to be mapped homomorphically to relations between states of the world.

He asks us to consider Bob (p. 81 ff.), who is trying to figure out who Alice's brother-in-law is. To do so, he consults a family tree diagram that has symbols such as "Alice" and "Jim" that map onto people, and other symbols, such as arrows and dotted lines, that map onto relations between people, such as being married or siblings. In this case, Bob presumably takes the symbols to *represent* people and the familial relations between them.

Ramsey then (p.84) asks us to suppose that we put Bob in a Searle-style Chinese room and ask him to solve the same problem. In this case, all he has access

to are symbols in a language he doesn't understand and a series of rules telling him how to manipulate the symbols according to their shape. He thus implements a computational process that recapitulates the relational structure of both the family tree diagram and the familial relations holding between actual persons. Only, in this case, Bob himself does not realize that the symbols he's manipulating bear this homomorphism to the world. Nonetheless, he is able to manipulate the symbols according to the rules in order to output the symbol that maps to Alice's brother-in-law.

Ramsey argues that we cannot explain Bob's successful performance on this Chinese room style task without appealing to the intentional content of the symbols he manipulates. He writes:

> if told only that familial relations were discovered through focused attention to shapes and marks on paper, we would find this explanation of Bob's performance grossly inadequate… [W]e would still want to know how he was able to achieve success. We would want to be told what this arrangement of marks could possibly have to do with a familial connection between two people, and how it is that making marks on a piece of paper, and focusing on their shape, could lead to a discovery of that relationship. (p. 85)

To me, it seems we could satisfy these wants simply by pointing to the fact that the computational structure Bob implements is homomorphic to the familial relations holding between Alice and her brother-in-law, Jim. That is what the arrangement of marks "has to do with a familial relation."

Ramsey himself initially seems to endorse this answer, writing:

> *why* those markings eventually provide mindless-Bob with a solution when he uses them in accordance with the instructions… is that the marks on the paper do, in fact, accurately model the real-world family tree… His scribblings on the paper help generate answers because those scribblings share a structural similarity to the relevant state of affairs (p. 85).

But, Ramsey then argues that simply noting the homomorphism between the computational states and familial relations is *not* sufficient to explain Bob's performance on the task. He claims:

> We can't fully understand how mindless-Bob performs the operation of figuring out how two people are related unless we understand his operations as involving the implementation of a model. And to understand his operations as an implementation of a model, we need to look at the elements of these operations-- in particular, the marks on the page-- as representations of people and kinship relations. (p. 85).

Again, this appeal to representational content is ambiguous between an instrumentalist and a substantive reading. On the instrumentalist reading, we would acknowledge that it's useful for us, *qua* theorists, to understand states of the model as being about states of the world, just as it's useful for us, qua navigators, to understand states of a compass as being about states of the world. But, just as we need not attribute original intentionality to the compass to explain its behavior, we need not attribute original intentionality to the model to explain its behavior.

But Ramsey and I are both after a substantive role for intentional content in cognitive processes. The above instrumentalist view won't suffice to provide a job description for intentional content by either of our lights.

Indeed, Ramsey writes:

> It should not be conceded that the only sense in which classical symbols serve as representations in computational processes is the artificial "as if" sense that is only metaphorical… a notion of representation can do explanatory work, *qua* representation, even in a purely mechanical problem-solving system (p. 90).

So, what is this substantive explanatory work that this homomorphic representational content does? Ramsey has two different answers to this question.

The first seems to once again beg the question. The idea seems to be that it is in virtue of corresponding homomorphically to states of the world that computational systems serve as models or simulations of the world. Since being a model or simulation is an intentional notion, we must ascribe intentional content to states of the model in order to explain how it is that the model is *about* the phenomenon it models. Ramsey seems to be making this argument the passage quoted above in which he claims that we must "look at" computational states as representations in order to "understand his operations as an implementation of a model." In a similar vein, he also writes:

> The content of the symbols is explanatorily relevant for their job because if the symbols don't stand for anything, the system in which they function can't itself serve as a model of simulation of the target domain. (p. 87)

If all that were meant by "model" or simulation" was the property of corresponding homomorphically to a target domain, it seems we could explain how a computational system is constituted as a model simply by noting that it instantiates

such a homomorphism. So, it would seem, Ramsey takes models and simulations to be intentional structures that are essentially *about* their target domains. In order to understand the way in which the model itself is about a domain, we must understand how its component parts are about parts of that domain.

But, if the job of the intentional content of computational states is to constitute the intentional content of a computational system as a whole, we have merely pushed the key question up one level. We want to know what role intentional content plays in cognitive processes. If the answer is that the intentional content of computational *states* explains how it is that computational *processes* are intentional, then we need an account of what role the intentional content of these *processes* plays in cognitive explanation. Ramsey doesn't offer such an account. Thus, his description of the job of intentional content begs the question. It *assumes* that computational systems have intentional content that plays a role in cognitive processes but does not tell us what that is.

Ramsey offers an alternative description of the job of these homomorphic representations that is not so question begging, but is nonetheless unsatisfactory. The idea here is that the intentional content of computational states explains how it is that the processes that use them do so "successfully."

Ramsey writes in reference to the Chinese-style room Bob is in:

> [T]he room produces appropriate outputs in response to the inputs it receives. Thus, a computational account of the room would need to include an explanation of how it consistently does this. Syntactic symbol manipulations are only part of the story-- we also want an

> explanation that tells us what it is about those manipulations that produces continued success (p.89)

Presumably, the intentional content of the states might help us in this regard. Indeed, Ramsey further writes:

> A syntactic story would reveal the process whereby the symbols come to be shuffled about in various ways. But it would not tell us what it is about those symbol shufflings that leads the system to consistently produce the appropriate responses. (p. 90).

Now, no doubt, we must say *something* else about the syntactic symbol manipulation to explain how it interfaces with the external world in a reliable manner. But, it's unclear how *intentional content* will be any help in this regard. Indeed, as we noted before, the homomorphic relation between the computational system and the world is sufficient to explain this correspondence to the world. It's because, *ex hypothesi,* there is a homomorphic mapping between the states of the system and familial relations that explains why states of the system can reliably be mapped onto states of the world.

Again, as we've seen, much of what Ramsey writes seems to concur with this view, e.g.:

> the overall system succeeds by exploiting the organizational symmetry that exists between its internal states and some chunk of the world (p. 89).

But, if it's the homomorphic *symmetry* that explains the success of the system, it's unclear what extra job intentional *content* is doing. Ramsey might say that it's in virtue of instantiating such a homomorphism that the computational states *possess* intentional content. But, even granting that, it remains unclear what role that content

plays in explaining the operation of the system. To this, he might retort that ascribing intentional content to the computational states explains how the system as a whole is *about* the domain to which it's homomorphic. But, this is just return to his first job description of intentional content, which we saw was question begging.

So, like Cummins and Gallistel before him, it's unclear that Ramsey has given an account of the role of intentional content that cannot simply be played by the relational homomorphic properties of the states to which he assigns it.

4. Conclusion

The above attempts at the horizontal project largely falter in that they take intentional idiom to merely gloss explanatory roles that have already been characterized independently of any appeal to intentional content. If intentional content is to buy us explanatory power, it must make available to us explanations that we cannot otherwise have. But, none of the accounts we've surveyed here give us that.

For Millikan, intentional idiom records information about Proper functions and Normal conditions. But, it does not buy us any explanations that these notions themselves couldn't give us. Similarly, Neander's teleosemantics tracks counterfactual dispositions to carry information had by ideally functioning organisms, but doesn't by explanations that can't be had simply in terms of natural information and ideal function.

Burge, in contrast, has a notion of intentionality that accords with the things organisms interact with as a matter of fact in their normal environments. But, again, it's unclear we can't then capture the generalizations of intentional idiom solely in

terms of *de facto* interactions organisms have with their environments. Similarly, the functional homomorphism accounts of Cummins, Gallistel, and Ramsey seem merely to record *de facto* correspondence between computational structures and the organism's environment. The intentional idiom doesn't buy us explanations we couldn't have simply by pointing to this structural homomorphism.

Now, for all that, it may be that the vertical, reductive proposals of any one of these accounts is in fact correct. Perhaps mental states come to possess intentional content *in virtue* of fulfilling a teleological function or instantiating a functional homomorphism to their representata. I don't take the considerations I've raised above to have damaged the vertical projects of any of the above accounts beyond repair. My argument here is simply that the explanatory work that these theorists have put this content to is otiose. If the intentional content they ascribe to mental states is to do explanatory work, it must make sense of how the correctness conditions of that content buy us explanations we could not have otherwise.

The next Chapter will examine several putatively intentional systems of increasing complexity in order to see at what point an appeal to intentional content is useful in explaining a computational system, like a mind. In doing so, we'll see another reason why several of the accounts above fail to offer a useful job description for intentional content.

It will turn out that intentional content is useful in characterizing counterfactuals that we could not otherwise. Insofar as Millikan, Burge, and Cummins give us accounts of intentional explanation premised on the *de facto*

relation of a creature to its environment, they will have difficulty characterizing these counterfactuals. Alternatively, as we shall see, because theorists like Neander and Fodor take intentional content to be sensitive to counterfactuals, they come closer to giving a non-vacuous explanatory role to content.

Chapter 3: A Role for Intentional Content

1. Computations Without Representation

As we've noted, given a computational account of a process, we can always abstract away from any semantic content and describe the process in merely syntactic terms. In Chapter 1, we detailed just how such purely syntactic characterizations might work, following suggestions by Piccinini, Stich, Chalmers, and Shagrir. So, if CTM is correct, intentional states can play an explanatory role only if they allow us to provide explanations or generalizations that we cannot make at this syntactic level of description. The theories of intentionality we explored in the last chapter failed to provide an explanatory role for intentional states because they failed to give an account of generalizations that could only be made by way of positing intentional states. The task of the following chapter is to account for just what generalizations intentional states can allow us to make that we could not otherwise.

In what follows, I do not presume to set out the only sort of explanatory function that intentional content might play. My goal is merely to demonstrate that-- hypothetically-- there are conditions in which intentional content does not add any explanatory power to what is otherwise a perfectly good computational explanation of cognition. However, there are likewise-- hypothetical-- conditions in which intentional content can indeed add real explanatory power to computational explanations.

So, our task is merely to demonstrate that some explanations that employ intentional idiom nonetheless do not require intentional content, while others in fact do. We want to discover roughly some of the factors that separate the first sort of explanations from the second. So, let's start with an instance of cognition that clearly does not require intentional explanation and see how we can modify the case until we get to a phenomenon for which positing intentional explanation is illuminating.

The conclusion we'll come to is that intentional content is explanatorily otiose in processes that can be counterfactually generalized over in terms of properties of their proximal input. Intentional content *does* play an explanatory role, however, in making generalizations over systems that are counterfactually sensitive to an indefinite disjunction of proximal inputs.

1. 1. Magnetotaxic Bacteria: Stimulus/Response Mechanisms

Recall that Millikan's magnetotaxic bacteria are single cell bacteria with magnetically responsive magnetosomes. In the northern hemisphere, these magnetosomes orient the bacteria toward magnetic north, which takes them toward the deeper, more anaerobic water in which they thrive.

On the face of it, there are several possible explanations of the bacteria's behavior. It may be that they move in the direction they do because they represent the direction North as such, or because they represent the property of being low in oxygen, and have a desire to move to places with these properties. Alternatively, it may be that the bacteria don't represent anything at all. We may be able to explain why it is they move as they do without attributing any intentional states to them.

Just which explanation in fact applies depends on whether we can generalize the bacteria's behavior in non-intentional terms. If all there was to the bacteria's behavior was this tendency to move toward geomagnetic north, where there happens to be more anaerobic water, then of course there would be no need to appeal to intentional properties. By hypothesis, the only generalization to make would be the one just stated in non-intentional terms!

But, of course, there are more generalizations to be made about the bacteria's behavior. In the first place, generalizations about the bacteria's mobility depend on counterfactual circumstances. Sometimes, they move toward geomagnetic north even though the water there is not more anaerobic. If a bar magnet is introduced in their environment, they will swim toward its south pole[10], regardless of its orientation relative to geomagnetic north. In the second place, there are generalizations concerning how the bacteria generate this behavior. It is not a fundamental physical fact of the universe that magnetotaxic bacteria swim as they do. They all swim as they do because they all instantiate some common mechanism that causes their behavior.

So, in order to capture these generalizations about the counterfactual supporting aspects of behavior, and generalizations about the implementation of that behavior, it may indeed be necessary to appeal to intentional states. It all depends on

---

[10] i.e. the pole that attracts compass needles. The canonical argot is a bit disorienting.

just what these generalizations are. So, we need to look more closely at just what generalizations there are to make about these magnetotaxic bacteria.

The magnetosomes in such bacteria are simply chains of magnetic crystals (usually iron oxide or sulfide) embedded in their cytoplasm[11]. These magnets orient the bacteria along geomagnetic field lines (or local field lines, if, say, a bar magnet is introduced in their environment). When the bacteria move by means of their flagella, they therefore move only parallel to magnetic field lines. It is as though they have tiny compass needles embedded in them that determine which direction they move. The bacteria are not *drawn* to areas of increasing magnetic force: their magnetosomes merely determine the direction that their mechanical movement will take them.

So, in general, the bacteria swim in a direction parallel to geomagnetic field lines because the magnetic crystals embedded in them orient them in that direction. We can thus make all the generalizations we want about the bacteria's magneto taxis without appealing to intentional states. We can make the generalization solely in terms of the proximal magnetic field stimulus impinging on the bacteria and the physical properties of the crystals embedded in the bacteria.

However, the reason that intentional properties are not required in this explanation is not because there is an explanatory level at which we can explain the behavior of an individual bacterium in terms of its physiology. Any particular

---

[11] The following account follows Bazylinski & Frankel (2004).

instance of a physiological organism's behavior can be explained physiologically---- or quantum mechanically for that matter---- without adverting to intentional properties. If intentional properties are to explain anything, it will be generalizations that hold across individual behaviors within or amongst organisms. I am not making the relatively facile argument that the mere existence of an account of behavior at a physiological or brute physical level obviates the need for intentional properties.

Rather, what's striking in the case of the magnetotaxic bacteria is that there don't seem to be any such generalizations that require intentional explanation. Any generalizations about magnetotaxic behavior occurring within one bacterium or amongst many can be captured in terms of the physiological makeup of the magnetosomes. The reason why at different times the same bacterium moves parallel to magnetic field lines, and its conspecifics do the same, is that they all have magnetosomes with a physical makeup that position them parallel to magnetic field lines.

Compare the bacteria's behavior to that of a human walking toward magnetic north. Of course, for any given instance of this behavior, there will be some physical story to tell concerning the physiological effects of stimuli impinging on the human's body. However, it seems on the face of it that this story will be very different in different cases. The human could be walking north because she's looked at a compass needle, the stars, moss growing on the side of trees, or an indefinite disjunction of different stimuli. Presumably, each of these relations to distal stimuli involve different proximal stimuli impinging on the human, eventuating in different series of

physiological changes within her. If we are going to explain, in each case, how it is she travels toward magnetic north, we're going to have to make generalizations over something other than the proximal stimuli impinging on her.

Contrasting the cases of the bacteria and humans suggests a hypothesis. In the case that generalizations about an organism's behavior can be made in terms of the proximal stimuli impinging on it (as with the bacteria), then intentional states won't be explanatorily efficacious. If, however, generalizations about its behavior cannot be made in terms of proximal stimuli, then intentional properties may help to do so.

To examine the first part of this hypothesis, let us look at some more complex cases of behavior that can nonetheless be generalized over in terms of proximal stimuli. In all cases, we'll see that although the behavior effected by the proximal stimuli may be much more complex than that evinced in the bacteria, as long as it can still be characterized in terms of the proximal stimuli that eventuate it, there will be no unique contribution for intentional states to make.

In section 2, we'll look at some examples of behavior that cannot be generalized over in terms of the proximal stimuli bringing it about and lay out an account of how intentional states can help to make generalizations over such behavior.

1.2. Physarum: proximal problem solving

An example of a creature with slightly more complex and flexible behavior that can nonetheless be explained without invocation of intentional content is the slime mold *Physarum polycephalum* (Reid et al., 2012). *Physarum* is a single celled creature

without a central nervous system. It moves by way of small oscillating units of its membrane that help distribute its cytoplasm. When attractants, such as food molecules, bind with receptors on its membrane, the surrounding oscillators vibrate more quickly, causing cytoplasm to flow toward them, and moving *Physarum* in that direction. When detractors, such as light and salt, bind with the membrane, the surrounding oscillators reduce the frequency of their vibration, preventing *Physarum* from moving further in the direction the detractors were detected.

Physarum's navigational abilities not only allow it to generally acquire food and avoid aversive stimuli, but to navigate a U-maze! Imagine a petri dish containing a well of attractive sugar solution at the bottom that diffuses through agar on the dish, so that the concentration of sugar molecules decreases radially as distance from the well increases. If Physarum is placed at the upper end of the petri dish, its chemotaxis should take it directly toward the sugar well, as its oscillators oscillate with greater and greater frequency the closer they get to the sugar solution.

If, however, a U-shaped barrier is placed between Physarum and the sugar, we should expect it to get trapped in the U, unable to reach the sugar. Physarum would head in the direction of increasing sugar concentration before running into the bottom of the U. It then might move to the left or right only to be stopped by the arms of the U. Any attempts to get out of the U would require it to move back up, in the direction of *decreasing* sugar concentration. But, its chemotaxis mechanism, as we've described it thus far, only allows it to move in the direction of *increasing* attractive stimuli. Physarum, it would seem, should get stuck in the maze.

But, that's not what happens. Physarum is able to get around the U-maze and continue toward the sugar. It so happens that Physarum leaves in its wake an extracellular slime, to which it itself is averse. So, Physarum tends not to return to places it has been before because they are coated in this slime, which inhibits the oscillation of its oscillators.

Suppose Physarum enters the U from the left. Once Physarum has explored the bottom of the U-maze, it can move up and out of it on the right side, avoiding its own slime trail. Even though the chemical attraction of the sugar is not as great toward the top of the U, the detraction of the slime Physarum leaves at the bottom of the U is sufficient for what little oscillation the sugar molecules toward the top engender to be greater than that had by parts of Physarum close to the slime at the bottom of the U.

Because they have shown that Physarum can navigate a U-maze, Reid et al. claim that it has a "spatial 'memory.'" Their own use of scare quotes around "memory" may indicate that they in fact balk at actually attributing intentional states to the slime mold. But, it's clear that we can use intentional *idiom* to describe the slime mold's behavior. Why is it able to solve the U-maze? Because it *represents* where it has been and *decides* not to go back. The question before us is whether that intentional *idiom* is picking out any intentional *properties* that actually do play a role in explaining Physarum's behavior, or whether it's just a convenient figure of speech.

Can we make all the generalizations we want about Physarum's behavior in terms of the proximal stimuli impinging on it? There does not seem to be much

known about the precise details of the mechanism by which these chemical stimuli influence the rate of oscillation in Physarum. One hypothesis (Whiting et al., 2014) is that the various stimulants variously inhibit and promote the production of the chemicals cAMP and Ca^{2+}, which may determine the rate of Physarum's membrane oscillation. There may be separate types of oscillator corresponding to different stimuli, or just one type, which can be manipulated in different ways by various stimuli. There's some reason to think the latter hypothesis is correct because of the way in which attractive and repulsive stimuli interact (*ibid.)*. When Physarum is stimulated by both an attractant (oat flake) and a repellant (light) at the same location, it tends to generally cease oscillating. Because the two stimuli seem able to cancel one another out, a possible explanation is that the light stimuli somehow blocks the mechanism by which the oat flake could cause the oscillation to increase (*ibid.*).

Despite not knowing the precise physiological mechanism mediating proximal stimuli and the response of *Physarum's* oscillators, we can characterize the counterfactual generalizations between them without resorting to intentional states. *Physarum* will always move in the direction of the greatest oscillation of its oscillators, where that oscillation is determined by the molecular make-up of the oscillators and the stimuli. This generalization describes how Physarum gets around in normal environments, and how it navigates the somewhat outré environments with U-mazes. No intentional properties are needed to explain its navigational capacities.

So, if we want a case in which intentional properties are doing some explanatory work, we may do well to look at cases in which there are generalizations

about organisms' behavior that seem to obtain despite differences in the underlying physical conditions of the behavior.

1.3. Cataglyphis: Computational Explanation

For such an example, look to the desert ant, *Cataglyphis fortis*, referenced by Burge (2010). More recent research than that Burge references (e.g. Steck et al. (2009) and Wittlinger et al. (2007)) indicates that the behavioral capacities of the ant may be a bit more complex than Burge makes them out to be. But, for present purposes, let's look at the behavior of this ant *as characterized by Burge*, as a useful thought experiment for examining the conditions under which intentional explanations become efficacious.

The ant takes a circuitous, random walk from its nest. Upon finding food, it is able to walk in a direct line back to its nest, rather than simply retracing its steps. Thus, the theory has it that the ant is able to compute a global homing vector by adding together the vectors constituting its outward walk.

Thus, we have a behavior we can make generalizations about despite differences in the physical conditions underlying different instances of it. In general these ants are able to walk in a straight line back to their nest from a food source, but they are able to do so despite changes in the route from their nest to the food source and the location of the food source relative to their nest. We cannot therefore make a physiological generalization about the ant that captures this generalization about its behavior.

The canonical explanation is that the ant has computational states that change

according to the number of steps the ant takes. Because the ant's steps generally traverse uniform distances, these states usually correlate well with the distance the ant has travelled. Other states change in response to changes in the polarity of the light hitting the ant. Because such changes usually result from a change in direction, these states usually correlate well with the direction the ant has turned from its previous heading. The ant instantiates a computational system that transforms these states that correlate with distance and direction so as to generate a state that causes the ant to walk in a straight line that corresponds to the addition of the distance-direction vectors that describe its outward walk.

So, we can account for the generalizations about the ant's behavior by positing a computational system that integrates states corresponding with the distance and direction of the outward walk, and transforming them to form a motor command that in most circumstances gets the ant in a straight line back to its nest. The question is whether positing states with intentional content in addition to this computational process is necessary to explain the behavioral generalization. Burge argues that it is not:

> Path integration in itself requires no spatial *representation*. It computes and utilizes information that *correlates* with spatial properties. The capacities evolved and *function* to enable an animal to find its way in space. But the information states need not represent spatial properties or relations as such. Veridicality conditions play no non-trivial role in explanations of the natures or formations of the states (p. 502).

The question before us is precisely *why* the ant's behavior does not require representational states for its explanation. The same diagnosis as to why intentional

properties were not required to explain the behavior of Physarum or the magnetotaxic behavior applies to the desert ant. In both those cases, we could generalize counterfactually about their navigational capacities solely in terms of the proximal stimuli impinging on them. We can do the same thing with the ant. Given certain proximal stimuli of its polarized light detectors and step counters, we can predict how these states will be combined to create a motor instructions for a return walk[12].

Once we describe how states of the any correspond with spatial properties of its environment and how those states transform via a computational process that corresponds to vector addition, we can explain why the ant is able to get back to its nest in a straight line over an indefinite number of scenarios. We don't need to posit states that can be tokened correctly or incorrectly to explain the behavior. We merely need states that reliably correlate with distance properties in the ant's normal environment.

We can further see why assigning intentional contents to the ant's states doesn't do any explanatory work by noting that whatever intentional content we could assign the ant's states would not change the explanation of the ant's behavior. Whether the states generated by the ant's leg movements represent leg movements, distances, or unicorns, what explains why those states reliably get the ant home is that

---

[12] This is not to say that the ant does not represent lots of other things. It may, for example, represent its burrow and its food as such. The claim Burge makes, and with which I concur, is merely that the *navigational capacities* of Burge's idealized ant can be explained without attributing to it intentional content as of *spatial properties*.

they reliably *correlate* with the distance it's traveled. This correlation explains everything there is to explain about the ant's behavior, so there is no further explanatory work for intentional content to do.

We're now in a position to diagnose more precisely why we can explain both the bacteria's and ants' behavior in non-intentional terms. Burge gets it right when he claims about his idealized ant:

> Explanation of formation of the vector and navigation according to the vector can remain strictly in terms of summing and updating proximal stimulation, and combining it so as to produce states that cause movement of specific parts of the body. (506)

Indeed, for both the ants and the bacteria, the same proximal input will always lead to the same behavioral outcome. In the case of the bacteria, magnetic fields impinging on their magnetosomes will always cause them to orient toward the strongest northern polarity. In the case of the ants, the same proximal stimulation of their leg receptors and polarity detectors will cause them to issue the same motor routine upon being properly stimulated by food. Therefore, as Burge notes, we can generalize their behavior solely in terms of the proximal stimulation impinging on them.

Given what we have said about the ant's navigational mechanisms, we should expect that were we to turn it over on its back, move its legs to and fro, and stimulate its polarized light detectors in just the right way, we could induce it to walk from any one position to another in a straight line. All we need to know are how the states engendered from this proximal stimulation combine with one another to produce the

motor output. The relevant experiments have not been done, but taken as a thought experiment, this would suffice to show that we can make all the generalizations we want about the ant's counterfactually supported navigational capacities solely in terms of proximal stimuli and computational transformations that take place on them.[13]

In retrospect, we can see the project the behaviorist psychologists were engaged in as an attempt to demonstrate that *all* behavior could be explained in terms of counterfactual generalizations over proximal stimuli. I hope to have given good reason to presume that if they had been successful at that project, they would have succeeded in eliminating intentional properties from psychological explanation. What examination of the foregoing bacteria, slime molds, and ants teaches us is that intentional properties are not necessary to make generalizations that can otherwise be made solely in terms of proximal stimuli.

2. Early Perception

Burge believes that whereas the desert any does not employ intentional states in its navigation, perceptual psychology nonetheless makes use of intentional properties in making generalizations characterizing early visual processing. This case is interesting

[13] As noted before, there is some evidence that *Cataglyphis* exhibits more complex navigational capacities than those characterized here and by Burge. So, it's an open empirical question whether actual *Cataglyphis* has intentional states. What is clear, however, is that our *idealized* ant does not require intentional explanation.

for our purposes because on the face of it, the generalizations of early visual processing seem just as easily cached out in terms of proximal stimuli as the navigational capacities we've surveyed above. So, it's initially unclear whether Burge has found generalizations made by this early perception that cannot be accommodated in the terms of proximal stimuli, or whether he believes there is a need for intentional explanations despite the availability of generalizations made over the proximal stimuli alone.

2.1. Burgean Constancies

Burge argues that intentional states play a role explaining the formation of perceptual constancies. Roughly, perceptual constancies are the capacity of perceptual systems to map multiple proximal stimuli onto the same mental state. For example, we exhibit a capacity to see a white sheet of paper as uniformly white even though the light reflected from its surface may give rise to varying proximal stimuli of our retina. The paper itself has a constant reflectance property-- its disposition to reflect light.

However, the proximal stimuli it provides to the retina are variable. The retina will experience intense stimulation from the side of the paper in the light, and weak stimulation from the side of the paper in shadow. Thus, the proximal stimuli impinging on the retina covary with luminance properties of the paper-- the amount of light illuminating the paper.

So, from an input of varying proximal stimuli, we are nonetheless able to treat the paper as having a constant property. The process by which the early visual system

effects this mapping from variable proximal stimuli onto a state that co-varies with one stable property of the world is the process of perceptual constancy.

Burge claims that the process by which the early visual system effects this constancy formation requires states with intentional content. At the outset, this conclusion must seem obvious-- indeed trivial. If constancy formation is described as the process by which we create a stable representational state as of the paper having constant reflectance, then of course constancy formation involves intentional states. It at least involves the intentional state that represents the paper as having uniform reflectance!

However, Burge's claim is not merely that because perceptual discrimination tasks seem to involve representational states that people with such perceptual systems have intentional states that represent, e.g. reflectance properties. I suspect it may well be true that people are able to represent reflectance properties. Indeed, it is just this ability which on the face of it would seem to allow perceptual psychologists to make hypotheses about them! But, Burge's point is more provocative: that intentional properties are involved in the explanation of the *process* that takes proximal stimuli as input and maps it onto representations as of constant reflectance.

This latter claim, I contend, is false-- at least if the process of visual constancy formation operates as Burge claims that it does. It's interestingly false, however, because it is another example of how intentional explanation is unnecessary as long as we can make all the interesting generalizations in terms of proximal stimuli alone.

So as not to bias our investigation, let's characterize the process of constancy

formation in a way that remains agnostic to any intentional attribution. It may be that states of the constancy mechanism (including its output) *have* intentional content. The question before is whether that content plays any role in explaining the operation of the constancy mechanism itself. I contend that it does not, whereas Burge claims that it does.

At base, a constancy must be an ability to sort proximal stimulations into equivalence classes. If constancies allow us to treat distal particulars as the same despite changes in proximal stimuli, then the deployment of constancies must at least require the ability to group different sorts of proximal stimulation all under the same equivalence class.

So, suppose we have a mechanism that takes in diverse proximal stimuli as input, and outputs computational states. Proximal stimuli of type P7, P3, and P5 all produce state S0 and proximal stimuli of type P2, P4 and P6 all produce state S1. Thus, all the proximal stimuli are segmented into one of two equivalence classes picked out by the computational state to which they give rise.

Further suppose that things are set up just so that state S1 tends only to arise in the presence of squares in the organism's normal environment, and state S2 tends only to arise in the presence of triangles. That is, there is a correlation between states S1 and S2 with squares and triangles, respectively.

Can we thereby say that S1 is a constancy that represents squares despite differences in proximal stimulation, and S2 is similarly a constancy that represents triangles? As Burge is continually reminding us, mere correlation is not sufficient to

require explanation in terms of intentional content. So, by Burge's own lights, the answer would seem to be, "no."

We could give an affirmative answer if these constancies played some *further* role in our psychological explanations. From Burge's discussion of the desert ant, it seems that at least a necessary condition on computational states having content is that the truth conditions of the states play some role in explaining the organism's behavior. But, it's not clear that any intentional content the S1 and S2 states might have explains how they come to be activated by the constancy mechanism.

A similar process is at play in the account Burge gives of lightness constancy. Many different patterns of photon stimulation on the retina will cause us to perceive an object as having the same lightness. For example, we can perceive a sheet of white paper as having the same whiteness both in direct sunlight or when it's lit obliquely by a lamp indoors, even though the two situations will cause very different proximal registrations of information on the retina. Burge notes that this capacity to map many proximal stimuli onto the same perceptual constancy does not depend on background knowledge or conceptual capacities (p. 351). It is all bottom-up, as we say[14]. Moreover, it requires that the perceptual system somehow discriminate cases in which different proximal stimuli are caused by differences in illumination on an object with

[14] Chapter 4 examines recent research indicating that early perception may not be so encapsulated. I argue there that the evidence against at least some degree of encapsulation in at least some perceptual processes is not decisive. For now, however, my purpose is just to trace out some in-principle conditions on intentional explanation. Whether those conditions are in fact met or not is of course a matter of empirical investigation.

uniform lightness, or reflectance, and situations in which these differences are caused by differences in the reflectance of the surface of an object, though it is illuminated uniformly. Namely, the perceptual system must distinguish between a zebra and a white horse lit by a projection of stripes of light and shadow.

It is a happy fact about our normal environment, Burge tells us, that sharp changes in light intensity are usually caused by differences in the reflectance of a surface, whereas more gradual changes in light intensity are usually caused by differences in the illumination of a surface. It is a further fortuitous fact that our perceptual system maps proximal registrations of sharp changes in light intensity onto one sort of perceptual constancy, and maps proximal registrations of gradual changes in light intensities onto another type of perceptual constancy. Thus, under normal conditions, our perceptual system is usually able to distinguish pretty well between changes in luminance across a surface of uniform reflectance, and changes in surface reflectance under uniform illumination.

I take all this to be straightforward psychological fact about how our perceptual systems work. Burge goes further to assert that such facts give reason to think that our perceptual states have representational contents that have veridicality conditions. This inference is not warranted. Burge writes:

> The formation principle that I cite in this example describes and explains a law-like process that tends to yield veridical perceptions of distal conditions. (p. 354)

The "formation principle" is simply the principle whereby our visual system maps sharp discontinuities of light intensity onto one constancy and gradual

discontinuities onto another. Insofar as the distinction between these two constancies correlates well with differences between changes in reflectance on the one hand, and changes in illumination on the other, early vision has a computational system that maps homomorphically onto differences between luminance and reflectance.

So, here's our explanation of lightness constancy computation in early vision:

1) Our perceptual system is able to map proximal stimuli onto one of at least two different perceptual constancies, call them "A" and "B."
2) The proximal stimuli the system maps to A tend to be caused by changes in the luminance properties of a distal percept. The proximal stimuli the system maps to B tend to be caused by changes in the reflectance properties of the distal percept.
3) Thus, in environments like ours, the system can reliably discriminate proximal stimuli caused by changes in reflectance from proximal stimuli caused by differences in illumination.

But, as Burge should well know, an ability to reliably discriminate distal causes neither requires nor entails an ability to *represent* those distal causes in such a way that the veridicality conditions of the representations play a role in explaining the behavior of the system. The desert ant can reliably discriminate between a random walk with one trajectory from a random walk with another trajectory. But, this discriminative ability is not enough to ascribe representations to the desert ant.

So, it's not obvious to me that an ability to reliably discriminate distal causes in the manner described above requires representations of those distal causes that have veridicality conditions. Now, of course, we might presume that the outputs of this constancy mechanism possess intentional content that plays an explanatory role

outside of the constancy mechanism itself. But, what Burge has not established is that any intentional content these states might have are playing a role in explaining the operation of the constancy mechanism. Burge, however, begs to differ:

> The explanatory principle that describes and explains the process makes non-trivial reference to representational states with veridicality conditions. (p. 354)

I've searched in vain for these references to these representational states with veridicality conditions. Moreover, given the explanations above, I don't see why such states would be required to explain our perceptual capacities.

Let's look at some potential places at which Burge might think representational states with veridicality conditions may need to be invoked. Burge notes that the formation principles used explain our ability to discriminate proximal stimuli caused by differences in luminance from proximal stimuli caused by differences in reflectance of distal surfaces "can yield illusions" (353). Perhaps Burge thinks that since the perceptual system can generate illusions, it must generate some representations that are veridical, and others that are not veridical, and thus illusory.

Insofar as we can have illusions such that we perceive differences in luminance as differences in reflectance, or vice versa, there is no doubt that we experience illusions. Insofar as our perceptual capacities to *discriminate* distal differences in reflectance and luminance help determine whether we *represent* a surface as having a difference in luminance or reflectance, the perceptual process described above can indeed "yield illusions."

However, all these points do not settle whether intentional content with

veridicality conditions is necessary to explain the *formation* of representations of reflectance and luminance. Even if constancy mechanisms eventuate in the output of representations as of reflectance and luminance, the mere fact that we experience illusions does not determine that *early perceptual systems* themselves utilize these intentional properties with veridicality conditions.

In any case, the mere process by which we discriminate lightness differentials from luminance differentials in early vision does not seem to require states with intentional content any more than does the desert ant's navigational system. So, by Burge's own lights, it's not the case that representations are needed to explain the process of constancy formation.

So, Burge's account of representations in desert ant navigation is at odds with his account of representations in constancy formation. In the former, he demonstrates how we can explain the correlation between mental states and properties of an organism's normal environment without appeal to correctness conditions. In the latter, he seems to argue that mere correlation between mental states and distal stimuli requires explanation in terms of representational states with correctness conditions.

In the end, it does seem that the process of reflectance constancy formation can be explained without appeal to intentional states. It's consonant with the working hypothesis that intentional states are not necessary when all the generalizations we want to make about a system can be done in terms of properties of their proximal stimuli. We should now start refining the hypothesis a bit.

The argument for our hypothesis is quite straightforward. If intentional states

are to play an explanatory role, they must allow us to make generalizations that we could not otherwise. If all the generalizations there are to make can be captured by rules ranging over the properties of proximal stimuli, then there will be no additional generalizations for intentional states to explain. Therefore, any process that can be explained solely in terms of proximal stimuli does not admit of intentional explanation.

2.2 Natural Constraint Inference

This conclusion, of course, flies in the face of Fodor's (1983) claim that it is precisely encapsulated perceptual systems where we should expect to find intentional inferences at work. In this contention, he follows in the long tradition of Helmholtz (1867) and Rock (1983) in characterizing early perceptual processes as "unconscious inferences."

Fodor's reasons for thinking that these processes are intentional seem simply to be a corollary of his view that computations are essentially intentional. Insofar as these early perceptual processes are computational, they must also be intentional. So, for example, in noting that the operations of the language faculty are computational, Fodor argues:

> ...the notion of computation is intrinsically connected to such semantical concepts as implication, confirmation, and logical consequence. Specifically a computation is a transformation of representations which respects these sorts of semantic relations… such semantic relations hold only among the sorts of things to which content can be ascribed (1983, p. 5)

Because the operations of perceptual modules is computational, it must *ipso facto* be

intentional as well. But, we saw in Chapter 1 that there are good reasons not to buy into this premise.

Fodor gives a more detailed sketch of the sort of intentional processes he takes to be operating in early perception:

> ...the character of the energy at the transducer surface is itself lawfully determined by the character of the distal layout. Because [of this] it is possible to infer properties of the distal layout from corresponding properties of the transducer output. Input analyzers are devices which perform inferences of this sort. (p. 45)

This is just the sort of characterization that Burge gives of the lightness constancy mechanism. There's a lawful relation between the sharpness in changes of retinal stimuli and changes in the distal layout due to either reflectance or illuminance. Marr (1982) talks about these sorts of processes as relying on what he calls "physical" or "natural" constraints on the one hand, and "matching constraints" on the other. Natural constraints are regularities in the distal world-- such as the tendency of differences in reflectance to be sharp and differences in illumination to be graded. Matching constraints are constraints on computations that track these physical constraints-- such as the constraint that maps sharp and gradual changes in retinal stimuli to two type-distinct computational states.

As we saw above, absent Fodor's appeal to the essentially intentional nature of computation, there seems to be on the face of it no good reason to attribute intentional content to states involved in these processes. Talking about them as unconscious "inferences" seems to be merely metaphorical. Given our own knowledge about the lawful regularity between reflectance differences and retinal

stimuli, *we, qua* theorists, can infer what distal phenomenon caused a particular retinal stimulus. But, it does not follow that our early vision systems themselves engage in such inference.

Silverberg (2006) provides additional argument for the view that perceptual processes that make use of such matching constraints involve intentional states. He argues that because parts of Marr's theory of vision make use of assumptions about the external world, the computations described by the theory must have content. He takes Marr's account of stereoptic vision as an example. Somehow the early visual system must match up states generated by the left eye that are due to some external feature with states coming into the right eye that are due to the same external feature.

To do this, Marr says that three constraints determine which states may be matched with each other from eye to eye. An "edge" state in the left eye can be matched with an "edge" state in the right eye if and only if these three constraints are met. If the process of matching states obeys these three constraints, Marr claims that the "correspondence is physically correct" (114–15). That is, Silverberg writes, each state in fact corresponds with the same feature in the external world. As Silverberg points out (510), this claim that the process yields "physically correct" states is part of Marr's computational theory. The theory is supposed to explain how vision processing connects up with the external world, and this is one way in which it does so: in our normal environments, states in each eye will in fact co-vary with the same distal stimulus.

However, its explanation of how computational states hook up to external

input is still in terms of "correspondence" and co-variation. Neither of these sorts of relations between external features and computational states require that the states *represent* the external features. The desert ant has states that correspond and co-vary with spatial properties in its normal environment. But, Burge and I both argue that this fact alone is not sufficient to attribute representations of space to those states.

Of course we would need a notion of representation if the *constraints* appealed to in the above account needed to be represented by the early vision system. But, that's not what the theory proposes. It merely claims that the early vision system *respects* the constraints. So, we can't conclude from the fact that the system respects constraints that is representing something.

But, neither can we conclude from that fact that the system is *not* representational. Silverberg chides Egan (1992) for jumping from the observation that the vision system does not explicitly represent the constraints to the conclusion that the relations between external features and computational states are not part of Marr's computational theory (511). He critiques her for then inferring that because the relations between the external states and external features are not part of the computational theory, then even if those relations are representational, the representational states would not be part of the computational theory. Silverberg's critique of this argument is spot on: the relations between the computational states

and external features are indeed part of the computational theory[15]. But again, just because relations between external features and internal computational states are part of the theory does not entail that the relations are representational. Silverberg needs to give us a further argument as to whether the computational states are representational or not.

Of course, if the constraints themselves ranged over representations, we'd need representations in early visual processing. Examples of such constraints would be what Marr calls "physical constraints," such as, "a given point on a physical surface has a unique position in space at any one time" (112–13). If these physical constraints were the constraints early vision operated under, it would indeed need to represent such things as "physical surface," "point," and "position in space." But, Marr says these physical constraints are not the constraints used by the vision system. Rather, he writes:

> We can therefore rewrite the physical constraints as matching constraints, which restrict the allowable ways of matching two primitive symbolic descriptions. . . . For the matching constraints to be valid, the elements of the matched descriptions must *correspond* to well-defined locations on the physical surface being imaged. (p. 114, my italics)

---

[15] Of course, even if we did not include the relations between the computational states and external features as part of whatever we denominated the "computational theory" it would not obviate those relations altogether. If they exist, and are representational, then the system is trafficking in representations, whether or not they are described in the computational theory or some other theory! Just because we can abstract away from mind-world relations or intentional content does not entail that these features play no role in our best explanation of a phenomenon.

So, the states of visual processing must merely respect constraints on which state can match with which other state. These matching constraints are happily constituted such that in most conditions, only states that correspond to the same physical feature will be matched with each other. But, since the constraints range over symbolic primitives, they need not represent the physical features these symbolic primitives correspond to.

So, when Silverberg claims that Marr's theory "contains descriptions of aspects of the physical environment that are external to the perceiver" (512), we can grant that he's right in regard to the *meta-language* of Marr's theory. Marr indeed provides an account of how the internal processing of the visual system links up to the external environment, and, in doing so, *he* describes that external environment. However, the object language of the theory, the computational process being described, does not possess descriptions—*qua* intentional states—of the external environment.

Pylyshyn (2003) presses a similar point about constraint based approaches to vision in general. He notes, as we have, that constraint based accounts of early perception have historically been thought of as "inferential" processes, but claims:

> ...there is another option… [a]ll that is needed is that the computations carried out in early processing embody (*without explicitly representing or drawing inferences from*) certain very general constraints on the interpretations it is allowed to make (p. 96, my emphasis).

Now, such proclamations are largely *obiter dicta* on Pylyshyn's part. He does not mount much of an argument as to why such constraint based processes are non-

intentional. Most of his work is devoted to arguing that such early perceptual processes are cognitively impenetrable. In fact, we might be prone to interpreting the above remarks as merely a recapitulation of the claim that such processes are encapsulated from global cognition-- not that they are non-intentional *tout court.* In fact, Pylyshyn at times stipulates that by "inference" he simply means a cognitively penetrable process: "I prefer to reserve the term 'inference' for a process that Stich (1978) calls 'inferentially promiscuous'-- i.e. that is not restricted in a principled way to what information it can use" (p. 124, n.4).

But, Pylyshyn does make other remarks that seem to indicate he does understand these early, encapsulated processes as non-intentional. In reference to such processes as lightness constancy, the construction of Kanizsa triangle percepts, and others traditionally seen as "intelligent," he writes:

> ...no additional regularities are captured by the hypothesis that the system has knowledge of certain natural laws and takes them into account through "unconscious inferences." Even though in these examples the visual process appears to be "intelligent," it may be carried out by prewired circuitry that does not access encoded knowledge. Notions such as knowledge, belief, goal, and inference give us an explanatory advantage when sets of generalizations can be captured under common principles such as rationality or even something like roughly semantic coherence (Pylyshyn 1984). In the absence of such over-arching principles, Occam's Razor and Lloyd Morgan's Canon dictate that the simpler or lower-level hypothesis… is preferred (pp. 119-120).

Here, Pylyshyn seems to be recapitulating the claim I've argued for above. Appealing to states with correctness conditions in otiose in explaining the operations of early perceptual processes such as the constancy mechanism described by Burge.

All that is necessary to explain such systems is a computational description of the system's states, and the observation that, *ceteris paribus*, there tends to be a correlation between those states and features of the distal environment.

3. Bayesian Constancies

Even if Burge's and other constraint based accounts of constancy formation do not require appeal to intentional content, some, such as Rescorla (2013; forthcoming), and Gładziejewski (2015), have argued that Bayesian processes of constancy formation *do* require intentional content. After all, Bayesian accounts make at least nominal appeal to "hypothesis testing," a paradigm intentional process.

Bayesian models of perception model the processes governing early constancy formation in terms of Bayes' theorem. It is an open question whether Bayesian theories merely do a good job characterizing the inputs and outputs of constancy formation in extension, or whether the theorem models the process of constancy formation in intension as well[16]. Most Bayesian theorists seem not to wrestle very much with the distinction. You might expect that if Bayesian models accurately model the process of constancy formation in extension, they are a good working hypothesis of how the process should be characterized in *intension*.

In any case, I'll remain agnostic on the question of whether or not Bayesian theories characterize the process of constancy formation in intension or not. More

[16] See Jones & Love (2011) and attendant commentary for a spirited debate on this topic.

precisely, for the purposes of argument, I'll presume that they do. For, if the process of constancy formation is indeed described in intension in terms of Bayes' theorem, then on the face of it, it would seem to require intentional (with a "t"!) states. For, Bayes' theorem requires hypothesis testing. Insofar as hypothesis testing is an intentional procedure, if the early visual system implements it, it would seem to require intentional states.

Nonetheless, I contend that even if Bayesian procedures accurately characterize the process of constancy formation in intension, it nonetheless does not require intentional states. For the type of hypothesis testing implemented by Bayes' theorem is not the sort of hypothesis testing that requires intentional properties. Seeing just what sorts of hypothesis testing require intentional descriptions and which do not will help us further refine the types of psychological processes that require intentional explanation.

3.1 Overview of Bayes Theorem

Bayes' theorem is just a corollary of the standard definition of probability:

$$P(A|B)= [P(B|A) * P(A)] / P(B)$$

Simply, the probability of some A, given B, is equal to the product of the probability of B given A and the probability of A, over the probability of B.

If we take A to range over properties of the world and B to range over evidence for those properties, then we can imagine a creature who uses Bayes' theorem to fix her perceptual beliefs. That is, given her beliefs about the likelihood of available evidence occurring given some hypothesis about how the world is, and the likelihood of that hypothesis being true, she can update how likely she believes the world is in a certain state given certain evidence.

On this interpretation, P(A) is canonically called the *prior* and P(B|A) the *likelihood.* The prior is simply an hypothesis of how likely it is that A is the state of the world prior to collecting any evidence. The likelihood is the probability that the evidence would occur given that A is the state of the world. 1/P(B) just ends up being a constant corresponding to the probability that the evidence should show up at all. The probability on the other side of the equation, P(A|B), is called canonically the *posterior*-- the probability calculated after the prior and likelihood have been set.

On most psychological applications of Bayes' theorem, the theorem is applied across an entire hypothesis space. So, rather than considering only the probability of one possible state of the world, the procedure computes the probability that each of a set of states of affairs obtain. For example, instead of calculating the probability that a surface has a particular reflectance property, the Bayesian procedure would calculate the probability that the surface has each possible reflectance property. Thus the calculations involve setting probability distributions for the prior and the likelihood as well. Thus, the prior becomes a probability distribution across different states of the

reflectance property. The likelihood function computes a probability distribution across different possible occurrences of the evidence.

Allred (2012) gives a nice toy example to see how this would work. Suppose again that a creature is trying to establish a reflectance constancy. Here, we can take the evidence to be proximal stimuli on the retina. So, the creature is trying to establish the probability that an object has a given reflectance property given the proximal stimuli impinging on its retina.

Recall that the same proximal stimuli could result from an indefinite combination of reflectance properties and illumination. An object could have very high reflectance under low illumination or very low reflectance under high illumination and yet provide the same proximal stimuli to the retina. So, the probability that the object has any given reflectance will depend on the probability that it is illuminated highly or not.

If all we are given is the proximal stimulation of the retina, there is no way to tell whether it is the result of a surface with low reflectance under high illumination or one of high reflectance under low illumination. The Bayesian procedure gets rid of this ambiguity by hypothesizing that some combinations of illumination and reflectance are more likely than others. It sets a prior distribution that assumes, for example, that surfaces with higher reflectances are more likely to be encountered than those with lower reflectance. It sets a likelihood distribution that considers surfaces with reflectance properties consistent with the proximal stimulation to be highly likely, and those that are not to be highly unlikely. That is, a surface with a

reflectance property that could give rise to the proximal stimulus given some likely illumination would have a likelihood close to 1, whereas one that would not (say, one with zero reflectance) would standardly have a likelihood of 0.

The procedure can also account for noise in the system by assuming, for example, that sometimes the proximal stimulus is brought about in a fashion other than being stimulated from light reflecting off an object. For example, if the proximal stimulus in question sometimes results from spontaneous firing of retinal cells, then the likelihood that it occurs despite the target surface having a reflectance inconsistent with that stimulus can be assumed to be non-zero.

After applying the Bayesian calculations across the probability distributions of the prior and the likelihood, we get as output a probability distribution ranging over different possible reflectance properties of the surface. Insofar as one of those reflectance properties is most probable, we can choose to believe that this is in fact the reflectance property of the surface we're looking at.

As I've described it thus far, the Bayesian procedure sounds highly intentional. Given hypotheses concerning the probabilities of possible reflectance properties and retinal stimulation, we can calculate the probable reflectance property of the surface before us. Hypotheses are paradigm intentional states: to be an hypothesis is to be *about* something.

Nonetheless, I contend that whatever is meant by "hypothesis" in the idiom of some Bayesian perceptual theories, it is not an intentional notion. This view is not without precedent amongst cognitive scientists. Jones & Love (2011) write that:

> Bayesian probabilities are not necessarily psychological beliefs. Instead, they are better thought of as tools used by the researcher to derive behavioral predictions. The hypotheses themselves are not psychological constructs either, but instead reflect characteristics of the environment. (p. 175)

Jones & Love's characterization is a criticism of the evidence often provided for Bayesian models of psychological systems. Their contention is that there is often little evidence that postulated Bayesian models are actually implemented psychologically by subjects. Rather, the Bayesian accounts just accurately describe subjects' behavior in extension, abstracting away from whatever psychological mechanism actually drives the behavior.

In any case, Jones & Love's word does not establish the non-intentional construal of Bayesian perception any more than the ostensibly intentional idiom used by Bayesian theorists establishes the intentional construal. In what follows, I'll assume that Brainard et al.'s (2006; 1997) Bayesian account of color constancy[17] is actually implemented psychologically. Nonetheless, I argue it can be given a non-intentional construal while preserving all its generalizations.

3.2 A Non-Intentional Bayesian Process

Brainard's model goes beyond mere lightness constancy to model the process of color constancy. As with lightness constancy, the perceived color of an object is a function

[17] A non-technical overview of the account is in Allred (2012)

both of its reflectance properties and the illumination. The reflectance and luminance properties affect not just the intensity of light impinging on the retina, but the wavelength as well. An object that appears white when illuminated by white light will appear red when illuminated by red light. That same object illuminated by red light can provide the same proximal wavelengths to the retina another that is illuminated by white light, provided that their reflectance properties are different.[18]

Therefore, if we know the luminance properties of a scene and the properties of the light reflected off a surface in that scene, we can determine the reflectance properties of that surface. This would be difficult for the early visual system to do, however, because all it has access to are states that correlate with properties of the light reflecting off of surfaces onto the retina.

Brainard et al.'s proposal is that the visual system uses a Bayesian algorithm to estimate the luminance properties of the scene from the information about the reflected light and hypotheses about the luminance likely to have caused such reflected light. Put this way, of course, the proposal is rife with intentionality, but as we detail its operations, we can see how it can be cached out solely in terms of generalizations operating over proximal stimuli.

---

[18] The situation is further complicated by the phenomenon of metamerism, in which two different surface reflectance properties generate the same proximal stimuli under one illuminant, but different stimuli under another.

To start off, let's characterize more precisely the proximal stimuli used by the algorithm. Three different cone cells in the retina contain different pigments that change from one molecule to another-- or isomerize-- in response to contact with different wavelengths of light. Rates of isomerization co-vary with the wavelength and intensity of light impinging on the retina. Assuming Burge's principle that co-variation does not entail representation, there's no reason at this stage to assume that any properties of the light are *represented* in the early visual system. States of retinal cells simply co-vary with these properties.

Brainard et al. first estimated the rate of isomerization of the three different color cones when they are struck by different wavelengths at different intensities. They then created a probability distribution that describes the likelihood that any given light stimulation of the retina caused any particular rate of isomerization for each cone. For example, each cone is likely to isomerize at a particular rate when exposed to lightwaves of 650 nm. They will isomerize at this rate whether or not the light is coming from a white object illuminated by light of 650 nm or from an object illuminated by white light that only reflects back light of 650 nm. Thus, the likelihood that this particular isomerization rate of the cones is the result of either of these states of affairs is equally high. By contrast, the likelihood that the cones isomerize at this rate when, for example, being stimulated by light of 300 nm would be quite low on the probability distribution.

Thus, this probability distribution of isomerization given proximal wavelength stimulation constitutes the likelihood function in a Bayesian calculation. Now, as we

have just described it, this probability distribution just is a description of the actual probabilities of cone isomerization *as estimated by Brainard et al*. They do not detail how precisely such a distribution would factor into a constancy mechanism implemented by an actual visual system. So, it is up to us to consider how this could be done.

Obviously, the visual system does not rely upon the estimation of cone isomerization provided by Brainard et al.! There are two alternatives that seem more likely. The first is that the early visual system *represents* a probability distribution of cone isomerization rates given properties of light impinging on the retina. That is to say that there would be a state of the early visual system that represents cone isomerization rates as such, and further, represents wavelengths of light and the probability that any one of them caused the isomerization. This would be in line with the reading that the likelihood function constitutes an hypothesis in the intentional sense.

A second possibility is that the early visual system has states that are homomorphic to the probability distribution described by Brainard et al. Given that the output of the constancy mechanism is a representation of a particular luminance property, there could be states of the early visual system that tip the scales more in the direction of particular outputs than others. Under the intentional interpretation, the visual system would *represent* a probability distribution such that that, say, an isomerization rate of 5 units is believed to have a .98 probability of being caused by wavelengths of 650 nm and a .01 probability of being caused by wavelengths of 300

nm. Under the non-intentional, homomorphic construal, an isomerization rate of 5 units would *cause* the system to go into a state that would weight the system toward outputting a representation of 650 nm by .98 units, and weight it toward outputting a representation of 300 nm by .01 units. This would be consonant with the "vote-counting" interpretation given by Jones & Love.

There is no reason at this stage to think that the intentional implementation of the algorithm is correct. However, we can't rule it out until we consider how the likelihood function interfaces with the rest of the Bayesian algorithm. For now, suffice it to note that there is no reason to prefer the intentional construal unless the further operations of the algorithm require us to do so.

The second element of the Bayesian algorithm is the prior. To determine this probability distribution, Brainard et al. estimated the probability that different illuminations would be encountered by a typical human subject. They relied on data gathered by DiCarlo and Wandell (2000), who sampled the illuminant properties of light outside a window at Stanford every minute from dawn to dusk over 20 days in a variety of weather conditions. The prior distribution set the probability of given illuminant being present in line with the frequency with which it appeared in this sample.

Again, there are two ways in which we could imagine the prior functioning in the psychological implementation of Brainard et al.'s model. It could be that visual system represents many possible illuminants as such and the probability that each of them will occur in line with the way Brainard et al. have represented such

probabilities. Alternatively, the visual system could implement a series of states that is homomorphic to the structure of the probability distribution described by Brainard et al.

Again, the idea would be that different states of the visual system count more in favor of outputting some illuminant judgments than others. Before, any proximal stimuli have been received, the visual system is more disposed to treat them as coming from some illuminants than others. The question before us, then, is whether we can combine these non-intentional descriptions of the likelihood function and the prior in a way that describes how the visual system can generate a representation of illuminant properties (other than the intentional output, of course).

Suppose that the early visual system is composed of a series of tubes that can be filled with water. Since the Bayesian explanation of the visual system abstracts away from its physical implementation, we need not worry how realistic such a supposition is. There is one tube corresponding to each possible illuminant value.

Suppose now, that when lightwaves of 400 nm impinge on the retina, they cause the three cone types begin isomerizing at rates of units 1, 3, and 5 units, respectively. This isomerization causes water to be pumped into the tubes. There is a bijective mapping between the quantity of water pumped into the tubes and the probability values in the likelihood distribution described by Brainard et al. Suppose, for example, that the likelihood function ascribes a probability of .9 to these rates of isomerization being the result of either 400 nm illuminance and white light illuminance, but only a .2 probability of the same rates being caused by 500 nm

illuminance. Then the tubes corresponding to the 400 nm and white light illuminance will be filled with .9 units of water, and the tube corresponding to 500 nm illuminance will fill with .2 units. Thus, the amount of water pumped into the tubes is homomorphic with values of the likelihood function described by Brainard et al..

Each tube is designed such that a certain percentage of the water pumped into it leaks out each time[19]. Again, there is a bijective mapping from the amount by which the water in each tube decreases and the probability values of the prior distribution described by Brainard et al. For example, suppose the prior assigns a probability of .01 to illumination of 400 nm, a probability of .9 to white light illumination, and .05 to 500 nm illumination. Then the water in the tube corresponding to 400 nm will always decrease by 1-.01=99%, that in the tube corresponding to white light by 1-.9=10%, and that in the 500 nm tube 1-.05=95%.

At the end of the process, different tubes will hold different amounts of water. That corresponding to illumination of 400 nm will have .01*.9 units=.009 units; that corresponding to white light illumination will be filled with .9*.9=.81 units, and that corresponding to 500 nm illumination will have .05*.2=.0125 units. Hence, the water in the tubes at the end of the process will be homomorphic to the posterior probability distribution calculated by Brainard et al.'s formula, where

---

[19] You could imagine, for example, a system by which the weight of the water in the tube depresses a spring that lifts a gate on a hole in the bottom such that the opening of the gate, and hence the amount of water released, is proportional to the quantity initially in the tube.

$$\text{posterior}(H_i) = \text{prior}(H_i) * \text{likelihood}(H_i)$$

such that H_i ranges over different possible illumination values. Posterior(400nm)=.01*.9=.009, posterior(white light)= .9*.9=.81, and posterior(500nm)= .05*.2=.0125. The values of posterior(H_i) can be mapped onto the same values as those of the quantity of water in each tube.

The foregoing thought experiment is *not* intended to establish that because there is *some* possible physical instantiation of this Bayesian algorithm that intentional properties are not needed for it to be explanatorily efficacious. In fact, it's premised on the idea that the Bayesian explanation abstracts away from any particular physical implementation. Rather, the point of the thought experiment is that for any arbitrary physical implementation we choose, we can make the generalizations of the Bayesian explanation without attributing intentional states to the early visual system itself.

Bayesian accounts of early perception can be thought of just as the explanations of desert ant navigation that point out homomorphisms between states of the target system and some mathematical description of states of affairs. In the case of the ant, researchers represent spatial properties in terms of a vector algebra, and then demonstrate that states of the ant are homomorphic to states of that representation. In this case, researchers can represent the probabilities of various illuminance properties occurring in the world, and then point out that the early visual system could operate via states that are homomorphic to states of that representation. In neither case do

formal homomorphisms entail that these states are themselves intentional. But, the homomorphisms do account for how the states of the system interact and how they are able to produce output that co-varies with distal states of the world. There are no further generalizations that further appealing to intentional properties might hope to explain.

4.Intentional Explanation of Isotropic Processes

The point of this examination of Bayesian perception is not merely to provide another instance of an ostensibly intentional process that can nonetheless be construed non-intentionally. It also gives us insight into what sort of processes *would* require intentional states. The Bayesian explanation of perception just is to show how the process of constancy formation can be modeled as formally homomorphic with an intentional process of hypothesis confirmation. So, we can look to see what makes the process of Bayesian hypothesis confirmation intentional, while leaving the perceptual processes it can be used to model non-intentional.

It's true that when we *use* Bayesian reasoning to calculate the probability of something, we take the prior and the likelihood to be hypotheses about the probability of certain illuminant occurring in the world. For example, suppose a satellite imaging analyst is trying to figure out the color of some object from the light sensed by a satellite camera. He could make an hypothesis about the probability that any given light is the illuminant of the scene in question. He might start out just assuming that the object is illuminated by white light. But, then perhaps he reads DiCarlo's article

and revises his hypothesis to reflect their findings that there is a probability distribution over a variety of illuminants for outdoor scenes. He might then realize that the conditions at the location of the object are relevantly different from those at Stanford where DiCarlo gathered his data. Because of red soils and pervasive cloud cover around the location of the object, he concludes that the probability distribution of various illuminants is different for the location of the object than in Stanford. Thus, his hypothesis concerning the probable illumination of the object is open to revision in light of a variety of data.

Contrast this case of the satellite analyst with that of the early visual system. The Bayesian account of illuminant constancy formation does not include an account of how the prior probability distribution is set, let alone how it could be revised. The assumption is that the prior is fixed to be roughly homomorphic to the actual probabilities of various illuminants occurring in typical environments. All that's necessary for the explanation to go through is that this homomorphism obtains.

The same contrast holds for the likelihood. In the model of the early visual system, there is no mechanism to revise the values assigned to the likelihood. An analyst using a camera, however, can revise his hypotheses concerning how likely certain proximal stimuli are to occur on the camera sensor given a particular distal state in light of new information about how precisely the sensor works. So, while the prior and the likelihood can be thought of as hypotheses, they are not open to testing and revision in Bayesian models of the early vision system.

This contrast between the intentional construal of the Bayesian formalism in the case of the satellite analyst and the non-intentional construal in the case of the early vision system suggests a possible role for intentional explanation. On the face of it, the analyst's use of Bayes' theorem involves intentional states, whereas the Bayesian account of early vision does not. In the first case, the prior and likelihood are open to revision, whereas the latter is not. So, it may be that a sufficient condition on intentional processes is that they involve states that are open to revision.

Rescorla (2015b) adopts something like this view. For somewhat different reasons, Nico Orlandi (2014) also comes to the conclusion that early visual Bayesian processes are non-intentional. Rescorla (2015a) raises an objection to her view. He writes:

> If we adopt a realist perspective on priors, we can explain why various changes in environmental conditions yield various changes in the mapping from sensory stimulations to percepts.

By a "realist" perspective, Rescorla means an interpretation that priors are in fact intentional. So, for example, a prior in the Brainard et al. model would *represent that* light of such-and-such wavelength is 88% likely to be illuminating a scene. Such an intentional interpretation is supposed to explain data about changes in the response of the visual system in light of experience.

The data Rescorla has in mind here is that such as Adams, Graf, & Ernst's (2004) finding that priors can be altered in response to exposure to novel environments. Whereas most humans seem to assign high prior probability to light coming from above, this prior can change if subjects are

consistently exposed to haptic feedback that's at odds with what they initially expect given their visual perception of the object.

He contends that it's not obvious how a non-intentional Bayesian algorithm could explain this phenomenon. He complains that Orlandi's contention that the early vision system simply "rewires" itself in response to such stimuli does not explain "*why* a given stimulus history yields a given mapping from retinal inputs to percepts." Presumably, the puzzle here is why distal stimuli from haptic feedback causes the visual system to rewire itself in just such a way so as to assign lower weight to the state associated with light coming from above . Given the novel proximal stimuli it's receiving, the visual system could have rewired itself some other way or not at all. The negative evidence wants explanation.

An intentional construal of the Bayesian processes could explain this by simply noting that the visual system lowers the probability of downward projecting light because it *represents* that light is now less likely to come from above. But, there is also a possible non-intentional explanation on the table.

Take again the Brainard et al. model. Suppose that this system increases the prior probability that corresponds with ambient red light upon immersion in an environment consistently lit with such red light. This change amounts to increasing the resistance of the spring that regulates the outflow of water from the tube. The more resistant the spring, the lower the percentage of water will flow out from the tube.

The task Rescorla has set us is to figure out a way to increase the resistance of the spring in proportion to the amount of ambient red light the visual system is presented with. Suppose that after each operation of the constancy mechanism, the remaining water drains out of the tubes. The water draining out of each tube coats the spring attached to that tube. Suppose further that this water bath causes a change in the chemical composition of the spring such that it gains more resistance the more water to which it is exposed.

If the visual system is consistently exposed to an environment with lots of red ambient light, the water in the red tube at the end of each operation of the mechanism will be higher than that in the tubes corresponding to, say, green and blue light. Upon repeated exposure to this red ambient light, therefore, the resistance of the spring associated with red illumination will increase relative to other springs (save that associated with white). This effects an increase in the state corresponding to the red illumination prior in response to repeated exposure to red ambient light. We need not have supposed that any of the states *represent* the frequency of that red light in order to effect the change.

Now, this is a just-so story and I don't begin to suggest that it remotely explains the particular data concerning prior change given by Adams et al. and others Rescorla cites. I leave it open whether a more detailed analysis of the Adams et al. results do require intentional characterization. The point of the story is just to establish that it's *possible* to implement a Bayesian algorithm

in which the *prior* states are responsive to changes in the distal stimuli in the right way without appealing to intentional properties. The mere fact that a Bayesian algorithm changes in light of experience is not in itself reason to suppose that adjustment of the prior must be an intentional process.

Note, again, the argument is not the facile one that because we can give an implementational story that we need not appeal to intentional properties at a computational level of explanation. I talk in implementational terms only to demonstrate how a feedback loop at the computational level between the prior hypotheses states and the output of a Bayesian system could in principle adjust those priors in response to the relevant distal stimuli.

Yang (2002) provides a rather less outré example of how just such a feedback loop at the computational level can explain language acquisition. Yang's model proposes a prior hypothesis space of candidate grammars. Upon being presented with a sentence, children first analyze it with the grammar currently enjoying the highest probability value. If the grammar successfully parses the sentence, then it is rewarded with an even higher probability. If it fails, it is punished with a lower probability value. The prior hypothesis space for analyzing the next sentence is thus set by the posterior distribution of the current analysis.

Yang thus shows how children brought up, for example, in Dutch speaking environments can revise their priors so as to afford greatest weight to hypotheses associated with Dutch style idiolects, whereas children in English

speaking environments gradually accord the greatest weight to hypotheses associated with English style idiolects.

I take it to be an open question to what extent Yang's story requires that things like grammars, sentences, and the like be represented by the Language Acquisition Device. What seems clear is that the adjustment of the hypothesis space proceeds without needing to represent anything concerning the *frequency* of various E-languages, such as English and Dutch, in the environment. It simply responds to a one-off, up or down, ability to parse a sentence at hand. In that respect, we can explain how the hypothesis space adjusts to be in accord with its language environment without appeal to intentional content.

What Yang's study also makes evident is the considerable empirical research and computational modeling required to argue that such a process is in fact psychologically implemented. If we wished to vindicate Orlandi's account of rewiring in the visual system, we'd have to conduct just such work. Since we haven't, Rescorla's critique of Orlandi may still hold: it's *not* obvious that we can explain revisions of priors in Bayesian visual systems without appealing to intentional states.

But, my toy example and Yang's work also demonstrate that the mere ability to adjust priors such that they more closely track distal stimuli is not demonstrative evidence that a Bayesian system is intentional. A more narrow

hypothesis, then, is that processes open to certain *sorts* of revision are amenable to intentional explanation and others are not.

The key difference between the revisions open to the hypotheses of the satellite analyst and those open to the "hypotheses" of the Bayesian perception mechanism are that the latter can be characterized in terms of properties of the proximal stimuli impinging on the retina. The Bayesian perception system will output a given state if and only if the cones isomerize at a particular rates. Once those isomerization rates are fixed, so is the output of the system.

Therefore, we can generalize how the system works solely in terms of proximal stimuli and the relations amongst its internal states. Given any particular isomerization rate of the cones, downstream states will be tokened in a manner homomorphic to the formal statement of the Bayesian algorithm. This homomorphism will ensure that in most contexts the states output by the visual system co-vary with the luminance properties of the external world. There aren't any other generalizations about the operation of the system that intentional states would be required to explain.

By contrast, if the prior and likelihood are open to revision by a meddling satellite analyst, the states of whatever plays the same role as the cones in the spy satellite will not determine what the output of the satellite is. Therefore, on the face of it, we will not be able to generalize about the operations of the satellite's Bayesian system solely in terms of the proximal stimuli impinging on it and the relations amongst its states.

Nonetheless, there are still instances in which we could generalize about the operation of a Bayesian perceptual system even if the likelihood function, for example, was open to revision. Suppose, for example, there was a state that covaried with the degree of dilation of the pupil. This state in turn determines which of a variety of likelihood distributions will be used in calculation of the posterior. Perhaps it's a happy fact about the world that the likelihood that a particular rate of cone isomerization is caused by a particular incidence of light on the retina does in fact change depending on how wide the pupil is. For example, light of 500 nm could be slightly less likely to cause cone isomerization a particular rate when the pupil is highly dilated than when it is not. The last desideratum would of course be that the actual differences in likelihood roughly co-vary with the degree to which the likelihood function changes as result of pupil dilation.

In this case, even though the likelihood function gets revised in different cases, we can generalize about how it gets revised-- and how the system functions in light of that revision-- solely in terms of proximal stimuli. We need only expand the proximal stimuli over which we generalize to include the states that co-vary with pupil dilation. That is, the output of any state will be determined by the conjunction of cone isomerization rates and pupil dilation. There are not any additional generalizations about how the system works that would be amenable to intentional explanation.

What is interesting about the hypothesis testing of the satellite analyst is that the circumstances under which he is liable to revise the values of the prior and the

likelihood distributions are not characterizable in terms of the properties of proximal stimuli impinging on either him or his satellite. This is not because we imagine him to be capricious, changing the probability distributions at random. What's most interesting is the fact that he is very deliberate concerning the circumstances under which he changes the prior and likelihood distribution. So deliberate, in fact, that it seems that we should be able to make generalizations about the circumstances that would cause him to revise them one way or the other. However, what seems clear is that we cannot do so simply in terms of the proximal stimuli impinging on him.

The behaviorist psychologists certainly had hoped that we'd be able to account for his behavior in terms of the responses engendered by proximal stimuli. But, they turned out to be wrong. It's not immediately clear how we can characterize just what kind of stimuli are relevant to revising his hypothesis concerning the prior distribution of probable illuminants. In general, to pick up a Quinean theme, it looks as though his hypothesis is isotropically revisable. In principle, just about any proximal stimulus could cause him to revise his hypothesis. We therefore cannot generalize about the situations under which he would revise his hypothesis in terms of the properties of the proximal stimuli presented him in those circumstances.

Here, then, is a case in which mere computational explanation fails to characterize all the generalizations we'd like to make. When we have states that are so isotropically revisable there may well be room for intentional states to play an explanatory role.

This is just to acknowledge what everyone knows: that the behaviorists were

wrong about intentional properties and the scope of psychological explanation generally. What is not always fully appreciated is that they were wrong on at least *two* counts. In the first place, it turned out that psychological explanation could not get along without positing internal states that stood in computational relations to one another. In the second, it turned out that it could not get along without intentional states as well. What cases such as Bayesian explanations of perception demonstrate, however, is that sometimes it's sufficient for psychological explanation to rely on computational explanations while eschewing intentional explanations.

Why this is apparent already in Chomsky's critique of Skinner's *Verbal Behavior*[20] that helped to launch the cognitive turn in psychology. There are indeed two aspects of Chomsky's critique. First, if human language can be described in terms of laws generating output from input, the rule relating the inputs and the outputs must be much more complex than the simple Law of Effect posited by the behaviorists. The theory of computation provides a nice way of giving nomological characterizations of such complex interactions. The best hypothesis, then, is that language is the result of a computational mechanism.

Now, why posit that states of the language-speaking organism are *intentional* in addition to being computational? Chomsky noted that, in the presence of a red

---

[20] Though, see Collins (2007) for an argument that Chomsky's review does not endorse intentional explanation. I take Collins' position to be that Chomsky endorses the critique that the framework of stimulus, response, and control, etc. are insufficient for psychological explanation, while remaining agnostic on whether intentional explanations should supplant the behaviorists' project.

chair, you might either say, "red" or "chair." If Skinner is to be believed, whether you respond with "red" or "chair" depends on whether the redness of the chair or the chairness of the chair is the "controlling stimulus."

You might explain which stimulus is controlling by appealing to your newfound computational theory of mind. Whether the redness or the chairness is controlling depends upon your mental state upon encountering the chair. Were you in state 1, the redness would become controlling; were you in state 2, the chairness would become the controlling stimulus[21]. Thus, *prima facie*, it seems as though we could explain why we sometimes say "chair" and sometimes say, "red" by appealing only to our *computational* theory of mind, without adding in any representational theory.

This is precisely why a computational account of the desert ant's navigation works so well. After all, the same leg movement on the ant's outward walk will evince very different behavior on its return walk depending upon whether it has occurred after a change of state in the polarized light detector or after. We can explain the difference in output behavior in terms of how the input behavior computationally interacts with other states of the ant.

So what are the circumstances that would require an intentional theory in addition to this computational theory? There are interesting generalizations we can

---

[21] Presuming, of course, that redness and chair-ness are distal properties of your environment-- a presumption that is far from clear.

make about the circumstances under which we say "chair" and those under which we say, "red," that cannot be captured by our computational theory alone. It may well be that whether we say "red" or "chair" when stimulated by a chair depends on whatever computational state we are in upon encountering the chair. However, there seems to be an indefinite number of states in which we might either say one or the other. Saying "red" might be preceded by a conversation about the color scheme of a room, a scavenger hunt for objects of a certain color, or an attempt to escape a charging bull. There are thus indefinite computational routes to an internal state that may cause someone to utter "red."

In any given instance, of course, we could in theory discover just what particular computational route took a person from a beginning state, to being stimulated by a chair, to uttering "red." However, it's not clear that in the general case we can make a generalization about what all these computational routes have in common such that they engender an utterance of "red."

The situation is just the same as that of our satellite image analyst, who will revise his hypotheses concerning the prior and likelihood of various illuminants given a large set of stimuli that we cannot generalize over. Reading articles about the frequency distributions of illuminants, taking his own measurements of such data on a spectroradiometer, hearing about the ambient conditions of the sites his satellite is photographing; these could all lead him to revise the setting of his prior.

If we hope to make any generalizations about why he changes the prior distribution the way he does in these indefinite scenarios, we'd want some sort of

generalization about the *types* of computational states that may lead to the behavior. A good proposal is that the type of states that get him to revise the prior distribution are those that are *about* the frequency of illuminant properties. Thus, attributing to the analyst the ability to *represent* illuminance and frequency properties allows us to make generalizations about his behavior not open to us otherwise.

The types of computational states that can get me to say "red" are those that are *about* the redness of the chair as such. Attributing to me an ability to represent the redness of the chair allows us to make generalizations about my behavior despite there being indefinite computational routes to that behavior.

In the case of the ant, we could individuate the states over which its computations operated in terms of their counterfactually supported causal powers independent of any intentional properties. Why does the state engendered by a leg movement lead to one motor output if it is tokened after a change in input from the polarized light detector, and another one otherwise? It's because the state is one individuated in virtue of it having just those causal powers! This buys us all the explanatory power we want in the case of the ant.

This style of explanation does not look to be true of human verbal behavior. Why did I say, "red" upon encountering a chair? Well, of course it's because I was in some particular cognitive state that caused me to do so. But, notice, we can't individuate that state in terms of this causal power and still hope to make interesting generalizations about the *types* of states that cause me to utter, "red" in response to chairs. To say that the types of states that cause such behavior are the states that cause

such behavior in the particular circumstances I happen to be in is vacuous.

In the case of the ant, we were able to individuate computational states in terms of the proximal stimuli that give rise to them in a counterfactual supporting way. The state that gives rise to differential motor output depending on whether it's tokened by before or after a change in state of the light detector is one that is caused by particular proximal stimulation of the ant's leg across any manner of contexts.

However, we cannot type individuate the type of state that causes me to utter, "red" in terms of the proximal causes of that state. For, it turns out, an indefinite number of proximal stimuli could lead up to the state that causes me to utter, "red." So, there's something more to be said about the type of state that causes me to utter, "red." It's not simply a state with the particular causal properties of the particular state that caused me to say "red" in this particular instance. We want to individuate a state that has the property to cause me to say "red" in response to an indefinite disjunction of proximal stimuli.

We can think of this distinction on analogy with the common distinction between implementational and computational levels of explanation. On the implementational level, we can explain, say, a particular instance of the ant's behavior in terms of physiological interactions between its parts. But, once we notice that there are an indefinite number of physiological interactions that could have evinced just the same behavior, we generalize over them by individuating states of the ant in terms of computational role. Thus, there is a state of the ant, possibly variously physiologically realized, that is individuated in terms of its power to be caused by

particular proximal stimuli and to cause transitions to other computational states. Thus, at the computational level of explanation, we can say that what caused the ant to evince a particular behavior was not necessarily a state of a particular physiological type, but, more generally, a state of a particular computational type.

In the same way, we can note that there is some state, variously physiologically realized, that causes a person on a particular occasion to say "red" in virtue of encountering particular proximal stimuli. But, we can further notice that there are other states that could evince just the same behavior despite differences in the proximal stimulation or computational route that caused the state to be realized. So, at a level above the computational level, we can say that it's not necessarily a state of a particular computational type that caused the verbal behavior, but rather a state of some higher order type, variously computationally realized, that caused the verbal behavior.

The contention of this chapter is that the higher order type of state we should be talking about is a state type-individuated by its intentional properties. The explanatory role that intentional properties play is that they allow us to type individuate states at a level above computational role in just the same way that computational properties allow us to type individuate states at a level above their physical constitution.

What we see in Chomsky's review of Skinner is a need for two additional levels of explanation in psychology that were not acknowledged by the behaviorists. First, there were behaviors whose explanation could only be captured by

generalizations over computational relations holding between states. Further, there were aspects of cognition, such as verbal behavior, the generalizations of which could not be captured by computational relations alone. Another level of explanation was needed to type individuate states independently of their computational role.

It is only by talking about states that are individuated in terms being *about* redness as such and chairness as such that we can make generalizations about conditions under which humans talk about the redness of a chair rather than its chairness. An indefinite number of proximal stimuli can give rise to an indefinite number of computational routes to an utterance of "red." We can describe what they all have in common by holding that they all at some point token a state that is individuated in virtue of it being about redness as such.

5. Sufficient Conditions for Intentional Explanation

All this has been rather abstract, so let's get back to some more concrete examples to see how intentional, properties can function to make generalizations across computational properties. Let's go back to our desert ant and start building up its navigational capacities until we find intentional properties necessary to explain them.

Suppose that we increase the type of proximal stimuli that can get the ant to revise the motor instructions it outputs for its homeward walk. We could, for example, suppose that the ant has hairs that respond to gravitational field strength,

tokening states that co-vary with its vertical displacement.[22] If the states tokened by these hair movements enter into the right relationships with those tokened by the polarized light detectors and step counters, the ant could instantiate a straight-line walk back to its nest despite an outbound walk that took it up and down over hills. But, still, we could explain this ability in terms of states individuated in terms of their propensity to be caused by particular proximal stimuli and to enter into computational relations with other states so individuated.

In fact, it looks as though we can increase the number of proximal stimuli to which the ant's navigational system is responsive nearly without limit and still be able to explain its navigational capacities in non-intentional terms. We need only note just which proximal stimuli the ant is sensitive to and how the states those stimuli engender combine in order to describe the behavior the ant will exhibit. The only reason we'd have to step up to another level of generalization is if there's no way of characterizing the states of the ant in terms of the proximal stimuli they are caused by.

Now, it might be objected at this point that appealing to intentional properties of the ant's navigational states *will* help us explain how they interface with other aspects of its cognition that are clearly intentional. For example, suppose the ant has a *desire* to go home once it has found food. You might suppose that we can only explain why the ant walks in the direction of home after finding the food by

---

[22] It looks as though something like this might be the case (Grah et al. 2005), though see Grah et al. (2007) for evidence that complicates this view..

attributing a practical syllogism to it. It *wants* to get back home, it *believes* that home is in a straight line to the northwest, say, and therefore it walks in a straight line to the northwest. In order for the ant to implement such a practical syllogism, it would seem that we *do* need to attribute representations of the spatial layout of its surrounds.

Now, it is true that the ant does act *as if* it is relying on such a practical syllogism. The question before us is whether the utility of this locution is *merely* as a useful metaphor, or whether it provides us additional explanatory power. There is a way of glossing the ant's behavior such that we can account for how it fulfills its desire to travel home without supposing that the ant further represents the direction of its home, where it is now relative to its home in allocentric space, or any other spatial properties[23].

Think of the ant on analogy with a human riding in an automated car that implements the same navigational algorithm as the ant. The car's onboard computer registers rotations of its wheels, and a sensor on top registers changes in light polarity. The computer integrates these inputs so as to compute a constantly updated homing vector that, once activated, will take the car back to its place of departure. Suppose the human riding in the car sits back as it performs a random drive through, say, the Atacama salt flats. The human has a desire to see a flamingo, and then to drive back home once she has seen one.

---

[23] We are, of course, *ex hypothesi*, presuming that the ant *does* represent things such as 'home.'

Fortunately, the car as a button marked "Home," which, upon being pressed, will activate the car's homing vector and drive it back home. Thus, all the human has to do to accomplish her desire to get back home is to press the button. We need not think of any state of the human, or the car, or their mereological sum, as having intentional content as of the spatial properties of its terrain. Yet we can allow that there are states of the system, particularly in the human part of the system, that *are* about flamingos and home.

Here, we have a case of an intentional system (the human) interacting with a non-intentional system (the car) in such a way as to produce behavior *as if* the system as a whole is instantiating a practical syllogism. But, we need not actually attribute the practical syllogism to the car and driver in order to explain how the two work together. In fact, it seems to me, we have a better understanding of how the human and car accomplish their task if we recognize how the human's intentional system interacts with the car's non-intentional system.

The explanation is better on at least two counts. One, it explains more precisely the behavior of the system in this one instance. Simply attributing a practical syllogism to the car and driver glosses over interesting details about how the car's navigational algorithm interfaces with the driver's desire. More trenchantly, the explanation generalizes better over counterfactual behaviors of the car and driver system. Suppose that the upon seeing the flamingo, the human decides that rather than go home, she'd instead like to drive back to see the alpacas she passed on the random ride she's taken out to see the flamingo. Unless she herself, distrustful of the

car's navigational system, has been charting her position throughout the drive and has the means to take over and pilot the car manually, she'll be unable to drive back to the alpacas. Neither she nor the car have a representation of *where* they are! So, the car and human system has no general *capacity* to implement practical syllogisms over desires to go particular places in light of beliefs about where anything is. They only have a capacity to initiate the homing vector. Nonetheless, they can still satisfy a desire to get back home whenever they wish. We thus can think of the ant as being able to satisfy its desire to get home without needing to reference any intentional content as of spatial properties its states might have.

Now, you might hold a view such that states that have the ability to fulfill desires thereby, *in virtue* of that ability, *have* intentional content that matches the content of the desire. The foregoing thought experiment strikes me as a good gesture at a *reductio* of this view-- though I take it it's not decisive. Maybe the car's states do represent spatial properties in virtue of being able to satisfy the human's desire states. In any case, the story I've been telling could allow for such a view. Perhaps states of the ant's navigational system *have* intentional content as of the location of the ant's burrow in virtue of the fact that they are able to satisfy the ant's desire with the content *to go to the location of the burrow.* What I'm pressing here is that even if these states *have* such content, it doesn't seem to be *doing any explanatory work* in the context of explaining either how it gets back home, how it may satisfy its putative desire to get back home, or its counterfactually characterized capacity to do either of these things in different contexts.

But, now, suppose we fortify the ant's capacities to see when we might need to appeal to intentional content to explain its behavior. Suppose, for example, that the ant exhibited the following capacity. After being allowed to roam around the territory surrounding its nest for a while, we anaesthetize it and place it at some random location or another within the territory. Consistently, upon regaining consciousness, it's able to navigate in a straight line back to its home.

On the face of it, it seems difficult to explain this behavior in terms of the proximal stimuli impinging on the ant. After all, it would seem to be getting very different proximal stimuli from some sources, and precisely the same stimuli from others. If the ground around the territory is flat, the ant could not use stimulation of its graviceptors to guide its way home. On the other hand, whatever visual stimulation it might be receiving could be so diverse as to prohibit generalizing about its behavior in terms of that stimuli.

Suppose that on one occasion, we place the ant facing north on the east side of a rock, and on another, facing south on the west side of the rock. On both occasions, the rock will present very different visual stimuli to the ant, yet on both occasions, it will navigate successfully from the rock in a straight line to its nest. We can't make a generalization about the type of proximal visual stimulus that is allowing the ant to orient toward its home.

What would seem to be the case in each instance is that the ant is responding to some aspect of the distal stimulus-- the rock itself-- and the location of its nest relative to that rock. Crucially, it's doing so in a manner that can't be generalized

about solely in terms of proximal stimuli the rock engenders in the ant. It would seem that the rock can give rise to an indefinite number of proximal visual stimuli alone, given the ant's different perspectives on the rock. It would be difficult to characterize what all of these proximal stimuli have in common such that they could be used to individuate the same type of state in the ant.

Now, given the ant's behavior, there is still the possibility that it is relying on an easily definable class of proximal stimuli. It could be that the ant has stored in its memory a series of "snapshots" of the rock-- visual impressions from different perspectives. It could be that we can characterize these snapshots in proximal terms, along the lines of retinal intensity arrays. So, when looking at the rock from two feet north, the ant receives a particular pattern of retinal stimulation that is stored in memory. The ant would thus have a capacity for pattern matching: proximal visual stimuli that matched a pattern stored in memory would activate that stored snapshot.

Such a capacity would be akin to that exhibited by flies, *Drosophilia melanogaster* (Dill, Wolf, & Heisenberg, 1993). These flies are tethered in a flight simulator with their heads immobilized, such that their flying motion controls which visual stimulus they receive and experimenters can measure quite accurately what retinotopic stimuli the flies are responding to. During a training phase, experimenters negatively reinforce the flies with infrared heat when, during the course of their flight, the flies are looking at particular random dot textures or figures. During training, the flies change flight direction when so stimulated. During a testing phase, the flies turn away upon encountering the same random dot patterns or figures that

had occasioned heat during training.

But, the patterns provoke this response in the testing phase *only* when they are presented to the same spot on the flies' retinas as they had been during training. If the same patterns are presented a few degrees below or above the position they had originally appeared on the fly's retina, the fly will not respond. This suggests that the flies' behavior can be characterized in terms of registrations of proximal retina stimuli stored in memory.

This is the sort of retinotopic pattern matching that Burge (2010) also takes to be non-intentional:

> ...the visual systems of bees respond to retinal impresses from the shapes of landmarks, without utilizing shape constancies, or other perceptual constancies. The bee flies to a position where it receives a stored retinal impress from a landmark. Then a further stage in its navigational procedure is triggered. In such cases, even if the retinal impress from the landmark is broken in some ways, the bee's sensory system will complete the sensory template. In such cases, no visual perception need be involved. The bee's visual system, in solving this particular task, does not form a visual representational model of any aspect of its environment. It relies purely on the prototypical retinal registration of proximal stimulation. Such reliance could be counted an extraction of form. It is not visual perception (p. 419)

So, Burge seems to think such a capacity does not require appeal to intentional content, for much the same reason I've been pressing: we can make all the generalizations we want in terms of the proximal stimuli impinging on the bees, the flies, or the ants.

Suppose that the ants are similarly able to store such retinal patterns in

memory, and that along with these snapshots are stored motor programs that will take the ant in a direct line back to its home from the location at which the snapshot was taken. Perhaps, then, what the ant is doing upon being released near the rock is wandering around a bit before being stimulated in just such a way that matches one of the visual snapshots stored in its memory. At that point, the proximal visual stimulus activates the motor routine that takes the ant back to its nest. If this were the procedure the ant was following, then we could generalize its navigational abilities in terms of proximal stimuli. Whenever encountering proximal stimuli of the type stored in one of its visual snapshots, the ant will initiate the motor program stored with that snapshot.

Now, it might be objected that in this case, intentional states would be useful for generalizing the behavior of the ant. Sure, there may well be a finite number of snapshots stored in the ant's memory-- a thousand, say. But, in the absence of any characterization of what's common to these snapshots in terms of their proximal properties, it would be useful to find a property that they all have in common. That they are all snapshots *of* or *about* the rock-- albeit from different perspectives-- might be a useful property by which to type individuate them.

We could even use the attribution of states that are about the rock to generalize across the behaviors of many ants that may make use of slightly different snapshots. Suppose ant 1 has a database of one thousand snapshots of the rock, and ant two has a database of one thousand different snapshots of the rock. Nonetheless, both are able to use their snapshots to navigate back to their nest if placed next to the

rock. We can describe what's common to their behavior by noting that both of them *represented the rock* by means of some snapshot or another.

In the case of the constancy processes described above in Section 1, we could make all the generalizations we wanted in terms of the proximal characteristics of the stimuli. For example, lightness constancy was explained in terms of the abruptness of intensity changes in retinal stimuli. Under our current hypothesis, we can't characterize the snapshots in terms of similarities in their intensity gradients or other such proximal characteristics. So, in order to characterize generally the states that get the ants back home, it seems like intentional properties might be useful.

Whether intentional states really would be explanatorily useful here depends on what precisely the *competence* of the ant's navigation system consists in. For the precise operation of the ant's navigation system will determine in what ways we are able to type individuate the snapshot states, and therefore, whether individuating them in terms of intentional properties proves necessary.

Suppose, for example, that the ant's navigational capacity operates in the following manner. As the ant walks around, every half second, its visual system stores in memory a snapshot corresponding to its current retinotopic stimulation. Concurrently, its step counting system computes the motor instructions that would take it in a straight line back to its nest. These motor instructions are stored in memory with the snapshot-- in the same file, if you will. The snapshot mechanism thus works to get the ant back to its nest in the case described because, given it's motivated to get back home, its current visual proximal stimulus pulls up a snapshot

from memory that brings with it just the right motor instructions get the ant back to its nest.

So we can generalize over the circumstances that get both ants home in a variety of circumstances without positing states that are about the rock that caused the snapshots. What the snapshots have in common is that they are associated in memory with motor instructions that will take the ant back to its nest from the location the snapshot was originally taken. We can generalize over the states in terms of these properties rather than assigning intentional properties to them based upon their distal causal etiology.

All this is to say that generalizing over the ants' behavior in terms of the rock itself would be to *under*generalize. There's nothing particular about the rock itself that helps explain the ants' competence. They would behave the same way whether or not the rock happened to be where it is-- or indeed whether or not they happened to see it. We can see that an ant equipped with virtual reality goggles that stimulate its retina independently of its environment would behave in just the same manner in this context as an ant without such goggles. Each of them would walk about, taking snapshots and storing them with motor instructions generated by their step counters.

At a particular location, the snapshot generated by the ant without the goggles would be caused by some view of the rock. But, the snapshot for the ant with the goggles would be generated by the goggles themselves. In both cases, though, they will be able to return to their nests upon returning to the location because returning there will eventuate in a proximal stimulus that will recall a matching snapshot from

memory, which in turn is associated with motor instructions for getting back to the nest. Just what the particular snapshot happens to be, or its distal cause, is irrelevant.

Indeed, in order to generalize about the similarity of the behavior between the two ants-- even the one who does not receive stimulus from the rock-- we do well to avoid referencing the rock or states representing it. The more general description individuates the states in terms of their function in a computational process rather than the distal etiology of the states.

It might be objected that the better generalization in this case is to say that the ants both get home, but do so because one of them veridically represents the rock and the other *misrepresents* whatever has been displayed to it by way of the virtual reality goggles (a blade of grass, say). While such a characterization might comport well with our folk psychology, it doesn't buy us any generalizations that can't be had otherwise.

We explain the ants' locomotion by generalizing that (a) they have states of their visual, motor, and memory systems that are homomorphic to one another. That is, there is some state X, realized in ant 1 by a snapshot of the rock, and in ant 2 by a snapshot generated by its goggles. And, state X is related to states of other systems in ant 1 in a way that is homomorphic to the relations between state X and states of systems of ant 2. The generalization works in the particular environment the ants are in because (b) the computational relations described in (a) are homomorphic to the space the ants are traversing. To say that state X in ant 1 represents a rock and in ant 2 represents a blade of grass may *track* the fact that the state has a different causal

etiology and physical properties in each ant. But, the intentional ascription isn't necessary to *explain* this fact. The intentional idiom is merely a way of noting it.

You might allow that these considerations are enough to suppose that appealing to *broad* content is not necessary to explain the behavior of the ant, but nonetheless think that the foregoing account does depend on a notion of *narrow* content, along the lines that Carruthers (1987) construes it. You might suppose, that is, the ants don't represent the rock *de re*, but that the snapshots have narrow contents that serve to refer to the rock when they are embedded in particular contexts. So, for example, the ants that are hooked up to the virtual reality goggles fail to represent the rock, whereas the ants who have states actually caused by the rock represent it as such. In both cases, the ants have states with the same narrow content.

There are two things to point out here. First, notice that whatever narrow content the ants might have doesn't have correctness conditions in and of itself: token deployments of such narrow contents have correctness conditions only relative to the context of that deployment. Insofar as the notion of content I'm concerned with is the notion concerned with correctness conditions, ascribing such narrow content to the ants doesn't *per se* entail that content in the relevant sense factors into explaining the ants' behavior. Of course, we could note that each narrow content *token* would have content with correctness conditions: the ants interacting with the rock would have states about the rock, and those with the goggles would have states of goggle projections. For our purposes, we can allow that the ants may *have* such contents.

What's not clear is that these contents *play any role in explaining* the

navigational behavior of the ants. After all, the ants with correct narrow representations and incorrect narrow representations would behave the same in the same context. The difference in correctness conditions doesn't track any differences in the ants' behavior or cognition. Thus, the observation remains that we can characterize the ants' behavior quite well solely in terms of the proximal stimuli impinging upon them.

Now, were further cognitive states to influence the ant's navigational capacities, it would become increasingly difficult to describe the ant's behavior solely in terms of its proximal input. Suppose, for example, the ants with the virtual reality goggles could revise their route based upon haptic feedback as well as their retinal snapshots. Suppose the states of the ants with the virtual reality goggles were somehow defeasible in light of a large disjunction of further possible evidence. Then, we would be hard-pressed to characterize counterfactually the circumstances under which the ants' states are open to revision.

Now, in practice, a creature's states are likely only revisable relative to some finite disjunction of possible input. As Descartes made manifest long ago, in practice, even our lauded human cognitive capacities seem to give out when trying to figure out whether we're being deceived by an evil demon.[24] So, again, when teasing out

---

[24] I hazard a speculation that it's the very isotropic nature of our minds that leaves us vulnerable to Cartesian skepticism. No matter how good our evidence one way or another, we seem always able to think of new possible ways that evidence could be defeated. Desert ants don't seem to have this problem.

whether intentional content is necessary to characterize a creature's cognitive capacity, we should look at its idealized cognitive *competence* rather than its in-practice performance. The question is whether the idealized competence of a system is in-principle revisable in terms of an indefinite disjunction of input in abstraction from contingent performative constraints-- *not* whether the system is *in fact* open to an indefinite disjunction of input.

There are lots of evidence that could in-principle get the ant to revise its walk, if only it were sensitive to it. For example, if the ant knew all about humans and their nasty penchant for experimenting on desert ants, the ants might infer that when they are lifted up and placed back down again that they have been displaced by an experimenter. Thus, we might expect them not to just start walking along their homing vector as though nothing had happened. But, this is precisely what they do. After all, as a matter of performance, the cognitive architecture of an ant outside of its navigational capacities just doesn't seem sensitive to properties such as 'human experimenter displacing ants in space.' So, for all we've said here, the ant might have some *competence* to represent its location in space, but in the experimental conditions, performance constraints-- such as the inability to represent experimenters as such-- restrict it from evincing this competence in practice.

But, what's interesting about the ants is that they don't seem to have even a *competence* to revise the output of their navigational systems in light of other evidence that they *are* sensitive to. Their competence seems to be solely limited to proximal input from their step counters and polarized light detectors. They simply

don't seem to have the cognitive architecture to influence the output of their navigational system via input from other systems. Indeed, the ants in general seem quite sensitive to the position of food-- they stop their random walks upon encountering it. But, when they are transported from food they have discovered on their random walk to an area in which the food is absent, they still commence the same motor routine as they did as when the food was still present. Thus, their *navigational systems* seem insensitive to the absence of food.

This is all just to note that the navigational competence of the ant seems to be encapsulated from input other than the proximal deliverances of its step counters and polarized light detector. It's because the system is so encapsulated that we can generalize its counterfactual operations solely in terms of its proximal inputs and computational transformation of same. It's because we can make all these generalizations in terms of these proximal properties, that we need not appeal to intentional properties in order to generalize about the operations of the ant's navigation system.

Thus, if we have means for type-individuating mental states independently of intentional content, then, *prima facie*, we have no need to appeal to intentional content in our explanations. In all the cases above, we were able to type-individuate computational states either in terms of the properties of proximal stimuli that give rise to them or in terms of the computational role they play in cognition. In all such cases, once we had individuated the states in these terms, we could run all the explanations we wanted in terms of the syntactic processes relating the states to one another, and

the causal processes linking these computations to the world. There was nothing left for intentional content to explain.

All this is not to argue that modular processes, in some broad sense, are non-intentional. Indeed, as the next two Chapters will argue, there are many modular processes that may well be intentional, including that devoted to phonology. These modular processes will be un-encapsulated modules such as those postulated by Carruthers (2006), rather than the encapsulated modules of Fodor (1983).

These considerations do suggest, however, that highly encapsulated processes-- those not open to isotropic revision-- do not, as a general rule, require intentional content. Of course, general rules are defeasible, and thus there is always the possibility that intentional content may well play an explanatory role even in highly encapsulated systems. The foregoing considerations merely suggest that we would have to have quite good reason to posit such an explanatory role. It's not immediately obvious what it might be.

In general, encapsulated systems seem amenable to characterization in terms of their proximal inputs and the computational transformations performed on them. In contrast, we have at least two good reasons to appeal to intentional states in characterizing isotropic processes. Intentional content allows us to type-individuate the states these processes operate over, and allows us to characterize counterfactual generalizations made over those states. Of course, I leave it an open question as to whether there are good empirical generalizations that are only available if we characterize encapsulated systems intentionally. If so, then we'd have good reasons to

admit that intentional properties also play an explanatory role in encapsulated systems as well.

5.1 A Role for Content: State Type-Individuation

Now, suppose, alternatively, that the ant's navigation system *was* sensitive to changes such as the absence of food. Or, perhaps it's sensitive to whether the grains of sand at one position are different colors than those of another. Or, in general, suppose that the ant was able to navigate back to its nest after being displaced at arbitrary points around the nest, but did not rely upon a store of retinotopic "snapshots" to do so. In these cases, the ant would seem to have a very different cognitive competence. Indeed, the ant's navigational system would seem to be in principle open to an indefinite disjunction of possible influences. Different points of displacement would cause very different proximal stimuli. And, it would seem, there would be no way to generalize over such an indefinite disjunction solely in terms of the properties of the proximal stimuli.

An ant with such a navigational competence would be very like the honeybees studied by Menzel et al. (2005). Unlike Burge's idealized desert ant, these bees exhibit a navigational competence that is only characterizable in terms of states with intentional content. Menzel et al. demonstrate that honey bees are able to navigate back to their hive after being displaced at arbitrary locations around it.

Because the bees were displaced from the hive to a release point in the dark, they could not rely on path integration capacities. They had no information about the distance or direction of the path that took them to the release point. So, there was no

opportunity to integrate any such information into a homing vector, as with the desert ant.

Because the release points were arbitrary, Menzel et al. point out that it's fair to assume that in many cases the bees had never before flown from the point of release to the hive. Moreover, some of the bees first flew to the locations of feeders before subsequently heading back to the hive. Since all flights to the feeders originate from the hive, we can be sure that the bees had never flown from their release points to the feeders before. Thus, the retinotopic "snapshot" story also fails to explain the bees' homing behavior[25]. Since we can generally assume the bees had not been at their release points in the past, they could not have any stored snapshots from those locations that could guide their navigation.

Indeed, since the bees exhibit a capacity to return to both the hive and to feeder sites from any arbitrary release point from their range around the hive, it would seem that they can return to the hive from an indefinite disjunction of sites. Of the indefinite disjunction of possible release points around the hive, pick any one arbitrarily and a bee will be able to navigate from it back to the hive (*ceteris paribus*). Since each release point corresponds to a different set of proximal stimuli that will impinge on the bee at that location, the bees exhibit a capacity to return to the hive in response to an indefinite disjunction of proximal stimuli.

---

[25] *Pace* Burge's assertion to the contrary quoted above.

Because the bees respond to an indefinite disjunction of proximal stimuli, we cannot generalize about their homing capacity in terms of properties of the proximal stimuli that gives rise to it. As I've argued, we *can* do this in the case of the magnetotaxic bacteria, slime mold, desert ant, and early visual constancy mechanisms. In those cases, we did not need intentional content to run our explanations since we could make our generalizations in terms of properties of the proximal stimuli. But, this route isn't open to us in the case of the honey bee, so appealing to intentional contents may well be useful.

Since we cannot generalize the behavior of the bees in terms of properties of their proximal stimuli, an alternative hypothesis is that all of the bees share some other downstream mental state that in each instance causes the bee to return to the hive. Our explanation of why all the bees are able to return to the hive will therefore depend on characterizing this state.

We won't be able to characterize the state in terms of the proximal stimuli that give rise to it. After all, *ex hypothesi*, an indefinite disjunction of proximal stimuli corresponding to the various release points are capable of giving rise to the state. Since we can't type-individuate the state in terms of its proximal causes, neither can we individuate it in terms of its computational role. After all, part of its computational role is constituted by its interaction with proximal states. Insofar as different proximal states give rise to the same state in different bees, the state we're after plays a

different computational role in each bee.[26] So, our hypothesis is that the bees share a state that can be type-individuated neither by the proximal stimuli that effect it nor its computational role in the bees cognition.

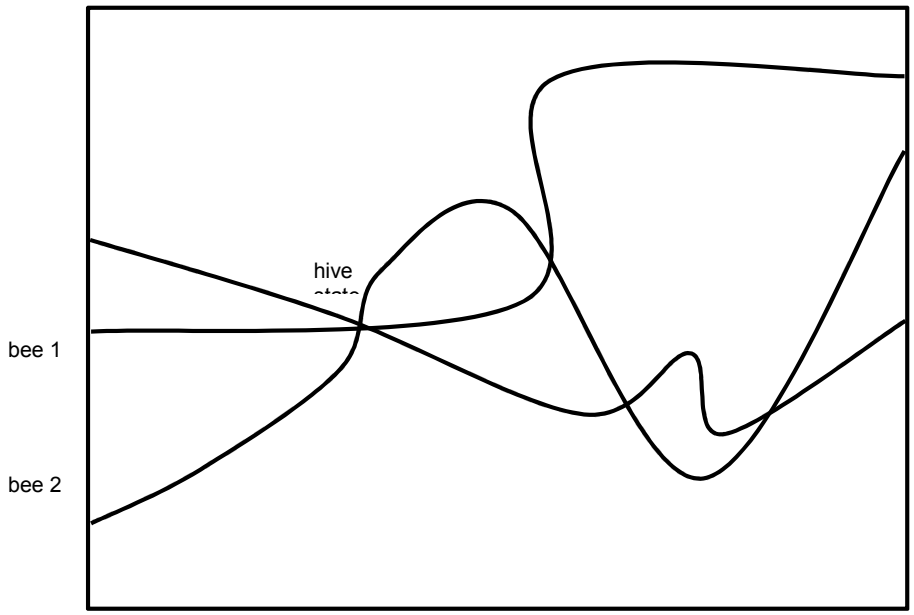
A solution is to type-individuate the state in terms of intentional content. We can attribute to all the bees a state that is *about* the hive, even though it may, on different occasions and in different bees, play a very different computational role. Thus, we can offer an intentional explanation of the bees' homing behavior. The bees are able to get back to the hive from arbitrary release points because the stimuli they receive upon being released prompts them to represent the hive and their own location relative to its location. Since all the bees receive very different proximal stimuli upon release, the causal-syntactic processes that will eventuate in them flying home may all be quite different. Yet, at some point in these syntactic processes, each bee will represent the hive, its location, and the bee's own location relative to it.

One way of visualizing such a process is to think of the computational process of individual bees as a path through an x-y coordinate space (fig. 1). Suppose that each (x,y) coordinate corresponds to a different mental state. The computational path

---

[26] This account leaves open the possibility that the state is individuated by some *particular* aspect of its computational role along lines of Rey (2009). In this case, some aspects of the state's computational role would remain constant across bees despite changes in its overall computational role. It's in such a manner that we might suppose that a 'bachelor' state is essentially one that has a computational role of being appropriately connected to 'unmarried' and 'man' states in every cognizer in which it's tokened, despite the fact that its computational role in regard to other states may be variable. That is, Mary may believe that John is a bachelor, whereas Susan may believe John is not a bachelor. In that case, both Mary and Susan's 'bachelor' states are connected to 'unmarried' and 'male' states, but have very different computational relations to their 'John' states.

of each bee's cognition through this space will likely be very different. They will start at different points because they start with different proximal input. They end with different output because the motor plan to get back to the hive will be different for each bee since they are at different locations. Since they start with different input and end with different output, they will likely take different syntactic paths from that input to their output. However, each path through cognitive space intersects at the same point. We can thus explain that this point plays a key causal role connecting the bees' different inputs with their different outputs. And, we can type-individuate this point by pointing to its intentional content as being about the bees' hive.



(fig. 1) A map of the cognitive processes of three bees, progressing from input

on the left to output on the right. Each cognitive path relies upon a particular state individuated as being about the hive.

The diagram also allows us to see why we can't individuate the same point in terms of its syntactic role: in each individual bee, the location of the point relative to other upstream and downstream states is very different for each bee. For example, in bee 1, the hive state is positioned below upstream states and above downstream states, whereas the reverse is true for bee 3. Being higher or lower on the y axis doesn't have any particular analogue to any actual feature of cognition. It's just a concrete way of pointing out that the syntactic structure in which the same state is embedded in bee 1 is radically different from the structure in which it's embedded in bee 3. Because of this, the state cannot be type-individuated in terms of its relations to the other states that make up the respective bee's cognitive processes.

In this way, intentional content can play an explanatory role type-individuating mental states that cannot be type individuated in terms of the totality of their computational roles. Intentional explanations can thus abstract away from token syntactic processes to describe mental processes shared across individuals that might instantiate very different computational routes from input to output. An intentional explanation of the bees cognition would type-individuate states that they share in terms of intentional content. An intentional explanation of their homing behavior would then recount the transitions from one point of intersection to the next, abstracting away from the computational details of how each individual bee itself arrives at each point of intersection.

Though this method of intentional explanation may at first pass seem a bit outré, I think we can see that it is a pervasive feature of many folk-psychological intentional explanations. Consider, for example, our explanation of why twice the number of people as usual visited the grocery store this Saturday.

A plausible explanation is that everyone in town believed there was a big sale at the store and many of them wanted to buy things at sale prices. But, of course, different people came to believe there was a sale at the store via quite diverse cognitive routes. Some read an advertisement in the paper, others heard about it from friends, and still others saw a lot of people driving to the store and inferred that there must be a sale going on. In each case, the proximal stimuli that prompted store-going behavior was different. Moreover the interactions with other mental states that led to the belief about the sale was different. Hearing about it from a neighbor causes one person to activate mental states concerning the trustworthiness and reliability of that particular neighbor, seeing many cars headed to the store alternatively interacted with beliefs about the number of cars usually on the road and reasons why more might be out than usual.

When we give the intentional explanation that generalizes over the behavior of all these different people, we abstract away from these particular details of individual subjects to characterize just the states common to everyone: a belief that there's a sale at the store, a belief about where the store is, and a desire to buy things on sale. So, one explanatory role for intentional content is to type-individuate mental states across subjects in which they play very different syntactic roles.

5.2 A Role for Content: Counterfactual Generalization

As we noted in Chapter 1, correctness conditions are the *sine qua non* of intentional properties. If intentional states are to do any explanatory work, these correctness conditions should play a role in explanation. Otherwise, it would seem that whatever work intentional states are invoked to do could be done by non-intentional states.

You might worry that the type-individuating role we've opened up for intentional states in this chapter won't require reference to correctness conditions. If all we need is some method of type-individuating states, then it seems as though we can do that without invoking correctness conditions. Given that we want to type-individuate a state that's engendered by indefinite proximal stimuli, we could just call it state X and be done with it.

If intentional properties type-individuate mental states, we should expect them to explain something else in virtue of their correctness conditions. So, what we want to establish is that intentional properties can help individuate mental states in such a way that the correctness conditions of those properties help explain aspects of those states' operations. As Milkowski (2013, p.154) puts it, "error should be system-detectable": whether an intentional state is correct or incorrect should make a difference to how the system itself operates, not just to how we, *qua* theorists, *interpret* its operations.

The most obvious candidate for explanation in terms of the correctness conditions of intentional states would be the interaction amongst those states. Once

we've type-individuated those states, what we need is an account of how they interact with each other. That is, the semantic properties of mental states would explain the interactions amongst them.

However, explaining the interactions amongst intentional states in terms of their semantic properties flies in the face of canonical formulations of the computational-representational theory of mind. In these formulations, the interactions amongst mental states are supposed to be explained by the syntactic properties of those states alone. That the computational theory of mind could explain how truth-preserving inferences could be effected independently of the semantic properties of their constituents was supposed to be the nicest thing about it.

The proposal now on the table would seem to be that we reverse the direction of explanation. Semantic properties of mental states will explain their syntactic operations. It's in virtue of a mental state having particular correctness conditions that it interacts with other mental states as it does.

Fortunately, we can have our cake and eat it too. There is a way in which we can still appeal to syntactic properties of mental states to explain their inferential relations, while also appealing to the semantic properties of mental states to generalize about their interactions.

The semantic properties of a state determine the counterfactual influence other semantically individuated states have upon it. Suppose, for example, you are lost in the forest. You've been travelling in a direction that you believe is north, but you come to notice that there is moss growing only on the left side of the trees. Recalling

that moss tends to grow only on the north side of trees, you revise your belief. Now you believe that you've actually been travelling east, and come to believe you must turn to your left in order to head north. In this case, the mental states eventuated by your spotting of the moss had an effect on your direction of travel because they effected a belief about the direction of north. Because this belief was about north, it was able to influence your belief about your direction of travel.

Notice that an indefinite disjunction of proximal stimuli could have influenced your belief. Had you seen the north star and believed it appeared in the northern sky, this would influence your belief. Had you found a compass with a needle pointing to the right, you would have come to believe that north was to your right. There's no way to characterize in proximal or computational terms the disjunction of stimuli that could *possibly* revise your belief in your direction of travel. It is only other beliefs *about* directions that influence your current belief about the direction you're heading.

Notice that the correctness conditions of these states determine what sort of influence that they have on your belief about your current direction of travel. If you believe north is the direction of the north star, and believe that the north star is ahead of you, your belief that you are travelling north will not change. After all, the truth conditions of your beliefs about the north star are logically consistent with your beliefs about your current direction of travel. But, if you find a compass that you believe is pointing north and to your right, these beliefs will contradict the belief that you are already travelling north. Thus, *ceteris paribus*, you'll revise your beliefs in order to make them logically consistent. So, you are liable to suppose that north is

actually to your right (or, alternatively that you've found a defective compass).[27]

Thus, attributing intentional content allows us to type individuate mental states independently of their computational role or contingent relations to the external world. The correctness conditions of such states allows us to make counterfactual generalizations about the type of states to which they are sensitive. States that are about the direction north are counterfactually open to revision in light of other states that are also about the direction north.

But, we can say all this while allowing that in each token case it is causal syntactic properties that *implement* these logical operations. It is in virtue of its syntactic properties that whatever mental state you use to represent north is causally open to revision in light of other mental states that also represent north. Given just this individual bit of cognition, we can describe it in purely syntactic terms, just as we can describe it in purely physical terms.

But, as we've pointed out, since there are indefinite syntactic routes from proximal stimuli to revisions of this state, we cannot counterfactually characterize the *type* of states to which your belief is sensitive in terms of their syntactic roles. Thus, the only way to generalize about the type of state to which your belief about north is

---

[27] Of course, just *which* beliefs you'll revise in the face of a contradiction is a subject of perennial dispute. Fodor (2000) despairs that computational psychology can give an adequate account of this problem. But Carruthers (2006) and others suppose that we can appeal to heuristic processes that determine which beliefs to revise. I remain neutral on what is the correct account of this problem. My point is that whatever process instantiates such belief revision, it must be characterized in intentional terms.

sensitive is to point to the semantic properties of those states. To put this principle concisely: states that are about north have the syntactic properties that allow them to revise other states that are about north.

This mode of intentional explanation is not just good for such folk psychological accounts. It helps explain phenomena such as the navigation capacities of the honey bees. The conclusion that Menzel et al. come to about the bees is that they not only represent the hive and its location, but that they represent the location of the hive relative to features of the surrounding landscape and feeders. In this sense, they have what Menzel et al. call a "map-like" spatial representation of the hive and landscape. The bees use their beliefs about the location of the hive relative to these landscape features to orient themselves upon being released.

Often, upon being released, bees initially fly at the same compass bearing they had been pursuing upon capture (p. 3042). It's thus plausible to attribute to the bee an initial belief that it is located just where it had been upon being captured. Once it sees a feeder and represents it as such, it must square this new representation with the antecedent belief that it is still located back where it was captured. Our story would have it that the bee changes its flight path once it revises its belief about its own location to be consistent with its belief about the location of the feeder relative to itself and the hive.

Attributing intentional states to the bees explains (1) how it is that different bees receiving different proximal stimuli can all fly back to the hive, and (2) why it is that the bees alter their flight patterns in just the way they do upon getting visual

information that's inconsistent with their beliefs about their current location.

5.3 Fodor's Alternative

Fodor (1994, pp. 51-53) paints a similar-- though, *n.b.*-- non-identical picture of the explanatory role of intentional content. The picture is different enough that I want to disavow it here, while borrowing some of its more felicitous insights. The idea is that propositional attitudes are relations to mentalese *sentences*. Intentional generalizations hold across creatures that take attitudes toward *different* mentalese sentences by characterizing the sentences as all having the same referent. So, for example, three people might variously stand in the relation of desire to the sentences: (1) 'what Lucy's eating' (2) 'the dessert your mother made for you' (3) 'Turkish delight.' Fodor talks of these sentences as "modes of presentation" of the same referent: Turkish delight. We might generalize over these three subjects by attributing to them a desire for Turkish delight. We would explain the success of all three in obtaining Turkish delight by reference to an intentional law that, e.g., *ceteris paribus*, one gets what one wants[28]. Thus, psychological laws range over the intentional contents of mental states, while abstracting from the syntactic properties of the sentences that possess those contents.

Fodor's story is problematic on two counts. First, it makes it unclear how the computational theory of mind is to be useful in explaining the implementation of such

---

[28] Set aside for the moment the plausibility of this particular psychological law.

intentional generalizations. Secondly, such intentional generalizations seem to miss out on counterfactuals that we can describe only by reference to the compositional structure of the mentalese sentences.

As Fodor constantly reminds us, the computational theory of mind is so terrific because it gives us an account of how we can implement transformational processes over mentalese sentences in a manner that systematically preserves their semantic properties. But, if psychological explanation is generally like the above case, it's hard to see why we need appeal to a process that preserves the semantic properties of mentalese *sentences.* After all, it's intentional contents that have semantic properties, and according to the above picture, intentional explanations *abstract* from mentalese sentences.

Consider an example. Person 1 wants "what Lucy's eating." Suppose he further comes to believe: "if there's any more of what Lucy's eating, it's in the fridge," and subsequently, upon Lucy telling him there's more of what she's eating, he believes: "there *is* more of what Lucy's eating." He thus deduces that he can fulfill his desire by going to the fridge.

It's clear in this case that the explanation requires an account of how the semantic value of "there's more of what Lucy's eating" can be transformed in order to instantiate the deductive syllogism. The computational theory of thought can take care of that for us, presumably. What's not at all clear is how the semantic value of "Turkish delight" comes into play. At no point does a mentalese token with content as of "Turkish delight" factor into the sentences over which computations are

performed. But, the Computational Theory of Mind is supposed to be useful because it explains how the intentional states invoked by psychological laws enter into truth-preserving relations with one another. But, according to the picture Fodor paints here, the mental states that enter into computational relations with one another (e.g., "there's more of what Lucy's eating") are distinct from those that figure in *intentional* generalizations ("Turkish delight").

Now, you might object to the above account by arguing that there is a higher, more abstract level of explanation on which we *can* give a computational account of thoughts with the content TURKISH DELIGHT. Sure, person 1 may instantiate *syntactic* states of the form "what Lucy's eating," just as person 2 instantiates syntactic states of the form "the dessert your mother made." But, insofar as they both have the same referent (Turkish delight), we should treat them as the same thought for purposes of intentional explanation. Thus, on a higher level, we can make the computational generalization that both person 1 and 2 want Turkish delight.

But, giving intentional generalizations on such abstract terms misses crucial counterfactual generalizations. Desiring, as she does, some of the dessert your mother made (*de dicto)*, upon learning that there's some Turkish delight in the fridge, person 2 won't thereby come to believe that the dessert your mother made is in the fridge. In fact, the situation is worse. Should she come to believe (contrary to fact) that your mother made the brownies on the counter, she will likely cut herself a brownie, *even if she also knows there is Turkish delight in the fridge*! Life is full of such small Sophoclean tragedies.

The lesson here is that the better explanation of person 2's behavior tracks such counterfactuals. Attributing to her a desire for Turkish delight *simpliciter* doesn't do so. A better explanation is that she revises her thoughts and behaviors to be consonant with the intentional properties of the states that compose the sentence to which her desire relates. It's because her desire is an attitude toward a sentence with contents about your mother and last night that it is sensitive to beliefs with selfsame contents, and not sensitive to beliefs about Turkish delight as such.

Though, as we have seen, Fodor is often quite sensitive to capturing counterfactuals such as the above, in the account at hand, he characterizes intentional explanation in terms of similarities in the *contingent* causal histories of different subjects:

> [T]he syntax of the mental representations which have the fact that P in their causal histories tends to overlap in ways that support robust behavioral similarities among P-believers. (p. 53)
> …
> computational-syntactic processes can implement broad-intentional ones because the world, and all the other worlds that are nomologically nearby, arranges things so that *the syntactic structure of a mode of presentation reliably carries information about its causal history* (p. 54).

Thus, besides failing to take into account counterfactuals such as those above, Fodor's account makes it unclear how psychology is supposed to deal with intentional inexistents, such as phonemes, triangles, unicorns, etc. We don't want to say that every child wants a unicorn because they all have a similar causal history as regards *unicorns*! What was so nice about Fodor's (1987) version of asymmetric dependence theory is that it allows us to type-individuate intentional contents independently of

whatever *de facto* relations a state might bear either to other mental states or to externalia. It appealed only to *counterfactual* relations to externalia. That left us free to generalize across subjects independently of their actual holistic mental states or contingent environments and histories.

Why Fodor changes his tune in (1994) is not clear to me. In any case, all I want to stress here is that the notion of intentional explanations abstracting up a level from computational explanations I'm propounding is *not* the same as that given by Fodor (1994), for good reason.

Nonetheless, despite its problems, there is quite a bit correct in what Fodor says here. In fact, we can give a gloss on the generalization over the three subjects above in a way that preserves much of his insight and skirts the problems enumerated here. We need only assume that on way to reasoning about their initial desires, each subject in fact tokens a desire toward the mentalese phrase, "Turkish delight" (which, of course, has the content TURKISH DELIGHT).

Subject 1 starts with a desire to eat what Lucy's eating and Subject 2 starts with a desire to eat the dessert that your mother made last night. Subject 1 asks Lucy what she's eating. She replies, "Turkish delight." In light of this evidence, Subject 1 comes to believe: "Lucy is eating Turkish delight." So, now he wants to eat Turkish delight. Believing that there is some Turkish delight in the fridge, he heads over there.

Meanwhile, Subject 2 is off on her quest to fulfill her desire to eat the dessert your mother made last night. She spots a plate of brownies on the counter and entertains the hypothesis that your mother made them. Then, she spots an empty

brownie package in the trash, and concludes that someone bought the brownies at the store, took them out of the package and put them on the plate.

In the trash, though, she also spots some walnut shells. She recalls your mother talking about how she hates shelling walnuts, and so only does so when she makes Turkish delight. She thus comes to believe that your mother made Turkish delight last night. So, now she wants some Turkish delight. She also believes (counter to fact) that Turkish delight melts unless refrigerated, so she reasons it must in the refrigerator. So, she heads to the fridge, where, lo, she meets Person 1.

Why did Persons 1 and 2 both go to the refrigerator? Well, because they both wanted Turkish delight and they both believed it was in the refrigerator. This explanation relies on the intentional properties of "Turkish delight" in at least two respects. First, it allows us to type-individuate states across both persons despite those states being embedded in very different computational processes in each individual. Secondly, since the state is individuated in semantic rather than syntactic terms, we can make sense of how it interacted with each subject's idiosyncratic mental states in each case. It interacted to as to make semantically valid inferences.

Maybe something like this is the picture Fodor had in mind all along. The idea is that intentional explanation abstracts from computational level explanations not by supervening on them in the usual sense, but by abstracting away from various intra-state relations that may vary from person to person, and from various interactions

they may have with the external world.[29] Thus, intentional content is explanatorily efficacious when we must generalize across subjects that have very different intra-state relations and/or are embedded in very different environments. By contrast, there is no need for intentional explanation as long as we can type-individuate computational states in terms of properties of the proximal stimuli that give rise to them and their syntactic relations to other states.

6. Conclusion

Before moving on, there is one last point we should clarify about this latter claim. The formulation above, remains ambiguous between two ways of construing the claim. On the one hand, there are generalizations we could make in terms of rather simple properties of proximal stimuli. The examples we've looked at so far have made just such generalizations. Early vision responds differentially to sharp versus gradual changes of intensity on the retina. The ant registers a discrete state for each movement of its leg. I take this Chapter to have given good reason to suppose that in all such cases, explanatory appeals to intentional content are otiose.

---

[29] How to deal with Frege cases? We can perfectly well allow that *contents* are atomistic and can't be characterized definitionally. It does not follow that propositional attitude states only relate to atomistic structures. Fodor is quite right to point out that under the CRT, attitude states relate to *sentences*, thus *ipso facto* to *non-atomistic,* structured entities. We can allow that Oedipus wants to *not* marry Oedipus' mother *qua* mother of Oedipus, but nonetheless *does* want to marry Jocasta *qua* Jocasta. The mentalese *sentence* "mother of Oedipus" has content characterized not by its *de facto referent*, Jocasta, but rather the composition of the semantics its constituent parts: "mother" and "Oedipus." Had he come to believe prior to the marriage that Jocasta was his mother, the latter desire to marry her would presumably have extinguished.

On the other hand, there could be states that are responsive to some finite, albeit very long, disjunction of very *different* proximal input. We could imagine, for instance, a creature that has 13 ways of perceiving a blackbird, all of which will cause it to engage in predatory behavior. Its predatory behavior might be occasioned, for example, by seeing black, smelling certain odors characteristic of blackbirds, hearing various blackbird related sounds, feeling downdrafts of wind often caused by flying blackbirds, etc. The creature's response to all of these proximal stimuli allows it, more often than not, to capture blackbirds and eat them.

Now, we can ask whether it's explanatorily useful to attribute intentional content as of blackbirds to the creature in order to explain its predatory success. In this case, we could, in principle, characterize counterfactually the workings of its predatory system in terms of properties of the proximal stimuli that serve as input to the system. But, it would be rather unwieldy to list out each of the 13 disjuncts that could possibly serve as antecedents to its predation behavior. A better explanation, one might think, would be one that attributes to the creature a representation with contents as of blackbirds and use that state to generalize the behavior of two creatures who attack a blackbird under different circumstances. One might have attacked because it had certain olfactory stimulation, and another because it had certain visual stimulation: in each case, we could say that they attacked because they represented a blackbird.

This may *seem* a more parsimonious explanation than one which says that they both attacked because they each received one of 13 possible proximal stimulations, namely:

1. certain olfactory stimulation
2. certain visual stimulation
3. ...etc.

Nonetheless, it's unclear that *intentional content* is doing any real explanatory work in this situation. For one, the correctness conditions of the blackbird representation don't seem to be playing a role. By hypothesis, if the creature sees black, it will pounce, regardless of any additional evidence that the sight of black is, in this instance, not caused by a blackbird. Thus, the creature's predatory system doesn't have a state that is sensitive to other states that might represent that there is *not* a blackbird around. It has no mechanism by which to make sure the correctness conditions of its states comport with one another.

Of course, we could engineer the creature such that its responses to proximal stimuli *are* defeasible in light of *some* other proximal stimuli. So, for example, it could be that the creature pounces when it sees black, unless it concurrently hears the growl of a panther-- in which case, its auditory stimuli defeats the visual stimuli and suppresses its attack behavior. This merely extends the length of the finite disjunction of stimuli to which the creature is sensitive.

Nonetheless, the competence of the system could still be described in terms of this rather long disjunction of properties of the proximal stimuli. The creature may

thus behave in many cases *as if* it represents blackbirds as such. But, *strictu dictu*, attributing *correctness conditions* to a state that results from these stimuli will not buy us any counterfactual generalizations. Even, idealizing away from performance limitations, the system is *ex hypothesi* non-responsive to stimuli that outstrip the finite disjunction. So, listing out the disjunction captures all the counterfactuals. Attributing intentional content as of blackbirds would only serve as a helpful *gloss* on this disjunction, but not buy generalizations beyond those recorded by the disjunction itself.

Nonetheless, you might suppose that this intentional gloss is explanatorily useful as it would helpfully allow us to *type-individuate* the states so caused as a common kind. Intentional attribution would thus make our explanation of the creature's behavior more parsimonious. Rather than having to type-identify its states by listing out the long disjunction of stimuli to which they are sensitive, we could just identify them as "blackbird" states, where "blackbird" serves as a compact gloss on the disjunction.

But, notice, we can achieve similar parsimony without appealing to *intentional content*. We need only coin a term, "H," say, and type identify "H-states" as just those states caused by the long, yet finite, disjunction of proximal stimuli that "H" stands in for. Thus, we can have parsimony in our *description* of the creature's operations without appealing to intentional content *per se.*

This is what we do, presumably, when we talk about "smoke detectors" (cf. Orlandi, pp. 52, 148). The detector may be responsive to a number of proximal

stimuli that are as a matter of fact associated with smoke (e.g., the presence of a light beam that may be diffracted by smoke, or a bimetallic strip that contorts in the presence of heat that may co-occur with smoke, etc.) . As a result, it tends to beep in the presence of smoke. It's simpler to say that it's a "smoke detector" rather than "a device that enters into an electronic state that causes it to beep when light usually incident on a detector no longer is so incident, or when a bimetallic strip is deformed in a particular manner, or when…etc... all of which often as a contingent matter of fact are co-incident with the presence of smoke."

But notice, we may just as well call it an "H-detector" where "H" stands in for all the proximal stimuli the smoke detector is sensitive to. To explain the operation of the smoke detector, we need just note that it's sensitive to a disjunction of proximal stimuli, H, each of which tends to covary with the presence of smoke. Attributing intentional content as of smoke to the smoke detector does not buy us any explanatory virtues.

Thus, intentional content is not on the face of things explanatorily useful in explaining the operations of systems that can be generalized over in terms of a finite disjunction of properties of their proximal input. It *is* explanatorily useful to generalize over systems that are subject (after idealizing away from performance constraints) to an indefinite disjunction of possible proximal input. It is so useful on at least two counts. First, it allows us to type-individuate the states of such systems independently of their computational role taken as a whole. Secondly, in so doing, it

allows us to capture counterfactual generalizations that we would not be able to make otherwise.

The next chapter addresses three objections to this account of the explanatory role of intentional content. First, it considers the claim that the more pertinent distinction between intentional and non-intentional processes is one between finite-state machines and read-write memory computers, rather than between encapsulated and isotropic processes. The second is the worry that my distinction between encapsulated and isotropic processes just recapitulates the old idea that the distinction between intentional and non-intentional processes maps onto the distinction between person-level and non-person-level states. Finally, I consider the objection that while my distinction might be correct *in principle*, it has no empirical bite because, as a matter of fact, there just is no distinction between encapsulated and isotropic processes operative in cognition.

Chapter 4: Architectural Objections

1. Introduction

Three objections to the picture laid out in Chapter 3 center around how it connects up with the overall cognitive architecture of the mind. The first objection concedes that I may be correct that explanation of many encapsulated processes requires no recourse to intentional states. The objection is that I've got the diagnosis wrong. It's not simply because these processes are encapsulated that precludes their intentional explanation. Rather, it's that these processes are instances of finite state computers. Following Gallistel & King (2009), you might suppose that it's computers that employ read/write memories are those we should look to for intentional explanation. That is, the difference between computational cognitive processes that involve intentional states and those that do not is best characterized as a difference between finite state processes and processes that rely on read/write memory. It is not, as I have been arguing, best characterized as the difference between proximally encapsulated processes and isotropic, abductive processes.

Not to worry, however: the differences between finite state and read/write memory machines on the one hand, and intentional versus non-intentional processes on the other, double dissociate. Finite state processes can require intentional explanation, but can sometimes do without, as can processes that make use of read/write memory. I'll argue for this conclusion in section 2 below.

The second worry also calls into question drawing the line between intentional and non-intentional processes in terms of encapsulated versus isotropic processes. The worry here is multifarious. At one level, the worry is that while I insist not to have eliminated intentional states from psychology *in principle*, I may have succeeded in doing so *de facto*. Against the background of a roughly Fodorian architecture, the distinction between non-intentional encapsulated modules and an intentional global workspace may make some sense. If, however, the architecture of the mind is massively modular, as Carruthers (2006) has argued, we might start to worry that to the extent that I'm correct, *all* cognition is in fact non-intentional.

Even if cognitive architecture is not *exhausted* by modular organization, you might further worry that insofar as my picture allows for intentionality, it burdens it with old fashioned restrictions to the effect that intentional states end up being at least *de facto* conscious, person-level, states. Suppose, for example, Carruthers' recent (2015) work is correct and that conscious, System 2, reasoning is co-extensive with the only non-modular part of the mind (a sensory-based global workspace), then you might worry that intentional states only play a role in conscious, higher level thought processes.

I could, of course, just bite the bullet and allow that given the theory of intentional efficacy given here, we would be forced to accept these conclusions if Carruthers' architectural theories are correct. I could also point out that my claim is merely that cognitive isotropy is a *sufficient* condition on intentional efficacy. I leave

it open in principle that there may be other cases in which intentional attribution is required even though the cognitive capacity is not isotropic.

Both of these responses seem rather unsatisfactory, however. The more interesting response is that we can allow for a Carruthers-style architecture *and* keep to my insistence that, generally, it is cognitive capacities that are open to isotropic revision that require intentional attribution-- all without having to *de facto* eliminate intentionality from human cognition, or accept that intentional states cannot be unconscious or part of "sub-personal" level processes. In section 3 below, I'll argue for this claim. Then, in Chapter 5, I'll give an example of a Carruthers style modular process that is nonetheless intentional: speech perception.

One last worry is that if the distinction between intentional and non-intentional explanations hinges on a distinction between isotropic and non-isotropic processes, it may turn out to be without any *practical* consequence. After all, if it turns out that people like Gary Lupyan and Andy Clark are correct, human cognition might not implement *any* bottom-up, encapsulated processes. Perhaps I am right that encapsulated processes can often be characterized non-intentionally. But, that observation, you might think, becomes uninteresting if there just aren't any encapsulated processes in the human mind as a matter of fact. In that case, my thesis amounts to showing that human cognition is intentional, end of story. That, it would seem, is old news. The response here is just to observe again that we might allow for cognition that has massive top-down influences, while still being non-isotropic. In

this way, we could well have mental capacities open to lots of top-down revision that are nonetheless non-intentional. We'll see how that can be in Section 4 below.

2. Finite State versus Read/Write Architecture

Gallistel & King (2009) propound a view that "representations" figure into cognition once it comes to rely on an addressable read/write (ARW) memory. A machine with an addressable read/write memory system, "must be capable of symbolizing its own memory locations (addresses) using encoding symbols" (p. 149). The idea is that the machine must represent that the '1' digit in the symbol, '10,' stored in memory, represents the product of the number one and the number two raised to the first power in virtue of its position to the left of the '0' digit. If further computations change the '1' symbol to a '0,' presumably the machine must represent that the '0' now occupies the same location as the '1' and therefore represents the product of the number zero and two to the first power. In this sense, ARW machines represent the syntactic relations amongst binary symbols.

Gallistel & King contrast such ARW machines with finite state architectures. Finite state architectures, they claim, do not rely on "representations" in the same sense as ARW architectures. Thus, Gallistel & King have a possible competing view as to what marks the difference between computation with and without representation: the latter occurs in finite state architectures, whereas the former occurs in ARW architectures.

At first glance, this alternate diagnosis does seem compatible with some of the cases we've discussed so far. The Bayesian account of color constancy we examined

in the previous chapter implemented a finite state architecture. That is, each subsequent state of the machine was determined entirely by the immediately preceding state. Recall, for example, that the level of water in the tubes depended entirely upon the level of water in the tubes immediately before new water input was pumped into the system. As long as the water level was as it was, the machine would transition to a new state regardless of what particular history of state transitions had caused the water levels to have the particular prior distribution they had.

Gallistel & King explain the notion of a finite state architecture as that implemented by a Turing machine that can only move its tape in one direction. It can produce outputs by writing on its tape, but the outputs are determined entirely by the current state of the machine head and whatever input it has just read off of the tape. That is, just like our Bayesian color constancy mechanism, such a finite state automaton will produce the same output whenever it is in the same state and given the same input, regardless of differences in the causal history of how it came to be in that state, receiving that input.

A fully fledged Turing machine, on the other hand has an ARW memory in the form of a tape that can move in two directions and be written upon by the machine head. Such a device has the ability to write to its tape, store what it has written, and then advance the tape so as to make use of this stored symbol in later computations. The output of such a device is not fixed entirely by its present state and current input. Given the same state and current input, what such a machine outputs

may change based upon what it has written on its tape in the causal history leading up to its current position.

Devices with such an architecture, Gallistel & King argue, makes use of representations, whereas the finite state architecture does not. Gallistel & King put the difference in representational content between these two architectures by writing that there is:

> a distinction between knowing in the symbolic sense and the "knowing" that is implicit in a stage (state) of a procedure. This is in essence the distinction between straightforward, transparent symbolic knowledge, and the indirect, opaque "knowing" that is characteristic of finite-state machines, which lack a symbolic read/write memory (p. 100).

Precisely what they mean by these two senses of "knowing" requires some explication:

> We put the state-based form of knowing in quotation marks, because it does not correspond to what is ordinarily understood by knowing. We do not place the symbolic form of knowing in quotation marks, both because it corresponds to the ordinary sense, and because we believe that this symbolic sense of knowing is the correct sense when we say such things as "the rat knows where it is" or "the bee knows the location of the nectar source"[30] (100)

Gallistel & King take these two senses of "knowing" to correspond to a nominal, anti-representational sense utilized by behaviorists on the one

---

[30] Of course, the premise of the present work is that it's not perfectly clear whether the "knowledge" of the bee or the rat is intentional.

hand, and the representational kind utilized by cognitive scientists in the post-Chomskian tradition[31]:

> The anti-representational tradition, which is seen in essentially all forms of behaviorism… regards all forms of learning as the learning of procedures… At least in its strongest form (Skinner, 1990), this line of thinking about the processes underlying behavior explicitly and emphatically rejects the assumption that there are symbols in the brain that encode experienced facts about the world (such as where things are and how long it takes food of a given kind to rot). By contrast, the assumption that there are such symbols and that they are central players in the causation of behavior is central to the what might be called mainline cognitive science (Chomsky, 1975; Fodor, 1975; Fodor & Pylyshyn, 1988; Marcus, 2001; Marr, 1982; Newell, 1980) (101)

Of course, just what this distinction amounts to and whether there is just one distinction to be made between the sense of intentional terms used by these two traditions is the central question of the present work. Traditionally, it has been assumed that cognitive accounts of behavior have differed from behaviorist accounts on at least two fronts. First, they posit internal mental states that are related computationally to one another and mediate between input and behavior. Secondly, there is the further hypothesis that these states have correctness conditions that feature in the explanation of behavior.

---

[31] Though, of course, as we've seen, Chomsky himself intends to use "knowledge" and "representation" in the nominal, non-intentional sense!

Gallistel & King have here carved out some conceptual space in the middle that corresponds to the gloss we saw Collins (2007) give to Chomsky earlier. This is the position that many cognitive processes rely on computational states that mediate input and behavior, but that these states are not "representations" in some sense of the term. Collins argues that the computational states posited by generative linguistics are not "representational" in the sense of having correctness conditions that feature in linguistic explanation.

Gallistel & King are not entirely explicit on how they take these internal computational states to be non-"representational," but I suspect they do not have quite the same distinction in mind as Collins. Recall that for Gallistel (1990), a "representation" just is a state of an organism that instantiates a functional homomorphism with respect to some feature of the world, in much the same way Cummins (1989) thinks of representations. This is a convention Gallistel & King carry forward to their present work (2009, p. 55 ff.).

The distinction between finite state and ARW memory architectures they appeal to in distinguishing between capacities that can exhibit "knowing" in some pertinent sense seems, then, to be a distinction between architectures that can instantiate such functioning homomorphisms and those that cannot. A machine with an ARW memory can hold symbols in that memory that can bear relations to one another such that there's a mapping from those symbols and relations to objects and relations holding between them in the world. On

the other hand, a finite state machine may well be able to function well in a world inhabited by such objects, but it simply won't have symbols and relations that satisfy a mapping onto the objects and relations in the world they serve to help the machine cope with.

For example, they point to a finite state automaton that instantiates a function mapping binary numerals onto the properties odd and even:

> 1 Read bit 1. If it is a '0', go to state 2. If it is a '1', go to state 3.
> 2 Read bit 2. If it is a '0', go to state 4. If it is a '1', go to state 5.
> 3 Read bit 2. If it is a '0', go to state 6. If it is a '1', go to state 7.
> 4 Read bit 3. If it is a '0', output '1'. If it is a '1', output '0'. Halt.
> 5 Read bit 3. If it is a '0', output '0'. If it is a '1', output '1'. Halt.
> 6 Read bit 3. If it is a '0', output '1'. If it is a '1', output '0'. Halt.
> 7 Read bit 3. If it is a '0', output '1'. If it is a '1', output '0'. Halt.

In this instance, they write:

> State 5 "knows" that the first bit in the input was a '0' and the second bit was a '1', not because it has symbols carrying this information but instead because the procedure would never have entered that state were that not the case. We, who are gods outside the procedure, can deduce this by scrutinizing the procedure, but the procedure does not symbolize these facts (p. 100).

So, in this finite state architecture, there is not a single symbol that co-varies (homomorphically or otherwise) with the digits of the input binary numerals. Therefore, there is no homomorphism (hence no *functional* homomorphism) between the binary digits and symbols of the algorithm that computes whether the numerals are odd or even. Thus, by Gallistel's lights, there are no "representations" of the binary digits. Perhaps Gallistel & King would concede that the *output* of this finite state automaton is "representational" insofar as the output states do homomorphically

co-vary with the evenness and oddness of the numbers symbolized by the input numerals. But, they argue that no representations (in their sense) figure in the procedure converting the input to the output.

By contrast, they point to a ARW architecture that outputs sums given the input of two numerals:

> It knows what the carry bit is because that information is carried forward by a symbol (the bit) placed at the top of the current column earlier during the computation. f+ can be in State 3 with a '0' in the carry position or a '1' in the carry position. This information is known explicitly. (100)

Here, the "carry position" can be put into different states such that those states differentially covary with the results of computations carried out earlier in the process. The state that the carry position is put into may not influence further computations until several steps after it gets set.

But, as we saw in Chapter 3, Gallistel & King's notion of "representation" just isn't the notion that this work is concerned with. In particular, we argued that mere homomorphic correlation is never sufficient to instantiate intentional states with correctness conditions that pull explanatory weight. Such states may be explanatorily efficacious-- but it is their *de facto* correlation with the environment that has

explanatory power. Any description of this correlation in terms of correctness conditions is otiose.[32]

It's for this reason that Gallistel & King's further contention that having a read/write memory architecture is a necessary and sufficient condition on implementing "representations" doesn't contradict the thesis I've been propounding here. They may well be correct that such systems necessarily rely on "representations" *as they understand them*. That is, states that can stand in homomorphic relations to things in the world, but do not necessarily have intentional content as of those things. But, I argue that some such systems *can* be understood without appeal to intentional states, in my preferred sense of states with essential correctness conditions.

For, notice that having a finite state versus ARW architecture double dissociates with the property of having states that are isotropically open to revision. We've already seen an example of an ARW architecture that I've argued is not *intentional*: the navigational capacities of our idealized desert ant. In order to navigate, the ant must store states that correlate with the number of steps it has taken since leaving its home. These states are then later accessed and transformed in accordance with a vector algebra in order to generate a motor procedure that takes the ant in a straight line back to its nest. Here, the states are stored in memory and indeed

---

[32] Moreover, homomorphisms can't account for the ability of intentional states to be about things that don't exist. If they don't exist, there's nothing for the representations to be homomorphic to!

bear homomorphic relations to geographic distances. So, they are "representational" in Gallistel & King's sense of instantiating a homomorphism. But, as argued in Chapter 3, they are not *intentional* in the sense that concerns us-- despite being states of an ARW architecture.

We can also modify the finite state Bayesian procedure used to implement color constancy in the previous chapter so as to create a finite state architecture that *does* rely on intentional content. Recall that in that procedure, particular states (which we modeled as water levels in tubes) co-varied with the probability that light waves of certain illuminances were present in the environment. The procedure was finite because the water level in the tubes at any given stage depended on what the water level in the tubes had been immediately beforehand. There was no way to siphon off water from the tubes, store it in "memory," and only later allow it to interact with the system.

We can preserve this same basic finite state architecture while opening the system up to isotropic revision simply by expanding the number of influences on the water level in the tubes. In our example, the only input that altered the water level was the deliverance of cone cell isomerization that pumped additional water into the tubes. If instead of just cone isomerization, some indefinite disjunction of inputs could possibly cause more water to be pumped into the tubes, we'd have just the sort of isotropically revisable system I've argued is a sufficient condition on requiring intentional content.

If input from just about any system could possibly cause more water to be pumped into the tubes, we'd have no way to generalize across the *type* of proximal stimuli that cause the water levels to revise. Positing that the tubes *represent* the likelihood that light of certain wavelengths is present, however, allows us to posit that the type of input stimuli that cause changes in the water level is just that of being about light waves of certain wavelengths. Notice that despite this alternation, we've preserved the finite state architecture of the constancy mechanism: the water level in the tubes is determined entirely by the level immediately before it receives input from some other system.

So, Gallistel & King's distinction between finite state and ARW architectures maps nicely onto their notion of "representation" as a functional homomorphism. ARW architectures are better able to implement such homomorphisms that finite state systems. However, the architectural distinction does not seem to have much bearing on whether a system makes use of intentional states. Both architectures can implement procedures that are open to isotropic revision and those that are not. Thus, drawing a distinction between mental processes that rely on intentional content and those that do not cannot be better drawn along the distinction between finite state and ARW architectures. My distinction between isotropic and non-isotropic processes still seems relevant.

3. Massive Modularity & Isotropy

My distinction between intentional and non-intentional computations in terms of processes that are respectively encapsulated and isotropic also elides nicely with the

Fodorian picture that the mind has a natural division between early encapsulated modules arrayed around an isotropic, Quinean global inference system. But, the distinction should hold up no matter what the actual architecture of the mind turns out to be. I stake no claim on the *correct* architecture, and would like for my account of intentional content to apply ecumenically.

What if it turns out that the mind just does not have a global isotropic belief fixation system? For example, Carruthers (2006, p. 356 ff.)) argues that the mind does not make use of isotropic belief fixation processes of the kind described by Fodor (2000). If Carruthers is right, then my argument seems to be a *de facto* case for intentional eliminativism *vis a vis* human minds.[33] I could bite this bullet, but fortunately I don't have to. For, the notion of isotropy that Carruthers attacks is just not the sort of isotropy that I've advanced as a condition on the efficacy of intentional states.

In the first place, as Carruthers is at pains to point out (p. 357), the modules that feature in his architecture are not the strongly encapsulated sort proposed by Fodor. Carruthers distinguishes between "wide-scope" and "narrow-scope" encapsulation (p.57 ff.). Narrow-scope modules are those that are counterfactually cut off from certain mental inputs: Pylyshyn-style early vision is just never open to input

---

[33] *Strictu dictu* the conjunction of Carruthers' claim and mine wouldn't entail eliminativism, even for actual human minds. For, recall, I only advance the present proposal as a *sufficient* condition on the efficacy of intentional states. But, absent a plausible additional sufficient condition, you might still worry the proposal points toward eliminativism.

from global belief states. Wide-scope modules are counterfactually, in principle, open to revision from a wide variety of sources. But, any particular operation of a module may access only a subset of the information it is in in principle privy to. Carruthers' wide-scope modules are thus in principle perfectly compatible with the sort of counterfactually openness to isotropic revision I've argued is a sufficient condition on the operation of intentional content.

Indeed, when Carruthers describes the process of belief fixation regarding whether Dusseldorf or Munich is the larger city, he seems to rely on a capacity that is isotropic in the sense germane to my argument. He proposes that such a capacity could be implemented by a "Take the Best" heuristic borrowed from Gigerenzer et al., 1999. This heuristic "searches for the piece of information concerning the two target items that has most often been found in the past to discriminate between items of that type along the required dimension" (55).

Any given application of the Take the Best heuristic may access only a small subset of beliefs. But, it does seem *capable* of accessing just about any belief in alternate applications-- and thus sensitive to an indefinite disjunction of different intentional states.

For example, it may well be that in this particular application, the heuristic led me to access my belief that having a top-division soccer team has been the most reliable guide to a city being large, which in turn caused me to search my beliefs about top-division soccer teams. But, had some other property-- say, the presence of a famous University-- been what I had judged as the most reliable guide to city size, the

heuristic would then have searched my University beliefs. In this way, the heuristic seems sensitive to representations of whatever I deem to be the most reliable guide to city size. *Prima facie*, the disjunction of such properties appears indefinite: I could come to believe that any number of properties have been most reliable guides to city size.

A module making use of the Take the Best heuristic is also open to revision in another respect. Insofar as the Take the Best heuristic is sensitive to my beliefs about soccer teams, universities, and the like, it also is going to be sensitive to all the states that influence my beliefs about soccer teams, universities, etc. My belief that Bayern Munich is in a higher division than Fortuna Dusseldorf is open to revision from a seemingly limitless number of sources-- things I read, what my friends tell me, what I see on television, etc. So, the disjunction of states to which the heuristic is sensitive becomes even more indefinite.

Thus, application of the Take the Best heuristic seems to require precisely the condition on intentional explanation set out in Chapter 3. We can characterize the mental states the heuristic is sensitive to only in terms of the intentional content of those states. Which mental states is the heuristic sensitive to? Well, an indefinite number, as individuated in terms of their computational relations to other mental states. But, we can generalize over this disjunction by simply referring to computational states that are *about* properties that are reliable guides to city size, such as the presence of a top-division soccer team. It's thus *because* one of my states is about Bayern Munich being a top-division soccer club, and another is about top-

division soccer clubs being reliable guides to city size that explains how the heuristic uses them to come up with a belief that Munich is larger than Dusseldorf.

Now as a matter of fact, as Carruthers points out, any token application of the heuristic need not access *all* of my mental states, let alone an indefinite disjunction of possible states! In any particular instance, the heuristic needs only access one particular mental state. In this sense, the heuristic is not Quinean, in that it is not sensitive to holistic properties of a set of mental states, such as coherence, simplicity, and the like. But, it is nonetheless still isotropic in the sense germane to making intentional content explanatorily efficacious. The heuristic relies on states, the constitution of which cannot be cached out in terms of their relation to other mental states. For, their relation to other mental states varies indefinitely in counterfactual situations.

If all we had to explain was this one-off instance of cognition, we could just identify this state with whatever physical substrate realizes it this once. But, if we want to explain this bit of cognition as one instance of a larger competence of implementing a heuristic, we're going to have to expand the properties that constitute the states to which its sensitive. We can't individuate these states in terms of their computational roles, because any number of states with indefinite computational roles *could* counterfactually modify the output of the procedure. The suggestion then is that we take having particular contents

This consideration is thus also a nice illustration of my suggestion we think of intentional explanation as a further abstraction in the hierarchy of cognitive

explanatory levels. Intentional explanation generalizes over multiple computational routes just as computational explanations generalize over multiple physical realizations. Given any particular application of the heuristic, we could individuate the states involved via their syntactic relations to other states used in the operation. But, the fact that the computational routes that can be used to implement the heuristic are counterfactually indefinite makes it such that we can't individuate the states employed by the heuristic in terms of their counterfactual role. When characterizing the competence of the heuristic, we can thus truly claim that the heuristic makes use of a particular state because it is *about*, say, Bayern Munich: it is just those states with such contents that the heuristic is sensitive to.

This is all just to recapitulate and sum up points made earlier. But, it goes to show that, yes, isotropic revision and hence intentional content can play a role even in heuristics implemented by Carruthers-style modules. So, even if the mind is composed entirely of such modules, there will still be room for intentional explanation. In Chapter 5 we'll examine in greater detail another case study that might be seen as being run on a Carruthers-style module that nonetheless requires intentional states for explanation: phonological perception.

For now, suffice it to say that we can grant Carruthers' claim that global cognitive capacities are not implemented by a Fodor-style global workspace subject to Quinean principles of revision that require assessment of beliefs as a corporate whole, such as coherence, simplicity, and the like. Nonetheless, the architecture he proposes to replace it, filled with modules running heuristic processes still preserves

the isotropy that I've argued is a condition on intentional explanation. So, the hypothesis I've defended should be compatible with the architecture Carruthers defends in his 2006 work.

3.1 A Sensori-Motor Workspace

Carruthers has updated his architectural theory in his most recent book (2015) so as to allow for a global isotropic workspace, which (for him) is uniquely sensory-based[34]. The picture here is that mental contents become globally available to any module that may use them when they are embedded in sensory format and enter into working memory. Carruthers' thesis is that the contents of this global workspace and the contents of conscious thought are co-extensive: a state is conscious if and only if attention is brought to bear on it, bringing it into the global workspace.[35]

Thus, according to the picture, the inferentially promiscuous states are conscious states. Insofar as I've been claiming that cognition that makes use of inferentially promiscuous states is cognition that's intentional, you might worry that I'd be committed to claiming that within Carruthers' architecture, intentional processes are co-extensive with globally available, conscious cognition.

But, as we saw in the last section, the modular systems Carruthers takes to be arrayed around and open to input from the sensori-motor workspace could well be

---

[34] Again, I prescind from endorsing any particular cognitive architecture.I consider Carruthers' latest account to demonstrate that my account of intentional content is compatible with it.
[35] I follow Carruthers in taking "conscious" to mean access-conscious.

intentional insofar as they are in principle open to revision from any state that might possibly become available to them in the global workspace. Once such a module takes up information from the global workspace, the operations it performs on it may be closed off from consciousness.

In fact, the central claim of Carruthers' recent work is that attitudinal states do not enter into the sensori-motor workspace. So, if any attitudinal psychology takes place, it must occur off in the wings within individual modules. The previous section demonstrated just how such intramodular reasoning might occur in a way that does indeed utilize the intentional contents of attitude states.

So, the massively modular architecture championed by Carruthers does not entail that all cognition is of the strongly encapsulated non-intentional sort I've hypothesized takes place in early vision. Neither does the addition of a sensori-motor global workspace to this architecture entail that the only intentional processes are conscious, System 2 operations.

4. Lupyan and Clark: Top-Down Perception

The flip side of these worries is that *all* cognition is of the isotropic sort I've associated with intentional processes, leaving no room for encapsulated non-intentional processes. Gary Lupyan (2015) argues that the sort of cognitively impenetrable early perceptual processes posited by Pylyshyn and others don't exist. Instead, early perceptual processing is often influenced by mental processes traditionally thought of as downstream of perceptual processes. If Lupyan is right, then it would seem that the distinction I've made between encapsulated non-

intentional processing and more global, isotropic intentional processing has little practical significance.

This section will argue, however, that in the first place, the evidence for Lupyan's view is by no means decisive. As Firestone & Scholl (2015) have recently argued, there are interpretations of the evidence he cites that are consistent with at least some mental processes being relatively cognitively impenetrable. Further, I'll argue that there is a reading of Lupyan's claims that is actually consistent with the distinction between isotropic and non-isotropic processes I've been drawing attention to.

4.1 Global Penetration

Lupyan's claims that perceptual processes are penetrated by higher level representations when doing so minimizes global predictions error. The claim is not that it is *because* certain episodes of penetration will so minimize prediction error that they do so. Rather, the claim seems to be simply that we should expect to see such penetration occurring when it would help minimize global prediction error: "There is no gatekeeper deciding how far down a cognitive state should penetrate perceptual processes" (6).

Lupyan first points to some of the evidence that phonological perception is penetrated from a variety of channels to exhibit his claim. In the next chapter, I expand upon this example, arguing that Lupyan is indeed correct that phonological perception is subject to isotropic revision and therefore makes use of intentional states. The present worry, though, is that *all* cognition is similarly isotropic, making

my distinction between intentional processes operating in these isotropic capacities and non-intentional processes operating in strongly encapsulated processes practically useless, even if it is an interesting theoretical distinction.

There is a deeper worry about using phonological data to demonstrate the cognitive penetrability of linguistic perception pointed out as early as Fodor (1983, p. 76ff.). Here, Fodor distinguishes between a phonological module that has top-down processes *internal* to its operations and a phonological module that is actually penetrated by belief states external to the module itself.

Fodor's example of the distinction is the phoneme restoration effect (ibid, p. 65). Replacing the utterance of "s" in an utterance of "legislature" with the sound of a cough causes subjects to perceive an utterance of "legislature" (complete with an "s" sound!) with a cough in the background. This phenomenon seems to indicate that phonological perception is subject to influence not just from acoustic input, but from our knowledge of the lexicon as well. He surmises that the system does something like construct a perception of "legi" and "lature," then goes off to find a lexical item containing those two segments. Upon finding "legislature," it constructs a perception of "legislature." Thus, lexical knowledge has top down influence on phoneme perception.

But, crucially, Fodor notes, the extent of this top down influence seems to be limited to "levels of representation that the language input system computes" (p. 77). That is, knowledge of the lexicon may influence phoneme perception, but extra-linguistic knowledge, such as that about the intentions an motivations of one's

interlocutors may not. It would only be if such extra-linguistic knowledge had a bearing on phoneme perception that we could say it's cognitively penetrable.

The distinction is perhaps even more salient in the McGurk effect (McGurk & McDonald, 1976). Here, *seeing* someone utter the syllable "ga" whilst simultaneously *hearing* an utterance of "ba" causes subjects to perceive an utterance of "da." The effect thus seems to trade on some knowledge concerning the motor and acoustic effects of phonemic utterances. Nonetheless, it is impervious to knowledge about the experimental set up. Subjects who are aware of the deception nonetheless have the same perceptual experience as naive subjects. In this sense, the effect exemplifies a process that involves intramodular top-down processing, but is not cognitively penetrated by extra-linguistic knowledge. So, any evidence of cognitive penetration ought to demonstrate not just top-down processes, but processes that cut across putative modular boundaries.

Perhaps in the spirit of providing such evidence, Lupyan points to inter*modal* influences in cognition. For example, moving your hand in front of your face can produce visual experiences (Dieter, Hu, Knill, Blake, & Tadin, 2013). Now, the mere fact that there are intermodal influences on early perceptual processing does not in itself entail that the processing is cognitively penetrated. Visual processing could take afferent proprioceptive stimuli as proximal input as well as retinal stimuli while still being closed off from the "beliefs" and "knowledge" Pylyshyn points to as top-down influence.

The observation that putative modular perceptual processes are not neatly divided in terms of sense modalities is an old one. Fodor (1983, p. 67) also makes this point! He points out that when you push your eyeball with your finger, you perceive apparent motion, even though you know nothing is moving. Thus, the visual system "has access to corollary discharges from the motor center *and to no other information that you possess.*" It's open to inter-*modular* influences, but is still encapsulated from more global belief states.

If Lupyan's point is just that more than one sense modality can provide proximal stimuli to the same perceptual processes, he's neither established that these processes are cognitively penetrated, nor has he undermined my claim that if they are cognitively impenetrable in the right way, their explanation requires no appeal to intentional contents. Previously, I've argued that we can fully generalize in purely syntactic terms the computational processes of a module that generates output that's fully fixed by proximal stimuli. Whether those stimuli come from the retina, the proprioceptive system, or both makes no difference to that argument.

To these worries point, Lupyan replies:

> It might be objected that such effects do not demonstrate [cognitive penetration], but rather reflect automatic intra-perceptual modulation. For example, Pylyshyn speculates that perhaps the putative "early vision" system takes inputs from other modalities (Pylyshyn, 1999; sect. 7.1). However, the evidence above suggests not a fixed, reflexive system where by some inexplicable quirk vision can be modulated by sound or touch, but rather a highly flexible system where these modulations can be explained as ways of lowering the overall prediction error (p. 8).

Now, if what Lupyan means by a "highly flexible system" is a system of the sort subject to isotropic revision that I've characterized as intentional, to the extent his evidence establishes that early perceptual systems are so open to revision, I'd have to concede that explanation of their operations may well require appeal to intentional content. But, it's not clear from these remarks precisely what Lupyan does mean by a "highly flexible system."

In explicating the view, he notes that "[i]mportantly, this influence of touch on vision is by no means automatic, but depends on the task-relevance of those movements (Beets et al., 2010)" (8). He doesn't explicitly explain why dependence on task-relevance is so important. But, if the early visual system is influenced by factors such a task-relevance, you might infer that it is influenced by *beliefs* and *knowledge* about task-relevance-- precisely the sorts of influence that constitute cognitive penetration.

So, how precisely does task-relevance influence visual processing? Beets et al. (2010) demonstrated that hand movements can influence the way subjects perceive an ambiguous visual stimulus-- but only if the hand movements are made in response to the stimulus itself. Subjects were presented with an array of moving dots that can be perceived as a cylindrical structure rotating either clockwise or counterclockwise.

First, subjects were asked to use the fingers of their right hands to spin a vertical rod attached to a horizontal turntable, either clockwise or counterclockwise independently of the direction they perceived the ambiguous stimulus to rotate. They indicated when their visual percept switched from counterclockwise to clockwise

rotation by pressing either a button with their left ring finger or index finger, respectively. In these trials, the direction subjects were spinning the rod had no influence on whether they perceived the dots as spinning clockwise or counterclockwise. Subjects who spun the rod clockwise saw the dots rotate counterclockwise just as much as clockwise, and *vice versa.*

However, when subjects spun the rod in response to their visual percept, the action seemed to influence the character of the percept. In one (incongruent) condition, subjects moved the rod in the direction *opposite* to that of their percept (and changed their response accordingly when their percept changed). In a second (congruent) condition, subjects moved the rod in the *same* direction as their percept. Beets et al. discovered that the durations of percepts of one direction or another were significantly shorter in the incongruent condition than the congruent condition. This suggests that when subjects in the incongruent direction rotated the rod clockwise in response to a counterclockwise percept, the movement *caused* the visual percept to flip to a clockwise percept (and *vice versa*).

Notice that this influence of motor action on visual perception occurred only when subjects were turning the rod *in response* to their percept. When subjects were asked simply to turn the rod counterclockwise independently of the direction of their percepts, the action did not succeed in attenuating the length or number of clockwise percepts. This effect is the task-relevance Lupyan cites in reference to the study. Motor activity influences vision in this case only when it has the task of responding to visual percepts.

So, what does this tell us about the cognitive penetrability of early vision? Beets et al. demonstrate that proprioceptive input must not *automatically* have an effect on visual processing. But, this fact alone does not entail that proprioceptive influence on vision is a matter of cognitive penetration. To establish an answer to that question we must figure out just *how* task-relevance modulates proprioceptive influence.

One possibility is that some aspect of "task-relevance" modulates the input sent along to early vision in the first place. That is, it is only when subjects are turning the rod in response to their visual percept that proprioceptive states of the turning are allowed to penetrate the vision module. This modulation could indeed itself be modulated by global belief-like states if the process implements a rule such as 'allow proprioception to influence vision iff you believe that your proprioceptive states are generated in response to your visual percept'. This might be cognitive penetration of a sort.

But, if this is the sort of cognitive penetration going on, it's not sufficient to undermine the *sort* of encapsulated processing I've argued is sufficiently characterized in non-intentional terms. Suppose that when proprioceptive states are not being generated in response to visual percepts, early vision is just as encapsulated as Pylyshyn assumes. In that case, as I've argued, the operations performed on the input can be described in purely syntactic, non-intentional terms. If proprioceptive states *do* serve as input to early vision when they result from responses to visual percepts, then there's no *prima facie* reason to suppose that the workings of the

module couldn't continue to be characterized in terms of the syntactic operations applied to the input states.

Here's where something like Chomsky's (2000) argument against intentional explanation *does* have some bite. He argues that visual modules don't traffic in representational states, because, after all, they could be wired up to take input from audition while preserving the syntactic characterization of the operations performed on its input states. Similarly, in this case, there's no reason in principle to think that simply changing the input of the module from visual sources to proprioceptive sources prevents us from characterizing the operations of the module itself in purely syntactic terms.[36,37]

Of course, Chomsky's argument isn't enough to banish intentional content from the mind entirely. For one, the explanation glossed here still relies on intentional belief states modulating the input to the vision module, even though we conclude that states of the vision module itself need not have intentional content. Secondly, as we noted in Chapter 1, the mere fact that the vision module can receive input from a variety of sources is not *sufficient* to establish that its states are non-intentional. The possibility pointed to here is merely that we can generalize about the activity of the

---

[36] Macpherson, in her reply to Lupyan (Ibid., p.27), raises a similar (though more outre) objection to his characterization of such inter-modal processes as "cognitive penetration" of any kind.

[37] Chomsky's argument seems to rely on assuming that intentional states must have their contents fixed via an overly externalist route. *This* line of reasoning, as argued in Chapter 1 doesn't carry much weight. But, his example is illuminating in the present case.

module in terms of the proximal stimuli to which it is sensitive-- irrespective of the causal etiology of those stimuli.

So, whatever non-automaticity and flexibility the influence of proprioception on vision exhibits might well be the result of intentional processes happening *outside* of a strongly encapsulated early vision module. Of course, Lupyan's assessment may ultimately be correct: but the experimental evidence reviewed above doesn't decide the question.

4.2 Attentional Explanations of Top-Down Effects

Other sources of evidence Lupyan cites for his claims similarly have alternate explanations in terms of downstream attentional processes. So, for example, he cites the fact that people are quicker to perceive red rectangular shaped objects when primed with the word, "brick." Of course, these phenomena could be easily explained by appealing to cognitive penetration of object recognition capacities-- largely assumed to be downstream of encapsulated, early vision. The word "brick" could prime the object recognition system, causing it to be on the lookout for states coming in from early vision that are consistent with bricks being in the vicinity. This interaction would leave the operations of early vision proper untouched.

To try and push the influence of such lexical priming further upstream, Lupyan & Ward (2013) gave subjects lexical primes while suppressing their object recognition system. They presented a steady image of, variously, a zebra or pumpkin, to one eye, while simultaneously rapidly flashing a series of images in the other eye, a procedure known as continuous flash suppression (CFS). CFS tends to suppress

conscious perception of the stable image and, crucially, also inhibits activity in the ventral visual stream associated with object recognition (Lupyan, p. 10).

Subjects were to respond either "yes" or "no" as to whether they perceived any stimulus whatsoever. Normally, the CFS made them unable to see the stable images. But, hearing the word "zebra" allowed subjects to see a zebra when they were presented with the image of a zebra. In addition, hearing the word "zebra" when an unrelated image was presented (e.g., a pumpkin), *further* suppressed their ability to consciously perceive the image.

Since the CFS protocol wiped out object recognition areas along with conscious perception, Lupyan & Ward have a good case that the priming effects are not operating on object recognition areas. So, they might conclude, the best explanation is that the word "zebra" is activating concepts that represent zebras, which then penetrate early vision and cause it to construct zebras it wouldn't otherwise without this cognitive influence.

There's an alternate interpretation, however, consonant with Carruthers' (2015) architecture, that doesn't require such cognitive penetration of early vision. We might suppose that the retinal stimulation from the zebra picture presented to one eye does indeed trigger the operation of an encapsulated early vision module that builds up something like Marr's 2 ½ -D sketch, which is then available to be taken up by downstream processes, like object recognition. If something like Carruthers' architectural picture is correct, in order to be taken up by downstream processes, the output of early vision would first have to enter into a global workspace.

On Carruthers' picture, mental states become conscious and enter the global workspace only when attentional resources are focused on them. We might therefore explain the suppression effect of CFS by theorizing that the flashing images eat up attentional resources (as changing stimuli are wont to do). So, whereas early visual processes have worked as usual on their proximal stimuli, the output of this process doesn't become conscious because attention isn't brought to bear on it. Hearing the word "zebra" causes subjects to direct attention to zebra-relevant states that are available to working memory. They therefore access their beliefs that Zebras are horse-shaped striped creatures, and thus end up focusing attention on the early visual output typical of horse-shaped striped things, bringing it into consciousness. The word "pumpkin" causes them to direct attention to pumpkin-relevant contents, thus pulling away even more attentional resources that could be used to make the zebra percept conscious, explaining why "pumpkin" further inhibits conscious perception of the zebra[38].

Yet again, we can explain how cognitive beliefs that operate outside of early vision have an effect on our conscious perceptual experience. But, we need not suppose that they do so by penetrating and altering the operations of early vision itself.

---

[38] As Carruthers notes (2015, p.69; Veillet & Carruthers 2011) this story need not require that the deliverance of early vision have intentional contents along the lines of ZEBRA or HORSE-SHAPED.

Lupyan rejects such appeals to attention because he has a view of attention at odds with the traditional "mental spotlight" approach. He instead adopts Andy Clark's (2013) conception of attention as a surprise-reducing mechanism. Conceived in this way, he argues, "attention warps neural representations throughout the brain, in a semantically coherent way...—the argument 'That's not penetrability, that's just attention' needs to be retired" (40).

But, it's unclear in practice how this view of attention would be inconsistent with the attentional explanation of the Lupyan & Ward results given above. Here's an example of the sort of attentional processes Lupyan takes to be at play in cognition. He notes that a reliable difference between images of people and images of cars is that the former tend to be vertically oriented and the latter horizontally. Therefore:

> An effective way of transforming a perceptual representation in the service of attending to people therefore might be to accentuate processing of vertically oriented features. The vague phrase "accentuating processing" corresponds in a predictive-coding framework to using the prior (pedestrians are likely to be vertically-oriented) to more accurately model the posterior distribution corresponding to locations in the scene with high-power in vertical orientations (p. 16).

He then cites Oliva & Torralba (2006) as evidence that attending to people indeed boosts processing of vertical stimuli.

But this story seems perfectly compatible with that we gave the Lupyan & Ward results. We can suppose that early vision remains impenetrable, and outputs something like a 2 ½-D sketch with states that

correlate with vertical and horizontal orientations of visual stimuli. Once this output is available in the global workspace, an intention to search for people primes a belief that people are vertically oriented, thus focusing attention on sensory contents that are consistent with vertical orientation. Whether we think of this as a process of spotlight focusing or prediction error minimization seems beside the point. The difference may well come out in other areas-- but *prima facie* it doesn't close off the sort of attentional explanation of the effect of cognitive states on conscious percepts given above.

4.3 Bottom-Up Explanations of Top-Down Effects

Firestone & Scholl (2015) point to these attentional effects on perception as one way amongst five others to defeat claims that early perception is cognitively penetrated. They make the very strong claim that " there is in fact *no* evidence for such top-down effects of cognition on visual perception" (p.8). We need not rely on a claim this strong. In order for the distinction between encapsulated and isotropic processes that I've been drawing to be interesting, we require only that at least *some* processes-- likely perceptual-- do not admit of isotropic influences. If Firestone & Scholl are only half correct, that will be enough to vindicate my claims.

As we've seen, much of the evidence for cognitive penetration is likely defeasible in light of considerations about attentional allocation. Firestone & Scholl point to other procedural errors that also might explain away much evidence of cognitive penetration. They point out that many of the experiments take to be

evidence for such penetration are perhaps best interpreted as revealing cognitive effects on judgment, or memory rather than online perception. For example, subjects who threw balls at a target reported the target to be farther away when they threw a heavy ball rather than a light ball (Witt et al., 2004). Firestone & Scholl (p. 26) point out that this effect may be driven by a change in the subjects' post-perceptual *judgment* of the distance rather than a change in the deliverances of visual *perception.* In fact, in a follow-up study (Woods et al., 2009), subjects who were asked how far away the target "visually appears" exhibited no change in their responses relative to the weight of the ball they threw. Subjects who were asked for "how far away you feel the object is, taking all nonvisual factors into account" *did* exhibit the effect. Such results suggest that subjects' beliefs about the difficulty of throwing the ball does not influence their visual perception, even if it does influence their considered judgment of the distance.

More substantively, Firestone & Scholl provide many examples of phenomena that have been given cognitive glosses, but can instead be explained in terms of bottom up operations on low-level stimuli. For example, faces with characteristically Black (i.e., "African-American") structural features appear darker than faces with characteristically White (i.e., "Caucasian") structural features, despite being matched in mean luminance (Levin & Banaji, 2006). This phenomenon is often interpreted as indicating that beliefs about race and skin color influence the deliverances of early vision. However, Firestone & Scholl (2015) blurred the faces such that subjects could no longer racially categorize them. Nonetheless, subjects still experienced one face as

darker than the other. This suggests that low-level features of the stimuli influence the perception of darkness independently of racial beliefs.[39]

Lupyan's objections to traditional sources of evidence for encapsulated perceptual systems admit of the same sorts of alternate explanations that Firestone & Scholl point to. The paradigm case here is the Muller-Lyer illusion. Because our percept of the lines remains constant despite changes in our belief states regarding the length of the lines, Pylyshyn (e.g., 1999) argues that at least some aspect of our perceptual processing must remain impenetrable to these beliefs.

The canonical explanation takes one of two forms. The first is that early vision stores hypotheses concerning the arrangement of lines to as to effect size constancy. The idea is that the angle of the arrows prompt the vision system to represent the arrowed-line as a concave edge and the divergent ends of the second line prompt early vision to present it as a protruding edge. Given the hypotheses that the two lines subtend the same retinal angle, that convex edges are closer and concave edges farther away, and that edges subtend smaller retinal angles the farther away they become, the visual system infers that the line with divergent edges is objectively larger. It thus represents it as such in the conscious percept.

---

[39] For, even though the average luminance of the two faces is identical, there are local differences in luminance between the two faces. For example, the Black face has a darker jawline and the White face has darker eyes (Firestone & Scholl, 2015)

Lupyan rightly dismisses this common explanation given the evidence that lines with circles and squares on the ends also give rise to the illusion. Instead, he picks up on data gathered by Howe & Purves (2005) that lines with "adornments" of various types closer to their center are, as a matter of contingent fact, actually shorter than lines with adornments farther from the center.

There are at least four ways these data could be used to explain the illusion. The first would posit that the hypotheses used by early vision are of the form 'lines with adornments closer to their centers are generally shorter." This of course wouldn't be evidence for cognitive penetration of early vision-- just evidence that early vision has intentional states that may well be immune to influence from doxastic states, such as the belief that in this case the line with adornments closer to the center is actually just as long as the one with adornments farther away.

A second interpretation would have it that there's standing belief in general cognition concerning the correlation of central adornment and length. When we look at the Muller-Lyer lines, this belief penetrates early vision, changing the way it processes the stimulus so that we get the illusory percept. Of course this interpretation leaves it unclear why tutored observers' standing belief that Muller-Lyer stimuli tend to have lines of equal length does not similarly penetrate early vision and reverse the illusory percept. It's also unclear why this interpretation should be favored over that above.

A third option appeals to the notion of "physical constraints" in the tradition of Marr and Pylyshyn that we examined in Chapter 3. Physical constraints just are the

contingent factual regularities such as that discovered by Howe & Purves. Recall, for example, the regularity that differences in reflectance tend to result in very sharp changes in the light reflected off of surfaces, whereas changes in illuminance tend to result in more gradual changes of same. Thus, the intensity of retinal stimulation that results from reflectance differences tends to vary sharply and that from illumination more gradedly. Thus, if the early visual system maps discrete retinal intensity changes onto one state and graded intensity changes on the other, it will end up with two states that correlate generally with changes due to reflectance and luminance, respectively. As argued in Chapter 3, we can explain the operations of this process without supposing that the visual system relies on something like a hypothesis with the content that 'reflectance differences are sharper, whereas luminance differences more gradual.'

In a similar manner, we could suppose that early vision responds differentially to edge-type stimuli with anomalies of some sort near the center and edge stimuli with anomalies farther from the center. If it did so by being sensitive to properties of proximal stimuli that correlate with "adornments," early vision could exploit the physical constraints discovered by Howe & Purves without either being cognitively penetrated or making use of intentional contents.

It's perhaps in light of these worries that Lupyan seems to embrace none of these interpretations, and instead propounds a third option, championed by Purves et al. (2011):

> Purves et al. have convincingly argued that the very concept of illusions is erroneous because calling them illusions implies the

> goal of veridical estimation of the current stimulus rather than the globally optimal estimation that perceptual systems appear to engage in. Illusions on this view can be explained as "optimal percepts" (Weiss, Simoncelli, & Adelson, 2002). (13).

On one reading, this interpretation seems to eschew appeal to intentional states of any kind. Insofar as the explanation does away with the intentional notions of *illusion* and *veridical estimation*, you might think that it's averting to some sort of non-intentional characterization of early perceptual processes altogether[40]. We'll examine this possibility more in Section 4 below. For now, suffice it to note that the interpretation leaves mysterious precisely how cognitive penetration comes into play.

Indeed, on another reading, Lupyan seems to be altogether prescinding from any stance on the precise mechanism involved in generating the Muller-Lyer percept. His positive account is that:

> insofar as the Müller-Lyer illusion arises from the visual system attempting to represent likely real-world sources… [a] bit of additional evidence in the form of training allows the system to reach a globally optimal state, making accurate local predictions while maintaining globally optimal performance. (13).

Here we are back with intentional terms ("represent," "evidence"). The training here is presumably familiarity with the correspondence of centrality of adornment and length. But, the explanation says precious little about *how* the system reaches a globally optimal state. Indeed, any of the three options canvassed above is

[40] While leaving open that illusion and veridicality might crop up outside of early perceptual processes.

consistent with his claim that what the mind is implementing in this case is a procedure that minimizes prediction error! All of them generate on the whole what seem to accurate representations of how the world is-- failing only occasionally, as with the Muller-Lyer stimulus. On this reading, Lupyan seems not so much to be offering an explanation of the Muller-Lyer illusion so much as redescribing the phenomenon in terms of prediction error[41].

4.4 Top-Down Perception as Non-Intentional

Lupyan's treatment of the Muller-Lyer illusion canvassed above raises some more fundamental concerns about ambiguities in his account of cognitive penetrability. There's an initial tension in Lupyan's characterization of what top-down cognitive penetration amounts to. He claims (p. 4) he is merely adapting Pylyshyn's (1999) characterization:

> A perceptual system is cognitively penetrable if 'the function it computes is sensitive, in a semantically coherent way, to the organism's goals and beliefs, that is, it can be altered in a way that bears some logical relation to what the person knows' (p. 343)[42]

---

[41] Inasmuch as Lupyan is concerned to show that our understanding of the Muller-Lyer illusion is *consistent* with his view that the mind is a prediction error minimizer, this redescription may be sufficient for his purposes. The stronger claim that Pylyshyn-style early vision is cognitively penetrable *might* be most charitably read as a bit of rhetorical excess.

[42] Note that this characterization leaves unresolved whether Fodor's cases of phoneme perception canvassed above count as "cognitive penetration" by Pylyshyn's lights. Whether they do or not is orthogonal to the present problem of just what *Lupyan* takes cognitive penetration to amount to.

But, he also claims that "my saying that something is 'represented' should in no way be interpreted to mean that the information is explicit or implemented in a symbolic form" (p.5, n. 5), and further, "the term 'theory' should not be taken to mean something explicit or rule-based or symbolic" (6).

On the face of it, Lupyan's characterization of "representational" states at play in instances of cognitive penetration seem to conflict with those Pylyshyn makes us of in his characterization of the phenomenon. To be sure, Pylyshyn tends to think of mental states as compositionally structured in a language-like manner. In this respect, he is clearly at odds with a view that would characterize such states in in terms that are neither rule-based nor "explicit" in the sense of having a basic atomic structure such that each atom of a complex thought represents a particular content.

But, perhaps Pylyshyn's views on the language-like structure of thought are irrelevant to his characterization of cognitive penetrability. We could allow that mental states represent in, say, an iconic manner similar to that described by Fodor (2008) such that they have no standard rule by which they decompose into constituent representations. All that's needed is that these states interact in "semantically coherent" ways that bear "logical relation to what the person knows." Iconic representations seem up to that task (in some minimal sense, at least).

But, it's unclear that Lupyan's intent here is merely to signal his neutrality between theories of language-like and iconic theories of representational *structure*. Indeed, there's some reason to think that his notion of "representation" is non-intentional in our favored sense of having correctness conditions. He writes that his

notion of representation is one "typically applied to *neural* states...it is a promiscuous term that can be applied to, e.g., the information encoded in the immune system" (p.5, n.5). Intentional properties, on the other hand, are usually ascribed to *mental* states individuated at a higher level of abstraction than the neural.

This, of course, doesn't in itself necessarily indicate that Lupyan is thinking of "representations" in non-intentional terms, but supposing that he does so also explains some other phrasing he uses throughout. For example, he writes:

> the processing in lower levels of the perceptual hierarchy is also constrained by lower level (more local) predictions that would be violated if neurons in those layers 'ignore' an object that suddenly disappeared. (p. 5)

Here, the intentional term, "ignore," is set off in scare quotes and is talking about "predictions" as states of a multi-level neural network rather than intentional processes being carried out by computations over mental states *per se*.

More intriguing is the idea that Lupyan is thinking of early perceptual processing as non-intentional in much the same way as I have been advocating-- and perhaps for some of the same reasons. He repeats approvingly the contention of Purves et al. (2011) that:

> the very concept of illusions is erroneous because calling them illusions implies the goal of veridical estimation of the current stimulus rather than the globally optimal estimation that perceptual systems appear to engage in (Lupyan, p.13).

More precisely, the claim that Purves et al. make is that:

> calling any perception of lightness or brightness a "visual illusion" is incorrect. Rather, the perceptions that arise are simply the signature of how the visual brain generates all subjective responses

> to luminance. In these terms, then, the conventional distinction between veridical and illusory perception is false; by the same token, *making inferences about the physical properties of objects or states of the world is not how vision seems to work* (p. 3. my italics)

Here, they are quite explicit that the process of optimal estimation that Lupyan champions doesn't involve intentional processes at all.

Moreover, the reason they draw this conclusion seems to be similar to the reasons I've raised to think that many such early perceptual processes are non-intentional. Their conclusion that early vision doesn't traffic in inferences about the world comes at the end of an explanation of our old standby example, reflectance constancy.

Recall that a grey patch appears lighter when presented against a darker background than when presented against a lighter background. In general, this phenomenon is attributed to the processes used to implement reflectance constancy in the visual system: the capacity we have to generally see objects has having the same reflectance properties despite being illuminated by lights of different brightness. We've already seen two accounts of such processes. One, appealed to by Burge, makes use of the gradation of the intensity of proximal retinal stimulation. The other, propounded by Brainard et al. characterizes the process as one of Bayesian hypothesis testing.

Purves et al. propose that the process involved result from an empiricist process of weighting neural nets. To test the idea, they first searched a database of natural scenes for occurrences of the sort of gray patches encountered in the illusory

stimuli above. They extracted statistics from this search so as to come up with percentile rankings of the likelihood that patches of different luminance occurred against a backgrounds of different luminances.

For example, one distribution would record that a patch of luminance P or lighter appeared against a background of luminance B**1** 80% of the time. Another distribution would record that the same patch would appear against a different background luminance, B**2**, 20% of the time. Purves et al. found that percentile rank on these scales tracks the lightness value the percept of the patch will appear to have. So, a patch of luminance P will appear much lighter against a background of luminance B1 (where it falls in the 80th percentile) than it would against a background of B2 (where it falls in the 20th percentile).

From this, Purves et al. conclude that the visual system implements reflectance constancy by extracting these probability distributions from its environment. They claim that a patch of a given luminance appears darker in some contexts and lighter in others "because the frequency of occurrence of the retinal projections generated by natural sources is different" (2). Presumably, the story is something like the following: the visual system outputs states that tend to correspond to patches of light reflectance when retinal stimulation of intensity P is surrounded by retinal stimulation of intensity B1. It outputs states that tend to correspond with patches of dark reflectance when that same retinal stimulus instead occurs surrounded by retinal stimulus of stimulus B2.

Reflectance illusions occur because the visual system rotely implements this procedure even in instances in which the states generated by early vision *don't* correlate with the actual reflectance properties of objects generating the proximal stimuli:

> It follows from this strategy of vision that calling any perception of lightness or brightness a "visual illusion" is incorrect. Rather, the perceptions that arise are simply the signature of how the visual brain generates *all* subjective responses to luminance (p. 3, italics theirs).

Their reasoning here seems to be that we should not characterize the operation of early vision in intentional terms because the system will respond to the same proximal stimuli in the same way regardless. There are no additional counterfactuals positing intentional content would help explain. That is certainly, it should be clear by now, the gloss I would give on the account.

Indeed their account of reflectance constancy differs from others we've examined, only in the etiological story about how the visual system ends up in a state that is so differentially sensitive to proximal retinal stimuli (and also, of course, precisely which aspects of the proximal stimuli it is sensitive to). Many theories, assume that such biased responses are innately hard-wired. Purves et al. argue that the visual system so responds because it rewires itself to become sensitive to the same statistics they use to derive the correlation between reflectance and background luminance:

> experience-dependent refinements of connectivity during postnatal development and adult life would allow individuals to contend with the challenge presented by the inverse problem more successfully than could be achieved on the basis of inherited circuitry alone (p. 2).

Just what evidence Purves et al. have for this strong developmental claim is unclear. Regardless, this debate about innateness in orthogonal to our concern about whether early vision is subject to cognitive penetration in a way that requires explaining its operations in terms of intentional properties. What is clear is that the account Purves et al. offer gives no reason to suppose vision is cognitively penetrated in a way that would undermine the claim that early vision is not open to isotropic revision.

If anything, their account is just another example of the sort of non-isotropic process I've been arguing doesn't require intentional explanation. This time, for what little it's worth, not only my conclusion, but my reasoning seems to come with the psychologists' explicit endorsement!

In conclusion, if the sort of explanation offered by Purves et al. is the parade case of the sort of considerations Lupyan takes to argue for his predictive coding view, there's no reason to suppose a mind implementing such a predictive coding architecture would not allow for certain processes that are best characterized in non-intentional computational terms.

If the general framework in which Lupyan is working is broadly speaking the sort of non-intentional hierarchical neural net architecture assumed by Purves et al., the claim that early perception is "cognitively

penetrable" would seem *not* to be a claim about the extent to which globally available intentional states influence early perceptual processing. For, such an architecture doesn't quantify over intentional states in the first place.

In this instance, the claim may be best read as one that neural activity commonly associated with non-perceptual processes often influences that associated more closely with perceptual processes proper. But, whether this claim challenges my assumption that at least some of the processes involved in early perception are not open to isotropic revision depends on the details of how best to characterize this influence. As we saw in sections 4.1-3 above, there are multiple ways of cashing out the interaction between high level and low level processing, some of which preserve an architecture with Pylyshyn-esque encapsulation.

Moreover, the psychologists cited by Lupyan appeal to some of the very same considerations I've pointed to in motivating the claim that early vision is best thought of in terms of non-intentional processing. In this way, Lupyan's architecture of multi-level hierarchical influence seems in principle compatible with the picture I've been painting in which at least some early perceptual processing is non-intentional because it is not embedded in a web of isotropically revisable intentional states.

The considerations I raise here against Lupyan are not intended as a devastating or comprehensive critique of his predictive coding project. They are merely to suggest that the idea that at least some mental processes are strongly

encapsulated is still a live empirical hypothesis. The success of the research programs of Scholl, Pylyshyn, and the like, in addition to the success of Marr-inspired theories of early vision processes, including Bayesian accounts such as Brainard et al.'s, evidence the fruitfulness of the hypothesis. Of course, they're open to defeat in light of further empirical investigation. But, as long as they are relevant, so too is the present proposal that such strongly encapsulated processes do not on their face require intentional contents for their explanation.

5. Conclusion

So far, my argument has had a conditional form. If intentional contents play an explanatory role in cognition, we would expect to see that role realized in cognitive systems that are open to revision from an indefinite disjunction of stimuli. By contrast, the operations systems that are not open to such revision can be explained entirely by generalizations over non-intentional computational properties of their input states.

The present chapter shows that this hypothetical proposal has some empirical bite. The distinction it points to is likely to be realized by multiple mental architectures put forth by competing theories. It is even compatible with some readings of recent theories arguing that cognition is cognitively penetrable. To the extent that it's not compatible with these theories, the data that they rely on still admits of interpretations that allow for mental processes that are variously strongly encapsulated and subject to isotropic revision.

The following chapter will provide a detailed case study of a putatively modular capacity that has been argued by some linguists not to involve intentional states: phonological perception. The framework I've argued for in the preceding chapters will allow us to see why, on the contrary, explanation of phonological perception must avert to intentional content.

Chapter 5: Intentional Content In Phonology

1. Introduction

In Chapter 3, we concluded that intentional content allows us to type-individuate states in a way that allows them to feature in counterfactual generalizations that cannot be couched in terms of states individuated solely by their purely syntactic role. We further saw that we need to rely on this method of generalization only within systems that are open to isotropic revision. It's in such systems that type-identifying states in terms of intentional content does explanatory work. Chapter 4 argued that the distinction between encapsulated, bottom-up processes and more isotropic processes is still a live one.

Chapter 4 also argued that this distinction between encapsulated and isotropic processes need not map onto the classic Fodorian divide between encapsulated modular processes on the one hand and global belief fixation processes on the other. It was a live possibility, we concluded, that there may be processes that are modular (in something like Carruthers' (2006) sense), and nonetheless open to isotropic influence. The following is an account of one such process: phonological perception.

As we've seen, Chomsky himself (2000) has famously declared that while the research program in generative linguistics appeals to "representations," the appeal is merely nominal. That is, these "representations" are not intentional in the sense of

being about things such that they can be tokened accurately or inaccurately. Collins (2004; 2007; 2008; 2009) has argued extensively for this claim in the realm of syntax and semantics, while Rey (2006; 2008) has been one of its most nuanced opponents.

This debate up until now has been largely waged over the extent to which *syntax* and *semantics* appeal to intentional states. Little attention has focused on phonology and phonetics. Collins argues against Rey that syntactic states can be type-individuated independently of any intentional content in terms of their computational roles in relation to one another. But, it's not obvious that phonological states can be similarly type-individuated. The phoneme /k/ can combine with /æt/ to form /kæt/--but /s/ can also combine with /æt/ to form /sæt/. So, it's not *obvious* how we can type-individuate /k/, /s/, and /æt/ solely on the basis of their relations to one another. The best alternative may be to individuate them in terms of their intentional content. Indeed, the following argues that this is precisely what we must do if we are to capture the generalizations of phonological science.

This argument for ascribing intentional content to phonological states as necessary to individuate them has not, to my knowledge, been raised in the literature. Rather theorists have raised other considerations that might lead them to think that phonological states must have intentional content. For example, Rey (2003) has pointed to generalizations concerning 'articulatory grounding' as examples of processes that can't be understood without attributing content to phonological states. I argue in Section 4 that if the 'substance free' phonology of Hale & Reiss (2000; 2008)

is correct, we can obviate these reasons for thinking that phonological states have intentionality.

An additional conclusion from Hale & Reiss' work, however, is that we indeed cannot individuate phonological states in terms of their internal relations to one another. They can combine so promiscuously that the very same phoneme may stand in radically different relations to others from one idiolect to another. Since we cannot individuate phonemes in terms of their computational role, we are left trying to individuate them either in terms of intentional content or alternatively their relations to external states.

Section 6, surveys the phonological perception literature to assess the plausibility of the hypotheses that we can individuate phonological states either in terms of the articulatory gestures to which they give rise or the acoustic/auditory phenomena that cause them to be tokened. If we could type-individuate a phoneme as just that state that is caused by acoustic phenomenon X, or just that state that causes articulatory gesture Y, then we would have no reason to ascribe intentional content to such states. The conclusion of Section 6 is that neither of these strategies will work, however.

I argue that the only way to sufficiently type-individuate phonological states is to ascribe intentional content to them as an essential individuating feature. This strategy not only gives us a principled way to type-individuate phonological states, but it also allows us to make phonological generalizations that would otherwise be unavailable to us.

2. Linguistic Content

Collins (2007; 2008; 2009) argues that generative linguistics gets along quite well without intentional content. That is, linguists can attribute structures to the mind such as:

[Who will [$_{VP}$ <who> beat Fred] and [$_{VP}$ be awarded <who> the prize]]

without supposing that the mind thereby *represents* things such as VPs and <who>s. The mind may well instantiate computational structures with tokens linguists type-identify as 'VP' and '<who>.' But, these syntactic tokens are not *about* anything, much less the properties of being a VP or a <who>.

Much ink spilled debating Collins' claim has dwelt on the status of external entities that could serve as the referents of putative linguistic representations. So, Chomsky (2000) sometimes argues that linguistic states could not have intentional content *because* there are no externalia that satisfy that content. In opposition, Devitt (2006a; 2006b; 2008) has argued (1) that linguistic entities, such as VPs, do exist external to our mental states, and (2) either that generative linguistics should be interpreted as generalizing over these externalia themselves, or, to the extent that linguistics generalizes over mental states, those mental states should be understood as representations of these externalia.

Rey (2008) and Collins (2009) have defended the view that referents are not necessary in order to attribute intentional content to mental states. Collins adopts a convention that intentional states have content insofar as they have correctness conditions, *which may or may not ever be actually realized*. As Collins puts it:

> For a state to have content [x] , there must be conditions C and C* such that [x] is satisfied under C and not satisfied under C*

As Rey (2008) points out, Chomsky (2000) seems to assume that the above conditions, C, must *in fact* obtain in at least some circumstances in order for states to be intentional. But, Collins and Rey both point out that we need not assume that these correctness conditions ever *in fact* obtain. Thus we can have a WITCH state that is satisfied by and only by witches-- and thus *about* witches-- without requiring that witches actually exist. Thus, for both Rey and Collins, the debate as to whether phonological mental states have intentional content is orthogonal to the debate concerning whether there are actual entities--sonic, articulatory, or otherwise-- in the world that might serve as their referents.

As Collins notes, there are some subtle worries about the extent to which something like the above criterion is correct-- particularly when it comes to cases of fictional entities, logical impossibilia, and the like. Both agree that the existence of externalia is neither necessary nor sufficient for attributing intentional content to mental states. In the following, I'll assume they are correct in order to adjudicate some more subtle points of disagreement between the two. I'm not engaging their debate as to the *existence* of mind external phonological objects. Rather, I'm asking whether the mental states posited by phonologists must have intentional content,

whether or not there are (or *metaphysically could* be) objects that actually satisfy that content.

3. Phonological Explanation

In order to see why it's necessary to attribute intentional content to phonological states, it will be useful to canvass in brief just what sort of work they are supposed to do. Standard theories of phonology start with the observation that we perceive many speech sounds as the same despite differences in their acoustic properties and the manner in which they are produced. For example, native English speakers tend to perceive the vowel sounds in:

save
made
maze

as the same as the vowels in:

safe
mate
mace

This, despite the fact that the vowels in the two groups have reliably *different* acoustic properties: vowels in the latter group have shorter duration than those in the second. The task, then, for the phonologist, is to sort out how it is that language users reliably sort a wide diversity of sounds (and in the case of signed languages, gestures) into equivalence classes.

Standard phonological theories, then, consist of rules that attempt to systematically account for which utterances get sorted into which equivalence classes. For example, a classic phonological account explains the above phenomenon by postulating a rule to the effect that:

> /eɪ/ is realized as [ĕɪ] when followed by a voiceless consonant, and as [eɪ] otherwise

where the *phones*, [ĕɪ] and [eɪ], reference the short-duration vowel of 'safe' and the long duration vowel of 'save,' respectively. The *phoneme*, /eɪ/ references the equivalence class that English speakers tend to sort these two sounds into. It's because words in the latter group have voiceless consonants that explains why we produce the phoneme /eɪ/ as [ĕɪ].

Phonemes and phones are both further characterized in terms of a feature geometry. Features, are often glossed as features of the articulatory apparatus. For example, the consonants above varied in terms of their [voice] features. [v] is [+voice] whereas [f] is [-voice], indicating that the former, but not the latter, is articulated with the vocal folds engaged. Another set of features-- [place] features-- indicates the place in the mouth at which a sound is usually articulated. The feature [+bilabial], for example, is a feature of the articulation of the phone [m]. [Manner] features, on the other hand, describe other positions of the vocal apparatus during articulation. Both [m] and [n] have the manner feature [+nasal], indicating that they are articulated with the soft palate lowered (so air can flow through the nose while the mouth is closed). In contrast with [m], however, [n] has the features [-bilabial] and [+alveolar], indicating that it is articulated with the tongue on the alveolar ridge of the

hard palate. Phonologists characterize both phones and phonemes as being composed of constituents corresponding to the above features. So, for example, /m/ and [m] can both be decomposed into [+bilabial +nasal +voice].

Now, rules such as (a) above confusingly elide a distinction between *mental states* (phonemic and phonetic states) on the one hand, and *properties* of the articulatory system or acoustic phenomena on the other. As Rey (2003, p. 172) points out, generative phonology has been mired in use-mention confusion since its inception. Chomsky (1955, as quoted in Rey, p.172) explicitly stipulates:

> We will henceforth apply the term `phones' to symbols of Pn, as well as to utterance tokens represented by them' (p. 159)

In an attempt to avoid this confusion, I stipulate that 'phoneme' and 'phone' shall refer instead solely to *mental states*-- and not to the acoustic or articulatory phenomena they may putatively represent. In this way, phonemic states, such as /el/, computationally transform into phonetic states, such as [ěl], which subsequently give rise to articulatory and auditory consequences. I'll refer to mind-external properties of the articulatory apparatus or acoustics in curly brackets, e.g.: '{ěl}.'[43]

We can think of both phones and phonemes as being composed out of components from the same set of features. Whether a particular set of features is identified as a phoneme or a phone depends on whether it's in a 'phoneme box' or a 'phone box.' That is, the phoneme box comprises sets of features that constitute

[43] c.f. Hale & Reiss (2008, pp. 107-112), who employ a similar convention.

underlying forms. The general phonological process is to take these underlying forms from the phoneme box, transform them computationally in terms of the phonological rules operating within a language (such as (a)), and put the result in the phone box. The contents of the phone box then cause articulatory gestures that result in acoustic phenomena.

So, to take our example from earlier, we might start with something like the underlying form, /selF/, in our phoneme box. The computational structure of our phonological system is such that in transferring from the phoneme box to the phone box, the feature [+short] gets added to the feature bundle, /el/. Thus, in the phone box, we end up with the surface form [sělF]. This phonetic output then transduces to the motor system to cause the motor/acoustic phenomenon, {sělF}.

Before moving on to discuss more substantive issues, it will be helpful to get a brief gloss on one of the dominant theories of how to characterize such phonological transformations: Optimality Theory (OT). Recall that phonological rules characterize how an underlying phonemic form is transformed into a phonetic surface form. The rules of OT all take the form of constraints.

Markedness Constraints constrain the form that the surface, phonetic form can take. The rule against voiced obstruents at the end of words is such a constraint. Given that the infinitive of the German 'rauben,' *to rob,* is pronounced with a {b}, we might suppose that the underlying form has as a constituent, /b/. However, the gerundive form, 'raub,' *robbery*, is instead pronounced with a voiceless {p}. Thus, a constraint against voiced final obstruents seems to be in effect. When nothing comes

after the final obstruent in the gerundive form, the phonetic output drops the [+voice] feature from the underlying phonemic form, and so the /b/ is realized as a {p}. Computationally, this is realized by transforming the underlying /b/ in the phoneme box to a [p] in the phone box, which then causes the utterance of a {p}.

Faithfulness Constraints on the other hand, constrain the extent to which the final phonetic form is allowed to diverge from the underlying phonemic form. For example, one possible constraint is that the value of the [voice] feature in the phonemic form must be preserved in the phonetic form. You might imagine such a constraint operating in English. This explains why the English words, 'rob' and 'robbery' are realized as {rɑb} and {rɑbəri}, respectively: both with a voiced {b}. This indicates that the underlying phonemic form, /b/, remains unchanged when transferred to the phone box as [b].

OT supposes that every language (I- or E-) has the same universal set of constraints. They even have constraints, like those given above, that are mutually contradictory. Each language, however, imposes a different strict ordering on the constraints. Differences in the phonetic generalizations holding within different languages (as that between German and English above) are the result of different orderings on these constraints.

The translation from underlying phonemic form to phonetic form is accomplished first by generating a list of possible phonetic form candidates, each varying in terms of their phonetic constituents. There is no mechanism for determining which forms are more possible than others: the list is supposed to be

indefinite in extent![44] Each candidate is then checked against each of the constraints in order. If a candidate form violates a constraint, it gets thrown out. The last candidate standing at the end of this process is the phonetic form generated by the underlying phonemic form. Here's a toy example:

/Raʊb/ 'raub'	*Voiced_Coda	Ident(voice)
[Raʊb]	!	
[χaʊp]	!	!
[Raʊp]		!

The phonetic form [Raʊp] managed not to violate any constraints for longer than the other two possible forms we considered, so it is generated as the phonetic form of /Raʊb/. We can explain the English data if we simply assume that in English the ordering of the constraints is reversed:

[44] This 'richness of the base' is sometimes touted by OT's proponents as a selling point of the theory- but is regarded by others as being the most psychologically implausible aspect of the theory! This dispute doesn't directly touch on our current question, so I set it aside.

/rɑb/ 'rob'	Ident(voice)	*Voiced_Coda
[rɑp]	!	
[ɬɑp]	!	
[rɑb]		!

In this case, the phonetic form of 'rob' comes out with a voiced [b], as required. It's important to keep in mind that in both grammars, the output form violates at least some of the constraints: constraints are not prohibitions on output, but merely guiding computational structures.

Note that OT suffers from the same ambiguity as to whether terms in square brackets refer to phonetic mental states or to the articulatory/auditory consequences thereof. The unfortunate ambiguity in reference between phones and mind external phenomena makes it ambiguous just which entities phonological constraints are quantifying over. A constraint against voiced obstruents could be a constraint limiting the construction of *phonetic* forms of the form [-son +voice][45], or it could be a

[45] Where '-son' characterizes obstruents.

constraint limiting the engagement of air in the vocal chords while constricting the vocal tract.

Hale & Reiss (2008, pp.149-154) demonstrate how our choice of interpretation might entail that phonemic and phonetic states have intentional content as of the articulatory gestures they cause. The next section airs these concerns and develops Hale & Reiss' reasoning that such considerations do *not* in themselves entail that phonemic or phonetic states have intentional content.

4. Phonology With and Without Substance

Given the systematic ambiguity canvassed above, there are at least two ways we could interpret OT constraints. On the one hand they could constrain possible configurations of the articulatory apparatus or acoustic consequences thereof. Hale & Reiss (2008, p. 149 ff.) call such a view *Phoneticist*. Alternatively, they could constrain the possible computational transformations between phonemic mental states and phonetic mental states. Hale & Reiss call such a view *Traditionalist* (ibid.).

The Phoneticist view of constraints would seem to entail that phonetic mental states have intentional content as of the articulatory apparatus or acoustic phenomena. Insofar as it's correct, we'd have good reason to attribute intentional content to phonological states. In what follows, we'll see that Hale & Reiss give good reason for supposing that it is *not* correct, so we need not ascribe intentional content to phonological states because of Phoneticist concerns.

The Traditionalist approach does not in itself entail that phonological states have intentional content. But, as Hale & Reiss point out, many Traditionalists suppose that constraints on phonetic forms are in some way 'grounded' by constraints on actual articulatory mechanisms. One way of making sense of this grounding relationship is to suppose that phonetic states reflect constraints on the articulatory apparatus because they *represent* aspects of that apparatus as such. While Hale & Reiss are not explicitly concerned with banishing intentionality from phonology, the conclusion of this section will be that the arguments they launch against Phoneticist and 'grounding' views exclude these considerations as sufficient to attribute intentional content to phonological states.

Let's tackle the Phoneticists first. They posit constraints *within the grammar* that operate on either articulatory systems or a speaker's acoustic production. Note that such constraints themselves are not mere properties of the articulatory system or acoustic phenomena, indicating that, for example, it's physically impossible for the back of the tongue to touch the bottom lip. They are supposed to be abstract properties of the cognitive processes that lead to speech production. For, remember, constraints are often violated. Nonetheless, such constraints are characterized in terms of articulatory production (or, alternatively, acoustic phenomena). For example, a constraint against voiced obstruents would be interpreted by the Phoneticist as a constraint against engaging the vocal chords while simultaneously obstructing the vocal tract.

As Hale & Reiss note, one way to make sense of the fact that these constraints operate on the grammar, but are individuated in terms of the articulatory system and acoustic phenomena is that the grammar has what Hale & Reiss call 'phonetic 'knowledge'' of the articulatory/acoustic consequences of implementing phonetic structures[46]. We can remove the scare quotes and just talk about phonetic states as having *intentional content* as of the articulatory system.

We could, for example, make sense of a constraint on voiced obstruents in Phoneticist terms as follows. The phonetic form [+voice] *represents* the property of the articulatory system such that air flows over the vocal chords and [-son][47] *represents* the obstruction of the vocal tract. Thus, the constraint on vocal cord engagement simultaneous with vocal tract obstruction is realized as a constraint against the phonetic mental state, [+voice -son]. This state, in turn, is characterized as a state with *intentional content as of* the relevant state of the articulatory apparatus. The Phoneticist also explains *why* constraints are operative by appeal to phonological knowledge. It's because the phonological system represents *that* it's relatively difficult to voice obstruents that it imposes a constraint against them[48]. Thus, if

---

[46] Insofar as Hale & Reiss reject the Phoneticist paradigm, they also reject this appeal to phonetic knowledge.

[47] [-son] characterizes obstruents.

[48] The pronunciation of obstruents is generally done by obstructing the flow of air through the larynx--which, given our anatomy, makes it difficult to articulate voicing by running air across the vocal folds! consolidate this explanation with explanation of -son

Phoneticist construals of phonology are correct, we may have good reason to attribute intentional content to phonological states.

Unlike the Phoneticists, Traditionalists take phonological constraints to be operating on phonetic states rather than on the articulatory system itself. Nonetheless, they often borrow the same reasoning as the Phoneticists to explain why particular constraints are operative in human phonology. They might assert that there is a constraint against voiced obstruents *because* it is relatively difficult for the articulatory apparatus to engage the vocal chords while the vocal tract is obstructed. Such accounts are appeals to *articulatory grounding*: the existence of phonological rules or constraints is explained in reference to properties of the articulatory system.

Articulatory grounding as employed by the Traditionalists is ambiguous two ways. On a diachronic construal, it may simply be an appeal to an historical-causal account of why certain constraints on human grammar were selected. It's because, in our evolutionary history, voiced obstruents were relatively difficult to pronounce that the grammar grew a constraint against the [+voice -son] forms that typically caused such articulatory processes. Such a process wouldn't necessarily require that [+voice -son] have *content* as of the articulatory actions it often produces. Alternatively, a synchronic account would have it that the grammar implements a constraint against [+voice -son] because it *represents that* voiced obstruents are difficult to articulate and it takes [+voice] to have content as of the voicing gesture and [-son] to have content as of obstruent gestures.

Thus, while there is one version of articulatory grounding that does not entail that phonological states have intentional content, there is a version of it operative in some forms of Traditionalist phonology and all forms of Phoneticist phonology that would seem to require that phonological states have content as of properties of the vocal tract.

Indeed, it's such characterizations of articulatory grounding that seem to lead Rey (2003, p.178) to conclude, *contra* Chomsky's protestations, that phonological states require content. He references Kenstowicz (1994) who notes that nasal fricatives are impossible because 'so much airflow is diverted to the nasal cavity that not enough remains to generate the turbulence required of a fricative' (p.16). Rey concludes that phonological explanations can't get on without referencing the articulatory apparatus. So, you might presume, features like [+nasal] must be features of the articulatory system itself. If features like [+nasal] are features of the articulatory apparatus, Rey argues, the only way in which they could be subject to mental computational processes is if there are representations *as of* such features that the mind computes over.

As we'll see, Hale & Reiss argue that phonology can dispense with such articulatory grounding. Instead, they argue that all phonological structures are computationally *possible*. That some are *constrained* or *unattested* can be explained without appealing to articulatory grounding. To the extent that we can do away with articulatory grounding, we can also do away with any appeal to intentionality it might bring with it, either by the Phoneticist or the synchronic Traditionalist.

4.1. Against Articulatory Grounding

Take for example the constraint against voiced obstruents considered above. An articulatory grounding of such a claim would have it that the constraint exists because the phonological system *represents that* it's relatively difficult to voice an obstruent. What if, Hale & Reiss suppose (2008, p.154 ff.), it was no longer difficult for humans to voice obstruents? Suppose humans develop expandable skin sacs below their oral cavities, much like bullfrogs. These sacs would allow air to flow over the vocal folds even as an obstruent blocked airflow to the upper vocal tract. Voicing obstruents would be quite easy.

Even if we had such necks, and our phonological system represented as much, Hale & Reiss argue, we'd still be able to acquire German, a language that, as we've seen, OT theorists have it relies on a constraint against voiced obstruents. Children raised around German speakers would continue pronouncing 'raub' as {Raʊp} even though they have no articulatory ground for not pronouncing it as {Raʊb} instead.

An intuitive commitment to articulatory grounding and a thought experiment pumping our intuitions about it does not really give us a basis for deciding on its truth either way. So, Hale & Reiss present evidence that phenomena purported to be explained by such grounding are better explained without appeal to it. For example, take the observation that, in some languages, vowels at the beginning of words tend to have the same value despite changes in the surrounding phonological context (due to, e.g., declension, surrounding words, etc.). By contrast, vowels occurring elsewhere in

the word are often subject to neutralization: they are pronounced differently as the surrounding phonological context changes.

As Hale & Reiss point out (2000, p.163), one way to account for this phenomenon would be to posit an explicit constraint *on the phonological system* that prohibits vowel neutralization at the beginning of words, but licenses it in other positions, along the lines of Beckman (1997). Since the beginnings of words tend to be more perceptually salient, such a rule would maintain distinctions between vowels in different words in just those positions that are more perceptually salient. Beckman cites this result as a point in favor of positing such a constraint.

Her reasoning seems to be that the data about perceptual saliency and neutralization is consistent with a view of the phonological system such that it represents facets of the articulatory/perceptual system. Given that the phonological system represents the beginnings of words as being more perceptually salient, that explains why it would therefore impose a constraint that limits vowel neutralization at that position.

In contrast, Hale & Reiss argue that the data about perceptual saliency actually argues *against* positing such a constraint on the phonological system itself. Instead, they argue that this psycholinguistic data itself *explains* the regularity that word initial vowels are less prone to neutralization, obviating the need to posit a constraint *on the phonological system*. It's because word initial syllables are so perceptually salient that children acquiring a language tend to maintain the contrasts between high vowels and other vowels present there. It's because non-initial syllables

are less salient that children acquiring a language might tend to gloss over distinctions in vowel values that might be present there.

Over time, then, a language (I- or E-) could well settle into a state in which distinctions between high and mid vowels at the beginning of words are stable across a wide variety of contexts, but there are few such distinctions at other positions. Patterns of vowel neutralization in a language are explained by facts of the acoustic/psycho-perceptual environment in which language acquisition takes place rather than facts about the language faculty itself.

Since their explanation requires fewer theoretical posits, Hale & Reiss argue simplicity considerations prefer it. Beckman's explanation is an instance of 'substance abuse' in that it needlessly posits that constraints on perception and production influence constraints we should attribute to the phonological system itself. The simplest hypothesis is that phonological features are allowed to combine without constraint. Any regularities in phonological structures that are never or rarely realized are the result of restrictions on the articulatory apparatus itself. If a speaker creates a phonological structure that is impossible to make manifest with her articulatory apparatus, then no other speaker will *learn* to make such a structure because her articulatory apparatus will not be able to create the data necessary for others to reconstruct the phonological structure. We need not posit that speakers *represent* that certain articulatory actions are difficult or impossible.

4.2. Against Phonological Functionalism

These considerations also give Hale & Reiss reason to dispense with another possible source of intentionality within phonology: functionalist reasoning. 'Functionalism' has a special sense here, apart from its usual use within cognitive science more generally. The basic doctrine is that the phonological system starts out with an *aim* to accomplish two goals: minimize ambiguity for the hearer while nonetheless minimizing articulatory effort for the speaker.

Given two broad goals like these, the phonological system establishes a set of constraints on phonetic forms that variously help to accomplish these goals. The view often goes hand in hand with OT. In general, faithfulness constraints, which maintain contrasts between forms at the phonemic level, help reduce ambiguity. Markedness constraints, which prohibit phonetic forms it would be difficult to pronounce, help minimize articulatory effort. Construction of phonetic forms then proceeds out of a competition amongst these constraints.

Obviously, if something like functionalism were true, the phonological system would have to represent a great deal of information: the two goal states, aspects of the articulatory system, hypotheses as to what phonological structures will best avoid ambiguity in speakers, etc. For example, the account would have it that there is a constraint against voiced obstruents because the phonological system represents

[+voice] as being about the action of engaging vocal chords, represents [-son][49] as being *about* the action of constricting the larynx, and represents that it's difficult to constrict the larynx while at the same time engaging the vocal chords!

Even more trenchantly, consider the constraint from Kirchner (1997, p. 104), referenced by Hale & Reiss (p. 184):

LAZY--Minimalize articulatory effort

To follow this constraint, it would seem the phonological system would have to *represent* as such just which articulatory gestures were more effortful than others. It might, for example, represent that a retroflex stop, articulated with the back of the tip of the tongue against the palate, requires more effort than an alveolar stop, which is articulated with the *front* tip of the tongue. In order to be sensitive to this knowledge, it would have to represent these two actions. If the phonetic state, [ʈ], represented a retroflex gesture, and the phonetic state, [d], represented a gesture that used the front tip of the tongue against the palate instead, it might then obey the constraint by outputting a [d] rather than a [ʈ].

The considerations raised above against articulatory grounding also indicate that such functionalist theory is otiose. Certain phonological structures may not

---

[49] I.e., the feature that marks obstruents.

manifest very often because they are *in fact* difficult to pronounce or perceive-- not because the phonological system *represents* them as such. There may well be pressures on speakers to reduce articulatory effort and minimize ambiguity in speech-- but these pressures need not operate via the articulatory system representing them as goals, as functionalist reasoning maintains.

Thus, Hale & Reiss eliminate many of the loci where intentional content has been thought to play a role in phonology. They show how phonological generalizations can be captured without constraints that could be construed as representational. And, they show how these rules can be implemented without supposing that they must be grounded in *knowledge* of the articulatory system or the perceptual systems of potential listeners.

4.3. Carr's Objections to Substance Free Phonology

Philip Carr (2000) seems to argue that phonological features must be intentional simply because they must transduce information to motor systems and from acoustic perceptual systems[50]. He claims that under Hale & Reiss' story, 'phonological

---

[50] Carr's claim is actually conditional. If phonological features are thought of as mental states along the lines of Hale & Reiss, then they must be intentional, he argues. Carr takes this conclusion as a *reductio* of the claim that feature states *are* computational states that are a part of UG. Rather, he champions, along with Burton-Roberts (2000) a picture of phonology as outside of linguistic cognition proper. Instead, their picture seems to be that phonological generalizations range over external E-language type phenomena rather than aspects of cognition. I lay this view aside, as it obviates the question raised in this paper: if there just *aren't* any mental phonological states, there's no question as to whether such mental states are in fact intentional! To the extent the reader is sympathetic to this view, she may take my argument to provide succor for her views insofar as they defend the antecedent of her *reductio*.

representations have intrinsic acoustic content' (97) and that 'intrinsic articulatory content also seems to find a place in Hale and Reiss'' account (98). Unless phonology is simply a formal system spinning in a void, without any connection to speech, it must have some systematic bearing on the sounds that get produced by speakers and the structures produced by hearers as the result of acoustic information reaching their ears. Given that this is the case, Carr argues that there are at least three reasons to suppose that phonological states are intentional. He's wrong on the first two counts. His third complaint can be interpreted as a recapitulation of the individuation worry we'll consider in Section 5. Ultimately, we'll show that this worry is damaging to the thesis that phonological states are nonintentional. But, we'll have to examine much more evidence than Carr adduces to demonstrate that this objection is warranted.

First, Carr notes that 'the input and output of a transducer are intimately connected, and that they stand in a non-arbitrary relation to each other' (97). Therefore, in holding that the relation between the formal process of phonology and the mechanisms of speech production and perception is one of transduction, 'Hale and Reiss undermine their view that they are not intimately connected' (97).

It's unclear how these considerations damage Hale & Reiss' account. Precisely what Carr's objection is here is unclear. It's true that transduction must be 'non-arbitrary' in the sense that the connection between the input and output must be at least somewhat systematic. Given, say, an input alphabet of numerals and a Latin output alphabet, there would be no information transduced if any numerical input had an equal chance of yielding any Latin output. Certainly, in cases of speech perception,

we must capture the generalization that certain acoustic inputs reliably give rise to the construction of the same phonological structures. We could not make sense of the observation that [+voice] is reliably tokened in hearers when they hear speakers produce sounds by engaging their vocal folds if [+voice] did not bear some systematic causal relation to such events in our normal environments.

This observation may be sufficient to establish that acoustic and motoric events are 'intimately related' to phonological states in some sense. But, this connection is not enough to establish that phonological states themselves *are about* such events or otherwise are constrained in their possible combinations with other feature states in virtue of this connection. There is a systematic connection between coat wearing and temperature in that more people wear coats the colder it becomes. But, we need not suppose that instances of coat wearing are *about* temperature in order to explain the phenomenon. Neither need we suppose that this connection constrains the conditions under which coats may be worn. It's perfectly possible to don your coat when it's hot out even though, in general, there is a systematic relation between coats being worn and cold weather.

So, yes, articulatory and acoustic phenomena may well be 'intimately connected' to phonological states, but not necessarily in any way that undermines Hale & Reiss' claims about their connection. Whether they do is a matter of just precisely the nature of their connection amounts to. For that, we must turn to more detailed accounts of the process of speech perception.

Against these considerations, Carr claims that as long as the tokening of phonological states 'depend' upon auditory perceptual states, phonological states 'have intrinsic acoustic content… they are a species of acoustic image' (97). It's just not obvious, though, that anything caused by proximal auditory stimuli must be *about* the acoustic phenomenon that caused that stimuli. Suppose, rather eccentrically, that the feature [+alveolar] gets tokened in hearers if and only if they receive proximal auditory stimulation of the kind usually caused by a pure tone of 440 Hz. This dependency of phonological state on auditory stimulus seems perfectly characterizable without appeal to intentional content. The simple statement of the causal regularity above is sufficient-- and that did not appeal to any intentional content of the [+alveolar] state.

Eventually, I will argue that to capture the actual relations between distal and proximal stimuli and phonological states, we *will* have to attribute intentional content to those states. But, to demonstrate that this is the case, we will have to establish that the relation between phonological states and the stimuli they are causally dependent upon is of a special type. The example above demonstrates that the mere existence of any sort of dependence between the two is not sufficient to require the attribution of intentional content.

Carr's final objection is that, on their face, the feature states posited within phonology *seem* to have intentional content given that they are named in terms of articulatory phenomena. Place features, such as [+alveolar] seem to be identified in terms of their relationship to articulatory structures, like the alveolar ridge of the hard

palate. Carr objects: 'if the features in question are defined in articulatory terms, then they are substantive' (98). We've already seen that it may well be difficult to type-individuate feature states in terms of their relations to one another. So, we may well have to individuate them in terms of their relationships to things such as the alveolar ridge.

But, type-identifying feature states in terms of articulatory apparati does not in itself damage Hale & Reiss' conception of SFP. Neither does it in itself entail that we must ascribe content to such states *as of* places like the alveolar ridge. It could be that when a speaker tokens [+alveolar] in her phone box, as it were, it nearly always puts into motion articulatory actions that cause the tongue to make contact with the alveolar ridge. We could thus individuate the state [+alveolar] in terms of its usual causal consequences without further supposing that the state is *about* those consequences. In the same way, we can individuate a spark plug in terms of its tendency to ignite the compressed gasoline in an engine cylinder, without supposing that the spark plug represents the cylinder.

We'll see in Section 6that such a simple story *can't* be told for phonological states. The present point is just that *contra* Carr, it is not enough to attribute intentional content to note that computational states interact with non-computational physical structures, like articulators and acoustic waves. Section 6 will explore accounts of these relations, arguing that most of them do not require intentional attribution to maintain the generalizations of speech perception theories. In so doing,

we'll see just which aspects of speech perception *do* require us to attribute intentional content to phonological states.

4.4. Against Internalist Individuation

The Hale & Reiss picture is consonant with the Chomskyan 'Galilean' program of abstracting away from variations in performance to characterize the idealized competence of the language faculty. Thus, they insist that phonology should be the study of the computationally *possible* states of the grammar, rather than simply states of the grammar that are in fact generated given further constraints imposed by the nature of language learning, the environment, vocal tract, or what have you.

In principle, then, phonological states can combine in any way: their combinatorial properties are not limited by aspects of the articulatory/perceptual systems[51]. Insofar as these two things are true, we may well be able to make phonological generalizations that don't appeal to states of production and perception systems. So, you might think, we need not characterize phonological states as *being about* sounds or articulatory gestures. In what follows, I'll argue that such a conclusion is not correct.

---

[51] Hale & Reiss go on to argue against constraints and OT more generally. I leave it a live possibility that OT theory is generally correct-- only accepting Hale & Reiss' reasons for rejecting interpretations of the theory that appeal to articulatory grounding. So, maybe the combinatorial computational processes of the phonological system are governed by constraints-- but these constraints need not be grounded in articulation.

Indeed, it is this very capacity for promiscuous and indefinitely extendable combination that requires that phonological states be type-individuated by their intentional content. The phoneme /k/ can combine with /æt/ to form /kæt/-- but /s/ can also combine with /æt/ to form /sæt/. So, it's not obvious how we can type-individuate /k/, /s/, and /æt/ solely on the basis of their relations to one another.

Now, if there *had* been restrictions on which phonemes are able to combine with other phonemes, we could use these properties to individuate them. So, for example, *suppose* there was a non-violable restriction on nasal fricatives (on the grounds, for example, that it's difficult to actually pronounce this combination)[52]. We might then be able to individuate [+nasal] in terms of its inability to combine with the class of features that constitute fricatives. This strategy would be akin to Collins' suggestion that we individuate NP states from VP states in terms of the constraints that allow the latter to check case on the former, but not *vice versa*. But, if we hew to the substance free line, this path isn't open to us in the realm of phonology.

So, it does seem that the substance free approach *prima facie* leaves us without a mechanism to type-individuate phonological states. An alternate possibility is that features are individuated in terms of simple causal connections to other states-- computational or otherwise-- *external* to phonology. So, for example, [+voice] could be individuated in terms of its tendency to be causally connected to the vibration of

---

[52] Such a restriction would be different than most OT constraints, which are, recall, violable.

the vocal folds. Or, it could be individuated in terms of certain acoustic phenomena that cause it to be tokened in perception. If we cannot find such a principle of individuation, there may be nothing for it but to individuate distinctive features by ascribing to them intentional content. The thesis of Section 5 is that there are not such principles. Thus, we attribute intentional content to feature states in order to type-individuate them. Before airing that argument, though, it will be helpful to appreciate in more detail just why type-individuation is a problem for mental states in general.

5. The Individuation Problem

This problem of individuation has played a role in the running debate about intentionality in the realm of syntax and generative semantics. A key point of contention between Collins and Rey seems to be the extent to which intentional content might be necessary to type-identify the tokens of linguistic computations. Rey (2008) suggests that, absent intentional properties, there is no principled way to individuate NPs and distinct from VPs. The worry is that any attempt to individuate linguistic states in terms of their relations to one another-- via Ramsification, say-- will fail to adequately distinguish collateral from essential properties of those states. Fodor raises a similar problem for individuating concepts.

Fodor is not worried about Ramsification as a strategy to individuate theoretical entities *per se.* Problems there may be. But, he does not find any problems that preclude, for example, individuating propositional attitude states via a

Ramsification. Beliefs may be beliefs because they interact with desires in a particular way, and *vice versa.*

Fodor's concern (1987; 1998) is that Ramsification just won't do for individuating conceptual states in particular. As Collins notes, the difference seems to be that relations amongst intentional contents are not stable enough to accommodate Ramsification. It may be just fine to Ramsify over the physicist's term, 'star,' adjusting our quantifiers as we learn more about stars and their relations to other entities, such as hydrogen, helium, and gravity. In order for this process to even get off the ground, however, the physicist must have some content, STAR, that can remain stable across multiple revisions to our ramsification of the theoretical term, 'star.'

The physicist may start out Ramsifying 'star' in relation to terms like 'heavenly canopy' and 'hole,' but after sufficient investigation, sever these connections and instead relate 'star' to 'incandescent gas' and 'nuclear fusion.' That's fine: properties we used to *think* were essential to starhood become either contingent or obviated: the stars themselves remain fixed, and we'll continue to investigate until we get things right. The physicist is able to continue to revise her Ramsification, though, only because she has a fixed target that she is after: starhood.

By contrast, insofar as the psychologist wants to characterize STAR *mental* states-- and not 'star's-- she has no stable state to fall back on as she does when investigating what makes something the theoretical term, 'star.' So, in order to individuate STAR states, she must assume they are *about* stars.

Collins (2008) criticizes the application of Fodor's (1998) argument to the current dispute as begging the question:

> Fodor's argument explicitly rests upon the presumption that psychology/linguistics is an intentional science that deals with represented contents (p. 286)

Indeed, Fodor (pp. 58-59) does argue that the relation of the psycho-linguist to theoretical terms such as AGENT is different from the physicist's relation to terms such as 'Higg's boson' precisely because AGENT tokens are *supposed* to be intentional. The physicist is free to leave the term, 'Higg's boson' undefined insofar as it's a term of the meta-language the physicist uses to make generalizations about Higg's bosons. But, when the psycholinguist attributes an AGENT state to a subject, she cannot leave the term, AGENT, undefined, because it's not merely part of her meta-language. Instead, she is using the meta-linguistic *term*, 'AGENT,' to attribute an intentional *state*, AGENT, to her subject. Thus, it i*s* encumbent upon her to give a gloss on just what it is for her subject's AGENT state to be *about* agents, as opposed to about anything else.

True, this argument presupposes that the primitives of psychological enquiry are intentional, and whether the terms of linguistic theory are intentional is just what is currently under dispute. So, as Collins notes, this particular argument from Fodor won't do much work for us. But, there's another reading of Fodor's argument that

does not so beg the question[53]. The premise here is *not* that there is some intentional state, STAR, that's about stars. Rather, the premise is that there is some mental state, X, shared both by Aristotle and a modern cosmologist. Put aside whether X has intentional content or not. In Aristotle, X is linked up to other mental states, call them canopy-states and hole-states. In the cosmologist, X is instead linked up to helium-states and fusion-states. So, if we're to preserve the assumption that X is common to Aristotle and the cosmologist, they cannot be type-individuated by their relations to other states.

There are at least two ways out of this impasse. First, we could type-individuate mental states by way of stable mind-external causal relations. For example, we might hope to identify STAR states as just those states caused by stars. This a no-go because, again, the causal responsiveness to distal stimuli of Aristotle's STAR state and mine seems just as promiscuous as their responsiveness to other mental states. Lots of non-stars could cause STAR states and sometimes stars won't eventuate in STAR states.

Alternatively, we could deny confirmation holism altogether and suppose that mental states stand in generally stable, non-promiscuous, relationships over which we can Ramsify. This is the suggestion that Collins takes up when he notes that VP-states

---

[53] The following is similar to Fodor's (1998, p. 13) circularity objection to inferential role semantics, or his (2000, pp.15-16) argument as to why rationalist psychologists can't individuate constituents of propositions in terms of their causal relations to one another. We need not accept Fodor's externalist conclusion of these arguments if we think there is some way of individuating states in terms of computational role that doesn't entail holism.

can be distinguished from NP-states in that the former check case on the latter, but not *vice versa* (p. 284). Thus, the theoretical term 'VP' is more akin to the theoretical term 'star' than it is to the mental state, STAR.[54] Insofar as the debate is restricted to the realm of syntax in matured language faculties, I'm sympathetic to Collins' arguments.

However, the considerations of Hale & Reiss raised above suggest that this strategy isn't open to phonology. The Hale & Reiss approach would have it that any phonological state can in principle computationally combine with any other. There aren't any restrictions on phonological states of the type Collins points out for syntactic states.

An alternative strategy would be to individuate phonological states in terms of their external relations to, say, the articulatory and auditory systems. It's, again, not clear that we can individuate phonological states in terms of their distal causes any better than we can do so for STAR states. The range of distal causes to which they are sensitive seems similarly indefinite. In Section 6, we'll survey evidence that both of these *prima facie* obstacles to individuating phonological states do in fact hold.

---

[54] It may seem *prima facie* that lexical items. at least, must trivially face the same difficulties Fodor proposes for concepts. Lexical items-- words-- seem to exhibit precisely the same indefinite productivity faced by concepts: they can be related to one another in seemingly endless variety. One might think, however, that the verificationist strategy may work out for words in a way it never did for concepts. For instance, you might suppose that lexical items are type-individuated by their relations not to one another but either to concepts themselves or to phonological structures. That is, it may not be characteristic of the word, 'cat' that it appear in the sentence 'the cat is on the mat.' But, it may be characteristic of it that it is causally related to the concept, CAT, or the phonological structure, /kaet/, in a way no other word is.

Thus, phonological states will be in the same boat as Fodor takes conceptual mental states to be in. Both are too promiscuously sensitive to one another and to external stimuli to be amenable to individuation by way of either of the methods canvassed above. I argue in Section 7 that the best available way to individuate phonological states, therefore, is to do so in terms of their intentional content. This is the same conclusion Fodor's reconstructed argument given above recommends for conceptual states. Given that we cannot individuate STAR states in terms of the totality of their internal or external relations, we appeal instead to their intentional content.

Now, specifying how terms *get* their meaning may well involve giving an account of their relations either to external states (as in, e.g., Fodor's externalist account) or to other intentional states (as in, e.g., Rey, 2009). Notice, though, that both these strategies entail individuating mental states in terms of their content.

Fodor's externalism gets around the individuation problem by specifying that the external relations *relevant* to individuating mental states are of a particular type. Namely, he argues that there are external conditions in which the state is tokened correctly, and those under which it is not. Insofar as states with correctness conditions just are states with intentional content, Fodor individuates mental states in terms of intentional content.

Alternatively, you could specify that particular privileged *internal* relations are relevant to determining intentional content. BACHELOR might have as its content 'unmarried man' in virtue of being related appropriately to other states

meaning ‘unmarried’ and ‘man’ (as, e.g., elaborated by Rey, 2009). But, as Rey notes, such an approach *presupposes* that mental states are already individuated in terms of their content. You can’t relate BACHELOR states to MAN states unless you have already individuated MAN states as having the content, ‘man.’

So, I’ll refrain from taking a stand on whether the appropriate theory of *content constitution* is internalist or externalist. What remains is the insight that individuating states in terms of their intentional content helps get us out of the quandary of being unable to individuate states in terms of the *totality* of their internal or external relations. This is, in Ramsey’s (2007) terms, the 'job description' of intentional content in phonology.

We have not yet, however, demonstrated that intentional content is necessary to individuate phonological states. There remain two possible individuation strategies that do not appeal to intentional content. One alternative is that phonological states are individuated in terms of the articulatory gestures that they cause. For example, a [b] might be individuated as just that state that causes the lips to move together in a certain way. Another alternative is that phonological states, such as [b], are individuated in terms of the acoustic properties that give rise to them.

Section 6 considers these two possibilities and argues that neither works. Section 6.1 argues that phonological states cannot be individuated in terms of their relations to acoustic phenomena. Section 6.2 argues that neither can they be individuated in terms of their relations to articulatory gestures. As we saw in Section 4, phonological states also cannot be individuated in terms of their relations to one

another. Therefore, we are left unable to individuate phonological states in terms of either (1) their relations to one another, (2) their relations to articulatory gestures, or (3) their relations to acoustic phenomena. Section 7 concludes that type-individuating phonological states in terms of intentional content is the best alternative to these failed strategies.

6. Individuation Via Phonological Interfaces

If we can't individuate phonological states in terms of their *de facto* internal relations with one another, you might hope that we could individuate them in terms of their relations to the proximal states that mediate their auditory and motor interfaces. It's a common phonological assumption that the states that cause us to produce speech sounds are the very same states into which we sort auditory speech stimuli. So, a /p/ state is a causal antecedent of my utterance of 'pepper' as well as a causal consequent of me hearing someone else say, 'pepper.'

This picture leaves open two strategies to individuate phonological states. One strategy would be to individuate /p/ in terms of the motor instructions to which it gives rise. In this case, /p/ would just be that state that gives rise to a set of motor instructions, such as to bring the lips together, *inter alia.* If this strategy fails, we could instead try individuating phonological states with respect to their relations to proximal auditory stimuli. On this hypothesis, /p/ would be just that state that is in fact caused by auditory stimuli of a particular type.

If either of these strategies were to work, we could describe phonological processes without attributing intentional properties to phonological states. We could,

after all, individuate the states in terms of properties of the proximal stimuli or receptors to which they connect, and, assuming Hale & Reiss are correct, describe their interactions amongst one another in purely syntactic terms.

However, these strategies *don't* work. There simply are no reliable correlations between phonological states and their interfaces with perception and production that could serve to type-individuate them. This section examines evidence from theories of speech perception that establish this claim.

Phonologists and phoneticians have never been explicitly concerned with the question of how to type-individuate phonological states, much less our larger question of whether such states have intentional content. Therefore, we must do some work to see how the dialectic within the speech perception literature maps on to our present concerns. Theories of speech perception generally fall into three major approaches.[55] Generally, they are distinguished by what they consider the proper 'objects' for speech perception. Motor Theory (MT) has it that these objects are the articulatory instructions of the speaker. Direct Realist Theory (DRT) has it that it is the actual articulatory actions themselves. General Auditory Approaches (GA) have it that the proper objects are the acoustic waves generated by articulatory actions.

Just what it means for something to be an 'object of speech perception' is unclear. Speech perception theorists certainly do not tend to write explicitly in terms

---

[55] See Diehl, Lotto, & Holt (2004) for a helpful review.

of establishing the intentional content of states. Indeed, the criteria they cite in favor of one possible set of 'objects' seems to indicate that they are often engaged in purely verbal disputes amongst one another. What theorists take it to be an 'object of perception' may well vary one from the other.

It is tempting to gloss the debate as one concerning the intentional content of phonological states. Under this construal, all theorists would agree that phonological states are intentional, but would disagree as to whether the states represent variously articulatory actions, instructions to implement those actions, or the acoustic results of those actions.

Alternatively, the debate over the objects of perception could be premised on the idea that phonological states are *not* intentional. It could be, for example, that phonological states regularly *covary* with the proximal stimuli caused by acoustic phenomena, or properties of articulatory gestures, instructions to produce the same, or auditory stimuli. In this case, the debate about objects of perception could be construed as a debate over just which correlations with external phenomena are appropriate to individuate states *independently* of any content. We'll see that all of these non-intentional strategies fail.

6.1. Motor Theory: Against Auditory Individuation

In explicating their Motor Theory, Liberman & Mattingly (1985) argue that *because* phonological states can't be individuated in terms of the proximal auditory stimuli that cause them, they must be thought of as states with intentional content as of motor instructions. As we'll see, the antecedent in itself doesn't warrant the consequent.

But, the considerations of Motor Theory *are* sufficient to eliminate one of our candidate hypotheses about non-intentional individuation: we cannot individuate phonological states in terms of their mere causal co-variation with proximal auditory stimuli that cause them.

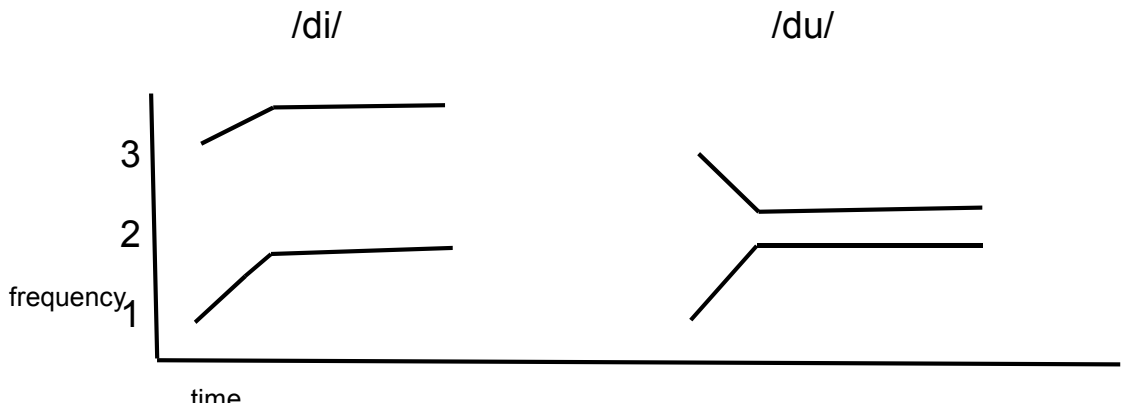
Motor Theory has it that:

> phonetic objects cannot be perceived as a class by reference to acoustic stigmata, but only by a r*ecognition that* the sounds might have been produced by a vocal tract as it made linguistically significant gestures. (Liberman & Mattingly, 1985, p. 24, my italics)

Their general reasoning seems to be that (1) there is no 1-1 mapping between either acoustic properties or the proximal stimuli resulting from them, on the one hand, and phonological states on the other. Because of this lack, (2) which phonological state the language faculty tokens as the result of proximal auditory stimuli is the result of it assessing how much evidence the proximal stimuli provide for the one state over another given its background representations of the vocal tract and its acoustic consequences. Considered as such, the theory is clearly intentional. However, the inference does not go through.

To see why, let's first get a grip on the task the auditory-cum-speech perceptual system must perform. Phoneticians often characterize sounds in terms of different frequencies, the intensity of which changes overtime. When these changes are mapped on a spectrograph, they appear as bands called formants. Each formant tracks an area of relatively similar sound intensity that may change frequencies over

time. For example, the spectrograph for an utterance of /di/ and that of /du/ looks something like this:



Formants are labeled numerically in order of ascending relative frequency as F1, F2, etc. So, here, the acoustic consequences of /du/ are such that F1 rises in frequency as F2 lowers.

These acoustic properties cause proximal stimulation of the auditory system. Somehow the speech perception system must construct phoneme representations given states that track the information given in such spectrographs. Liberman & Mattingly are correct to point out that the information recorded by such spectrographs does not match 1-1 with the phonemes that they give rise to. As seen above, two very different acoustic events can each give rise to the perception of a /d/.

However, this observation alone is not enough to establish that the process of mapping proximal stimuli engendered by such acoustic phenomena to phonemic representations is an intentional one. Instead, there could be regularities in the

mapping between acoustic cues and phonemic states that the perceptual system could exploit.

For instance, take an example from Liberman & Mattingly themselves:

> if the onset frequency of the transition of the second formant during a stop release is sufficiently low, relative to the frequency of the following steady state, the stop is perceived as labial; otherwise, as apical or dorsal (11).

To effect this regularity, the perceptual system would need (1) some proximal state that co-varied with the relative frequency of the second format, and (2) a computational process that mapped values of this state that correspond to formants with low initial values onto the feature [+labial], and not otherwise. The system would not have to *represent* that [+labial] states give rise articulatory gestures that result in acoustic properties such as the above. It would merely need to implement a computational system that transformed its proximal input so as to *track* this regularity.

There are many such regularities that seem to factor into our speech perception abilities. In the literature, they are called *cues*. Second formant transitions such as the above cue the perception of [+labial]. Liberman & Mattingly note that such cues are often in themselves sufficient for perceiving certain features, but never necessary. Moreover, they stand in complementary relations for one another, so that the absence of one cue can be compensated by the presence of another. So, for example, both differences in formant transitions and in silence can cue the difference between [sa] and [sta]. If differences in the formant transitions is degraded, increased

difference in silent periods maintains the perceptual distinction, and *vice versa* (Liberman & Mattingly, pp. 11-12).

Liberman & Mattingly cite the above phenomena as evidence that motor instructions, and not acoustic properties, are the appropriate *objects* of speech perception. Because of the variety of acoustic cues, 'there is simply no way to define phonemic categories in purely acoustic terms' (12). If the goal is to list necessary and sufficient acoustic conditions on the property of *being a phoneme*, they may well be correct. The list of sufficient conditions would be long, and the necessary non-existent. There simply may be no acoustic objects that could be identified as phonemes.[56]

All this is just to recapitulate the observation that began this section: that the phonemic states of hearers seem to vary roughly 1-1 with the phonemic states of speakers, but not with the acoustic events they generate. Thus, Liberman & Mattingly conclude that we cannot, as GA theorists would, individuate phonemic states in terms of the acoustic stimuli to which they respond. They write: 'An auditory theory that accounts for invariant perception in the face of so much variation in the signal would require a long list of apparently arbitrary assumptions' (14). It's unclear, however, precisely what would be problematic about such assumptions.

---

[56] See Rey's (2006; 2008; 2012) further arguments in favor of this claim, contra Devitt (2006a).

On the one hand, one might worry about the supposed 'arbitrary' nature of the assumptions about the structure of the perceptual system. Why should a rising second formant before /i/, and a falling second formant before /u/ both give rise to a perception of /d/? That the auditory perceptual system makes such a distinction would seem arbitrary unless we consider that the perceptual system is so constructed so as to be able to covary with utterances of /di/ and /du/.

But, this apparent arbitrariness problem is no problem at all. From the perspective of the theorist, of course, it may be obscure why the perceptual system *should* sort proximal stimuli the way it does. It's thus useful for the *theorist* to realize that the system is sorting stimuli into equivalence classes that co-vary with articulatory instructions. But, it does not follow from this that it is useful for the perceptual system itself to have this realization! It could just be a happy historical (i.e. evolutionary) accident that the speech perception system is structured in just such a way that it reliably is able to reconstruct the phonetic states of speakers out of the auditory stimuli they generate.

Alternatively, the problem may not be that the list of proximal stimuli that give rise to the same phonemic percepts is *arbitrary*, but simply that it is *so long!* You might think it an inelegant solution to the perception problem to suppose that it just relies on a look-up table by which it compares proximal input to a long list of possible input paired with the appropriate phonemic output. Each phonological state would have to be individuated in terms of a very long, hugely contextually dependent disjunction of proximal auditory stimuli.

Now, just because this method of individuation would be *inelegant* does not demonstrate that it's not in principle *possible*. In fact, we might not even have to reference all disjuncts of the stimulus disjunction in order to individuate phonological states. Even if there are not any acoustic cues that are necessary to engender a /d/ percept, we might be able to individuate the state in terms of just *some* subset of conditions we could demonstrate are sufficient. Thus, /d/ could be individuated as just that state that is tokened by, amongst others, those proximal stimuli eventuated by the /di/ and /du/ type utterances above, and any other states that listeners consciously perceive as being phenomenally *similar*. An arbitrary state, /x/, is identified as a /d/ just in case it is so perceptually similar to the state perceived to be in common in /di/ and /du/.

The problem with this method of individuating phonological states is that conscious phenomenal similarity does not seem to be a necessary property of phonological states. True, phonological state tokens of the same type usually have this property, and it's often used as a proxy by speech scientists to identify which phonological state a stimulus has engendered in a listener. Nonetheless, phonological states can be tokened even if listeners are not consciously, perceptually aware of their tokening. Subjects with Broca's aphasia may have difficulty with syllable identification and discrimination. But, this inability to explicitly respond to discrimination tasks doubly dissociates with speech recognition impairment (Holt & Lotto 2010, p. 1223). Such subjects seem to token phonemic states without

recognizing a phenomenal similarity between them. So conscious perceptual similarity can't serve as a necessary condition on phonological states.

It's fine to type-identify entities via description of properties that they may turn out to merely accidentally share rather than possessed constitutively. So, insofar as current syntactic theory has it that verbs check case on nouns, but not vice versa, it's fine to type identify noun and verb states by reference to this property-- even though we may in future discover nouns that check case on verbs. But this maneuver is kosher only insofar as this case-check asymmetry is, however contingently, a universal means for type-identifying out all states we currently identify as nouns and verbs. Perceptual similarity to an index phoneme perception just doesn't have this universal property.

So, if we are to type-individuate phonological states on the basis of proximal auditory stimuli, we'll have to do so by reference to a long, multifarious disjunction of such stimuli. But, the problem goes even deeper than this. We have evidence that the contextual cues that could cause us to revise our tokening of a phonological state constitute an *indefinite* disjunction. In this sense, there do not even seem to be *sufficient* conditions on the proximal stimuli that could be used to type-individuate phonological states. For any given proximal auditory stimulation, an indefinite disjunction of contextual information could change what phonological state it gives rise to.

These context effects are legion, so I'll survey only a few striking examples here.[57] In what follows, I'll continue the convention of containing IPA expressions in {curly brackets} when using them to describe external phenomena (i.e. acoustic blasts or articulatory gestures), thus leaving [square brackets] and /slashes/ to unambiguously refer to phonetic and phonemic mental states, respectively.

Famously, the same acoustic blast is perceived as /ga/ when preceded by {al} and /da/ when preceded by {ar}. This contextual effect is fortuitous because we are wont to pronounce /da/ and /ga/ differently relative to the preceding syllable. Producing {ar} shifts the tongue toward the back of the mouth, so the following {d} or {g} is also articulated further back in the mouth. Pronunciation of {al} likewise shifts articulation forward. Outside of this context, {da} usually has a higher F3 onset than {ga}. But, when produced after {ar}, further back in the mouth, the F3 onset of {da} lowers. Similarly, when produced forward in the mouth, after {al}, the F3 onset of {ga} goes up. Thus, the acoustic blasts produced in pronouncing /da/ and /ga/ in these contexts are extremely similar. Nonetheless, our perceptual system does not perceive these acoustic blasts as the same phoneme: it's able to compensate, perceiving it as a /da/ before {ar} and a /ga/ before {al}.

There's reason to think that this particular context effect is not mediated by representations of the human articulatory system. It's been found to obtain in native

---

[57] For a helpful review, from which many of these examples are taken, see Holt & Lotto (2010).

Japanese speakers (Mann 1986), who do not distinguish /l/ and /r/ phonemes, as well as quail (*C. japonica:* Lotto et al., 1997)! One explanation is that the perceptual system simply perceives F3 onsets as higher than they actually are after low F3 offsets, and *vice versa* (Lotto et al. 1997; Diehl, Lotto & Holt 2004).

But, the context in which perception of these sounds switches from one phoneme to another is not exhausted by F3 offset effects. Fowler et al. (2000) showed that the same sound can be shifted from being perceived as a /da/ to a /ga/ and *vice versa* by the preceding *visual* context. The sight of lips mouthing {al} before the sound leads it to be perceived as /ga/; the sight of {ar} changes perception to /da/. Fowler takes the result to show that articulatory representations *are* mediating phoneme perception.

But, taken together with the quail data, these results raise the possibility that the perception of /da/ versus /ga/ is mediated neither by proximal audio-visual context nor by representations of the human articulatory apparatus. Rather, the underlying contextual effect might be best described at the level of phonemes themselves. The rule might be something like: *ceteris paribus*, sound {%} is perceived as /da/ when preceded by /ar/, and /ga/ when preceded by /al/. It makes no difference which proximal stimuli occasioned the tokening of the /al/ or /ar/: be they auditory, visual, or something else, what matters is whether an /al/ or /ar/ has been tokened in the mental phoneme box.

But, whether such context effects are driven by low level auditory contexts or by top-down phonemic or lexical beliefs makes no difference to our ability to type-

individuate phonological states. You might think we could pick out /da/ under the description: 'the state that's tokened when proximal stimulus {%} impinges on the perceptual system in the case that it has just tokened an /ar/.' But, individuating /da/ in terms of /ar/ presumes that we have a good way of type-individuating /ar/: that's far from clear.

First, /ar/ would seem to be susceptible to the same sorts of contextual effects, making it difficult to individuate *it* in terms of proximal auditory stimuli. Further, context effects such as the above are defeasible because they can propagate backwards in time. For example, the same sound may be perceived as an /s/ when it's followed by a {u}, but a /ʃ/ when followed by {a} (Mann & Repp 1980). These effects could also be described at the phoneme level. Whether sound, {s}, is perceived as an /s/ or an /ʃ/ could turn on whether following stimuli are parsed into a /u/ or an /a/. But, this raises the possibility that the rule described above is defeasible. It's true that the sound {%} may give rise to a /da/ when following /ar/-- but that is subject to change relative to the following phonemic context.

Even if we could characterize a finite disjunction of phonemic context effects, there are nonetheless context effects that operate independently of phonemes. Auditory signals that are not perceived as speech, and thus would not influence phonemic context, nonetheless also influence phoneme categorization (Holt & Lotto 2010, p. 1222).

Furthermore, the generalization is also susceptible to revision in light of lexical, semantic, and more global context effects. For example, non-phonemic

acoustic properties of a preceding phrase can change phonemic perception of a target. The same sound will be perceived as either 'bit' or 'bet' depending on whether the preceding phrase-- 'please say what this word is'-- has its F1 frequencies raised or lowered. So, whether a particular sound, {%} is perceived as a /da/ may depend not only on the phonemes that are tokened before and after it, but also on the global nature of the speech sounds preceding it.

As the previous example suggests, phonemic perception is also subject to lexical context effects. In the Ganong Effect (Ganong 1980) an ambiguous sound is heard as /g/ when followed by 'ift' and as /k/ when followed by /iss/. Additionally, such lexical effects interact with more global context: they are enhanced in the context of lexical stimuli and attenuated in the context of non-lexical phonemic stimuli (Mirman et al. 2008). They also interact with phonemic context effects. An ambiguous sound that that shifts from an /s/ to an /ʃ/ due to the lexical effects from a preceding 'bli-' or 'bru-' engenders compensation for coarticulation in perception of the following phoneme. The same sound will be heard as a /t/ in the 'bru-' context and as a /k/ in the 'bli-' context because a preceding /s/ prompts the following phoneme to be heard as articulated backward in the mouth (/k/), and a preceding /ʃ/ as more forward (/t/) (Magnusson et al. 2003).

There also seem to be possible semantic context effects on phoneme perception. The same sound will be heard as 'goat' in the context of a sentence like:

'The dairy farmer forgot to milk the _____' and as 'coat' in context of a sentence like 'The tailor took care to press the _____' (Borsky, Tuller, & Shapiro 1998).[58]

So, phonemic, lexical, and global context effects are not only multifarious, but interactive. The lesson is that any attempt to circumscribe the conditions under which a particular phoneme will be tokened in perception is defeasible by some other contextual condition. Thus, any attempt to type-identify phonemes in terms of the proximal stimuli that gives rise to them will fail.[59] Moreover, as argued in Section 4.4, neither can we individuate phonological states in terms of their relations to one another. As we saw in Section 5, type-individuating conceptual states in terms of their intentional content allowed us to individuate them absent a similar inability to do so in terms of their conceptual roles or responsiveness to stimuli. The same strategy thus recommends itself for our present predicament.

6.2. General Auditory Approaches: Against Articulatory Individuation

Though individuating phonemic states in terms of either their relations to one another or the stimuli that give rise to them fails, we need not yet concede that we must type-

---

[58] It is, however, unclear to me whether these results indicate an effect on *phoneme* perception rather than just *lexical* perception.That is, the language faculty may have constructed a phonemic structure that ordinarily would pick out the 'goat' lexical item, but later revised its lexical retrieval without revising the original phonemic structure. Fodor (1983, p. 65 ff) addresses a similar worry in regard to interpretation of the phoneme restoration effect Here, an utterance of an {s} is replaced by white noise in a word like 'legislature,' viz.: 'legi{%}lature.' The white noise is nevertheless perceived as an /s/. It seems on the face of it that this perception is driven by the *lexical* context rather than the auditory or phonemic context.

[59] This is precisely the situation that undermined the Verificationists' attempt to type individuate conceptual states in terms of their verification conditions!

individuate phonological states in terms of intentional content. One last non-intentional alternative remains. That is that phonological states can be individuated in terms of the articulatory movements they give rise to in production. So, for example, it's possible that a feature like [+voice] could be individuated as just that feature that when appropriately tokened leads the vocal folds to vibrate.[60] However, we'll see that even this strategy proves inadequate because, just as with phoneme perception, there are context effects operating on speech production as well. Thus, in the end, we will have to type-individuate phonological states in terms of intentional content.

We've already seen that phoneme perception is modulated by a tendency to compensate for coarticulation. This perceptual context effect goes along with actual coarticulation on the production side. The phonemic context in which a phoneme is embedded influences the actual articulatory gestures it gives rise to. Thus, you might worry that individuating phonemic states on the basis of articulatory production will fail because any particular tokening of a phoneme will have a wide range of articulatory consequences depending on the context in which it's embedded.

---

[60] Something akin to this view might describe Fowler's (e.g., 1996) DRT view. She holds that the 'objects' of speech perception are articulatory gestures. But, she seems to deny that phonological states are mental states. She writes that vocal tract gestures '(not their neural control structures) *are* phonological components of an utterance' (p. 1731). Further, we can perceive such gestures *directly:* 'unmediated by processes of hypothesis testing or inference making and unmediated by mental 'representations' in the literal sense of mental stand-ins for real-world things' (ibid.), and yet also independent of variation in acoustic cues.The view seems to be that listeners instantiate states that co-vary with articulatory gestures in a manner that skips over any sort of processes that might mediate between the production of those gestures and the tokening of the states. She asserts that 'the theory of direct perception is not a theory about magic' (p. 1739), but I confess ignorance as to how it is to work otherwise!

Thus, phonemic features cannot be mapped one-to-one with actual articulatory movements. The movement to which any given phonetic feature gives rise depends upon the current state of the vocal tract at the time the feature is tokened. For example, 'lip rounding' may involve not just the lips, but the jaw as well (Liberman & Mattingly, p. 22). The same articulatory structure may also be recruited to enact different gestures at the very same time. So, the lips may be simultaneously rounding to produce [u] and closing to produce [b] when producing the syllable [bu] (ibid.). So, we cannot individuate phonetic features in terms of the particular articulatory movements to which they give rise, because these movements change depending on context.

There nonetheless remains the possibility that we can individuate phonetic features in terms of the *changes* they induce in the vocal tract regardless of context. Liberman & Mattingly point out:

> for any particular gesture, the same sort of distinctive deformation is imposed on the current vocal-tract configuration, whatever this 'underlying' configuration happens to be (p. 22).

So, for example, the feature [+open] initiates movement in the tongue, lips, jaw, and hyoid to varying degrees, depending on context. But, in all contexts, these articulators move so as to 'give the tract a more open, horn-shaped configuration than it would otherwise have had' (Liberman & Mattingly, p. 23). So, perhaps [+open] can be type-individuated as just that feature that gives rise to a more horn-shaped configuration of the vocal tract.

But, if this is the way in which we individuate features, it's unclear that we can do so without attributing to them intentional content *as of* the changes they effect. After all, if [+open] effects the change above, it does not do so through some simple causal reflex transmitted to articulators. In order to effect the open, horn-shaped configuration in any given context, the articulatory system would have to take into account how the articulators are positioned in that context, and calculate how to transform them so as to bring about the requisite shape.

More than this, [+open] isn't occasioning the production of the *same* shape each time, but simply a shape that is *more* open than it would otherwise be. Thus, it seems as though the articulatory system would have to represent counterfactuals about what is more or less open relative to an indefinite span of contexts. So, if we individuate phonemic states in terms of the articulatory stats to which they give rise, we must assume that they *represent* changes to the articulatory system.

It's still unclear, however, whether we can even type individuate phonemic states in that manner at all. First, there is evidence that non-human animals of diverse phylogeny seem sensitive to phonemic distinctions across contexts. So, for example, Chinchillas (Kuhl 1981; Kuhl & Miller 1975), macaques Kuhl & Padden 1983), zebra finches (Dooling et al. 1995), and Japanese quail (Kluender et al., 1987; , Lotto et al.,

1997) have all demonstrated abilities to discriminate phonemic contrasts[61]. It seems doubtful that such animals make use of representations of the human vocal tract to accomplish such feats!

Of course, just because non-human animals are able to discriminate phonemic distinctions in auditory stimuli, it does not follow that we must individuate the states they thereby token as the *same* phonemic states tokened by the human speech perception system. It could be that we share with chinchillas, Japanese quail, and other animals an ability to parse auditory stimuli into segments that correspond to the acoustic phenomena created by human speech utterances. It would not be surprising if humans capitalized upon this general auditory capacity when evolving a set of phonemic items for use in language. That is, it could be that phonemic states were selected that effected articulatory movements that eventuated in just those acoustic phenomena that our general auditory capacity could easily parse. Nonetheless, perhaps the phonemic states so selected are type-individuated in articulatory terms according to the manner described above.

However, there's a further worry that sometimes phonetic features don't seem to have any articulatory consequences at all. Phenomena such as assimilation and schwa deletion suggest that sometimes the phonological system tokens phonetic features in production, but that they nonetheless fail to have articulatory

---

[61] In many cases, the animals can do this within diverse phonemic contexts. For example, the quail can discriminate the /d/ in both /di/ and /du/.

consequences. In this the case, individuating [+alveolar] as just that feature that gives rise to contact of the tongue on the alveolar ridge would fail to pick out instances of the feature that fail to have these consequences.

Assimilation occurs when certain features that normally would have certain articulatory consequences instead have articulatory consequences associated with different features. In English, the word 'green' is usually pronounced with a final [n] sound. However, when uttered with certain other words, as in 'green boat,' speakers often pronounce the final consonant as an [m]-- so it sounds as though you might read the nonsense phrase, 'greem boat.' When saying 'green king,' they sometimes pronounce the final syllable as [ŋ], as one might read the nonsense phrase, 'gring king.' In doing so, they are changing the final syllable of 'green' such that it has the same [place] feature as the [k] in 'king' and the [b] in 'boat.' Just as [b] is articulated by bringing the two lips together, so is [m]. Otherwise, [m] is identical in terms of its feature geometry to [n]. The same goes for [n] and [ŋ]: they differ only insofar as the placement of the tongue in [ŋ] matches that in the [k] of 'king.' In this way, words like 'green' are susceptible to having their final consonant assimilate the [place] feature of the first consonant of the following word.

Assimilation seems to be an 'optional' feature of English idiolects. Some people do it, some don’t, and those that do it tend not to do it consistently: there seems to be no way of specifying the contexts in which it takes place. Through all that, few other than phoneticians notice that anyone is varying their speech in this way. For this reason, it’s often thought of as a regularity of the language that is

nonetheless not the result of any phonological rule. That is, it seems to be the result of performance factors rather than an aspect of phonological cognition proper[62]. So, it would seem that phonological processes would generate a surface phonetic form [grin bot] even though in production, the articulation produced is {grim bot}. Thus, if we individuate [n] as just that feature that gives rise to {n} articulations, we'd miss out on the instances in which it instead gives rise to {m} productions!

A similar phenomena occurs in schwa deletion. Words such as 'every,' 'family' and 'memory' generally have the same middle vowel: a schwa, or [ə]. In production, however, [ə] is sometimes deleted in these contexts so the words are pronounced 'fam'ly,' 'ev'ry,' etc. Again, this deletion seems to be an optional feature even of individual idiolects, so it's not likely the result of a phonological rule specifying that [ə] gets deleted in the phonetic form in certain contexts. Instead, it seems as though [ə] is always present in the phonetic form of these words, but does not always have articulatory consequences.

Browman & Goldstein (1992) survey evidence that such assimilation and deletion phenomena could be accommodated by a principle that individuates features

---

[62] Alternatively, we could account for assimilation in terms of the intentional content of phonological states. Because the assimilated sound depends upon the features of phonological states occurring later in production, it seems the phonological system must *anticipate* what these features will be in order to effect the assimilation. In order to anticipate, it may well have to *represent* what these features will be (thanks to Georges Rey for making this point in conversation). If this is the case, so much the better for my considered view that phonological states are intentional.

in terms of articulatory instructions in the terms given above. They note (pp. 36-39) that articulations in which assimilation or deletion has taken place often differ from articulations of segments that simply don't have the deleted feature in the phonetic form. For example, when the initial schwa in 'beret' is deleted, the articulation formed is nonetheless distinguishable from that that standardly produces 'bray.' Thus, they argue that assimilation and deletion are not cases in which the assimilated or deleted features have *no* articulatory consequences. Rather, they are simply cases in which the surrounding context makes the articulatory consequences of these features more imperceptible than usual.

Nonetheless, even if the Browman & Goldstein picture is correct, we'd be without a principle of individuation for phonological states. It may be that in instances of assimilation, assimilated phones still *have* articulatory consequences. But, nonetheless, these consequences are different depending on whether the phone is assimilated or not. Nonetheless, the case still establishes that any attempt to type-individuate the phonological states in terms of their articulatory consequences would have to appeal to a more or less unwieldy disjunction of possible articulations.

Alternatively, you might suppose that we could type-individuate phonological states in terms of *ceteris paribus* characterizations. For example, we might individuate [n] as the state that, *ceteris paribus,* causes the movement of the tongue against the palate. It's only when *cetera* are not *paris,* as when 'green' is pronounced as gree{m} before 'boat,' that [n] does not give rise to such articulations.

But, if we are to allow such *ceteris paribus* characterizations, we ought to have some principle for accounting for the instances in which exceptions to the rule arise. We need not be able to generalize all the cases in which we can expect exceptions, but for any given exception, we ought to be able to explain why the rule has failed.[63] These explanations don't seem available to us in the case of assimilation. As we noted, the contexts in which assimilation occurs appear to be largely arbitrary.

So, we cannot individuate phonemic states directly in terms of their articulatory consequences, because, as with acoustics, there is no one-to-one mapping between them. We may perhaps be able to individuate such states in terms of their context-relative effects on the vocal tract. But, doing so seems to require that we attribute to them intentional content as of aspects of that vocal tract. If we cannot individuate phonemic states in these terms, I argue in the next section that there is nothing for it but to individuate them in terms of intentional content.

7. Intentional Phonology

The lesson of the foregoing is that the contexts that may influence whether a given proximal input may give rise to one phoneme or another do not seem characterizable in terms of properties of those proximal stimuli themselves. We seem even incapable of delineating a finite disjunction of proximal stimuli that characterize the context effects to which stimuli impinging on the speech perception system are susceptible.

[63] See Pietroski & Rey (1995) for an account of *ceteris paribus* generalizations along these lines.

Similarly, we seem unable to individuate phonological states in terms of the motor instructions to which they give rise.

But neither, if Hale & Reiss are correct, can we individuate phonological states in terms of their *de facto* relations to one another. Like concepts, they combine too promiscuously to identify any *de facto* relation between phonemes as constitutive. We cannot take the strategy Collins adopts for individuating syntactic states. It may be true that VPs check case on NPs and not *vice versa*, but no such relational generalizations seem to hold for phonological states.

We thus cannot individuate phonological states in terms of either their relations to one another or in terms of their relations to states at their articulatory and perceptual interfaces. As remarked above, this is precisely the impasse Fodor takes us to be in with regard to conceptual states. They are too promiscuous to be individuated by their internal relational states, and too flexible to be individuated in terms of either their behavioral consequences or the proximal stimuli that give rise to them. Ascribing content to concepts makes psychology possible in that it allows us to generalize across subjects that exhibit diverse behaviors and arrangements of mental states. So too, attributing content to phonological states can make phonology possible.

Absent a way to individuate phonological states via their relations to one another or their proximal connections to the external world, we can individuate them in terms of their relations to intentional contents. Thus, a /p/ state is individuated by the fact that its intentional content is as of the property of being *p*!

This claim of course raises the question as to what exactly it is to be a *p*. Is it to be a sound of a particular type? A particular articulatory gesture? What? We can be non-committal on the answers to these questions while still buying that /p/-states are individuated in terms of p-contents. After all, we can hold that psychological laws hold across mental states that are individuated in terms of being about *cats* while remaining agnostic as to what exactly it is to *be* a cat. To be a CAT-state is just to be about *cats*, however *they* are individuated.

Of course, there are interesting questions as to what explains *how* it is that any intentional state-- phonological states among them-- has the intentional content that it does. The general answers are familiar. It might be some counterfactual disposition to co-vary with the things it represents in a particular way. Or, it might be the result of bearing some internal relation to other intentional states. You might worry, then, that the considerations I've raised against individuating *phonological states* in terms of either their internal or external relations would preclude us giving any standard story-- either internalist or externalist-- of the *content* constitution of those states.

This worry, however, is ill-founded. I've argued that we can't individuate /p/ states in terms of the acoustic phenomena that in fact cause them. There seems to be an indefinite disjunction of acoustic phenomena that sometimes will and sometimes will not cause the tokening of a /p/ depending on a host of contextual factors. Nonetheless, we could coherently claim that to have the content as of *p* just is to asymmetrically co-vary with a particular acoustic phenomenon under ideal conditions (á la Fodor, 1987).

Similarly, we could give an internalist semantics of the contents of phonological states consistent with the obstacles to individuation canvassed above. It may well be that two people share a /p/ state despite massive differences in their inter-phonemic state relations. Nonetheless, what constitutes a /p/ as such may be, say, its disposition to be tokened as a result of certain auditory mental states being tokened in some ideal circumstance. In any event, it's perfectly coherent to hold that the *content* of a mental state, /p/, may be constituted by its counterfactual tendencies to co-vary with internal and external entities in particular circumstances while denying that the *state* is individuated in terms of the entities it *in fact* co-varies with.

Alternatively, it could be that the intentional contents of phonological states are intentional inexistents, along lines argued by Rey (2012). This view has it that phonological mental states have as content objects that do not in fact exist. While acoustic phenomena may *cause* the tokening of phonological states, we could hold that the content of those states is not any particular set of acoustic phenomena. We could go even further to insist that phonological states have as content properties that couldn't *possibly* exist. Perhaps, for example, the content of /b/ states is a particular acoustic property that could not exist as a matter of metaphysical possibility.

Whatever of these stories of content constitution we might adopt, there are a number of other subtle questions to deal with concerning the nature of phonological content. For example, we'd want to address the possibility that phonological contents admit of Frege cases. They may even be hyper-inten*s*ional (with an *s*!), in that they are sensitive even to differences between necessarily co-extensive terms! My

suspicion (as aired earlier in Chapter 3, n. 29) is that Frege cases won't arise in our best theory of mental content. But, that's a suspicion only, and I won't attempt to argue for it here.

Indeed, I prescind from taking a position on exactly which of any the above positions we should take in regard to phonological states. They are all interesting questions, but far beyond the scope of the current work. All that's needed for our present purposes is to allow that /b/ states have content as of *b*s and /p/ states have content as of *p*s, whether *b*s and *p*s are particular acoustic phenomena, articulatory phenomena, or inexistent objects (whether possible or impossible). Just what *b*s and *p*s are will likely require at examining carefully the counterfactual generalizations of phonology.

The key property of intentionality that is useful for us is that it gives us a way of characterizing states independently of whatever contingent relations they may happen to enter into. Whichever answers we may adopt to the questions posed above, this unique property of intentional content will remain. And, it's this property that plays the key role of allowing us to type-individuate mental states despite vast differences in the causal and computational relations they enter into as a matter of fact.

My argument that phonological states are individuated intentionally is not demonstrative, but rather abductive. Given that we can't individuate them any other way, individuating in terms of intentional content would seem to be the best option.

But, ascribing intentional content to phonological states also buys us additional explanatory virtues.

Intentional content allows us to make counterfactual generalizations and predictions about just the sort of circumstances under which certain phonological structures might be tokened. So, for example, consider the McGurk effect. Subjects who see someone uttering {ga}, but synchronically hear the sound {ba} mentally token the phoneme, /da/. Why do these particular audiovisual stimuli cause a tokening of /da/? Well, it may be because the phonological system *represents that* /ba/ states usually result in closed lips, represents that the lips of the speaker are not closed, and so *infers* that the speaker is not producing a /ba/. Her phonological system may also represent that /ga/ states usually result in acoustic consequences different from what she's hearing. The acoustic states she *is* hearing, she may represent as more characteristic of /da/ states. So, all things considered, her phonological system infers that it is perceiving a /da/.[64]

This sort of explanation would not be open to us if we did not individuate phonological states intentionally. We can explain why the phonological system took a particular computational route from proximal auditory stimuli to representing a /da/ because this route preserved semantic properties of the phonological states. Given that the system represented the mouth as being open and something like the

---

[64] cf. Boersma (2012).

conditional: *{mouth open}* → *not-/b/*, it could not have tokened a /b/ and still have preserved the truth of the conditional.[65]

This is not necessarily to say that the semantic properties of the phonological states *caused* the system to transition as it did. The above picture is perfectly compatible with a methodological solipsism in which we assume that it was the syntactic properties of the mental states that caused one to transition to the other. Any particular semantic inference will be implemented by a syntactic process sensitive only to the local causal properties of the states bearing the intentional content.

Nonetheless, we can make sense of why a /da/ gets tokened as the result of the McGurk effect *and* as the result of seeing someone whisper, *and* as a result of hearing a certain sound over a garbled loudspeaker, by generalizing that in each instance, the computational process that led to tokening of a /da/ was truth preserving of the phonological system's representations *about* the property of being a *d*. Individuating in terms of intentional content allows us to generalize over the *types* of computational processes to which phonological states are sensitive. It is sensitive to computational processes that involve representations as of being a *d.*

Thus, intentional explanation provides two virtues to phonological theory: namely, the two laid out in Chapter 3. First, it allows us to type-individuate the states

---

[65] Of course, the picture also allows for characterization of cognitive processes as consisting of causal relations between semantically characterized states. I remain agnostic as to whether it's best to construe mental processes as causally sensitive to semantic properties or not.

over which the theory quantifies independently of either their relations to one another or phonology-external states. Secondly, the correctness conditions allow for counterfactual generalizations that would not be available to us otherwise

The voice synthesis model of phonology that Liberman & Mattingly propound seems to make use of intentional content in these two respects. They postulate a voice synthesis module within the perceptual system. When acoustic input arrives, it is compared against candidate articulatory antecedents generated by a synthesizer that

> incorporates complete information about the anatomical and physiological characteristics of the vocal tract and also about the articulatory and acoustic consequences of linguistically significant gestures (p. 26).

Thus, upon encountering the McGurk effect stimulus, the synthesizer may generate a list of phonemes likely to have generated the auditory stimuli and a list likely to have generated the visual stimuli. It will from these select the phoneme likely to have caused both the auditory and visual stimuli. It's hard to see how to describe this process if not in representational terms.

It cannot simply identify phonemes with particular auditory or visual stimuli because *ex hypothesi* there are multiple possible phonemes that could be the cause of the proximal stimuli. Thus, it must have a way of type-individuating the states independently of the proximal states that give rise to them. As we've noted, ascribing intentional content to the states would be a useful way to accomplish this individuation. Ascribing intentional content to the states hypothesized by the synthesizer would also allow us to describe in counterfactual terms the type of stimuli

that the synthesizer is sensitive to. We could, for example, explain the McGurk effect by reference to the intentional content of the synthesizer's states.

The key to explaining the McGurk effect is to explain why the same acoustic phenomenon is perceived as a /ba/ in isolation from visual stimulation, but as a /da/ in the presence of the visual stimulus. We can explain why the visual stimulus makes a difference to the synthesizer by appealing to its representation that the utterance of /da/ phonemes result in the sort of visual stimuli present in the McGurk stimulus, whereas /ba/ phonemes generally do not. Thus, the McGurk visual stimulus causes the acoustic stimulus to be perceived as a /da/ because of representations the synthesizer has *about* the visual consequences of /da/ and /ba/ phonemes. Thus, attributing intentional content to the phonological synthesizer allows us to capture counterfactual generalizations about its operation.

Now, Liberman & Mattingly identify the 'objects' of phonological perception as the mental states themselves that give rise to articulatory gestures. So, when I token a /d/ as the result of you making a particular utterance, I am representing that your utterance is the result of you yourself having tokened a /d/ mental state, which gave rise to your utterance. It's unclear that we are licensed to make this strong conclusion about what precisely the *content* of phonological states are. The contents may, for example, be intentional inexistents, such as those proposed by Rey (2006).

For example, even if something like the Liberman & Mattingly's synthesizer model is correct, it may not represent phonemes *as* motor instructions *per se*. In fact, it may not represent any metaphysically interesting characteristics of phonemes at all:

it need not represent them as mental states, or sounds. It need only postulate that they exist and that particular visual, auditory, and perhaps semantic information serves as good evidence as to whether they are present or not. That, at least, seems sufficient for running the explanations sketched above.

In any case, the individuation thesis I'm advancing here is metaphysical rather than epistemic. I don't for a moment presume that phonologists will use the criteria I've laid out here as a means of *figuring out* which phonological state a subject has tokened. The phonologists will likely continue to rely on, at least as a first gloss, their own 'impressionistic representation… of the acoustic or articulatory realization' of phonological states (Hale & Reiss, 2008, p. 146). The thesis that phonological states are intentional simply makes it intelligible how linguists' generalizations can metaphysically range over mental states subject to indefinite disjunctions of internal and external relational properties.

8. Conclusion

Thus, phonological explanation helps vindicate the contentions of this thesis. First, as argued in Chapter 1, there is a coherent distinction between computations with and without representation. Thus, it's a live question as to what extent a computational theory of mind must appeal to intentional content to run its explanations.

Chapter 2 argued that many theories of intentional content advanced over the last several decades assumed an explanatory role for intentional content that was vacuous. They assumed that appeal to intentional content merely vindicates our unreflective intentional *talk*, or provides a useful nomenclature for recording

generalizations that could be captured just as well in non-intentional terms. Neither of these functions, I hold, amount to an interesting explanatory role.

Chapter 3 argued that nonetheless there are instances of computational cognition that *do* require non-vacuous appeal to intentional content for their explanatory power. Namely, intentional content allows us to type-individuate states within and couch counterfactual generalizations over computational systems that are open to an indefinite disjunction of proximal input. By contrast, systems that are encapsulated and subject to just a finite disjunction of input do not require intentional content to type-individuate their states or make counterfactual generalizations over them.

Chapter 4 demonstrated that such a distinction still allows for intentional explanations of unconscious, modular processes that are not necessarily part of a person-level global belief fixation system. This final chapter has vindicated this claim by examining one such modular process: the phonological system. The phonological system is largely (though perhaps not wholly) insensitive to our conscious belief states. Nonetheless, it exhibits a competence such that it is sensitive to an indefinite disjunction of possible proximal input. Consonant with our conclusions from Chapter 3, intentional content is therefore necessary to type individuate phonological states and generate counterfactual generalizations over them.

Nonetheless, as we saw in Chapter 3, there may be a wide array of cognitive processes that do not evince such a competence, and thus do not require intentional explanation. Thus, we have vindicated the intentional theory of mind insofar as many

of our thoughts-- conscious, person-level, or otherwise-- are indeed *about* something. But, we have also seen that this is not always the case. There are many cognitive processes that are not *about* anything at all.

Bibliography

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. Nature Neuroscience, 7(10), 1057–1058. http://doi.org/10.1038/nn1312

Allred, Sarah (2012). Approaching color with Bayesian algorithms. In Gary Hatfield & Sarah Allred (eds.), Visual Experience: Sensation, Cognition, and Constancy. OUP: Oxford, p. 212.

Bazylinski, D. A., & Frankel, R. B. (2004). Magnetosome formation in prokaryotes. Nature Reviews Microbiology, 2(3), pp. 217–230. http://doi.org/10.1038/nrmicro842

Beckman, J. N. (1997). Positional faithfulness, positional neutralisation and Shona vowel harmony. Phonology, 14(1), 1–46. http://doi.org/10.1017/s0952675797003308

Beets, I., Hart, B. 'T., Rösler, F., Henriques, D., Einhäuser, W., & Fiehler, K. (2010). Online action-to-perception transfer: Only percept-dependent action affects perception. Vision Research, 50(24), 2633–2641. http://doi.org/10.1016/j.visres.2010.10.004

Block, N. (1987). Advertisement for a Semantics for Psychology. Midwest Studies In Philosophy, 10(1), 615–678. http://doi.org/10.1111/j.1475-4975.1987.tb00558.x

Boersma, P. (2012). A constraint-based explanation of the McGurk Effect. In Roland Noske and Bert Botma (eds.): Phonological Architecture: Empirical, Theoretical and Conceptual Issues, 299-312. Berlin/New York: Mouton de Gruyter.

Borsky, S., Tuller, B., & Shapiro, L. P. (1998) 'How to milk a coat:' The effects of semantic and acoustic information on phoneme categorization. The Journal Of the Acoustical Society of America, 103(5), 2670. http://doi.org/10.1121/1.422787

Brainard, D. H., Longere, P., Delahunt, P. B., Freeman, W. T., Kraft, J. M., & Xiao, B. (2006). Bayesian model of human color constancy. Journal Of Vision, 6(11), 10–10. http://doi.org/10.1167/6.11.10

Brainard, D. H., & Freeman, W. T. (1997). Bayesian color constancy. Journal Of the Optical Society of America, 14(7), 1393. http://doi.org/10.1364/josaa.14.001393

Brentano, F., & Kraus, O. (1874/1995). Psychology from an empirical standpoint. London: Routledge.

Browman, C. P., & Goldstein, L. (1992). Articulatory Phonology: An Overview. Phonetica, 49(3-4), 155–180. http://doi.org/10.1159/000261913

Burge, T. (2010). Origins of objectivity. Oxford: Oxford University Press.

Burton-Roberts, N. 2000: Where and what is phonology? A Representational perspective. In Burton-Roberts, N., Carr, P., & Docherty, G. J. (eds.). Phonological Knowledge: Conceptual and Empirical Issues. Oxford: Oxford University Press. 39-66.

Carr, P. 2000: Scientific realism, sociophonetic variation, and innate endowments in phonology. In Burton-Roberts, N., Carr, P., & Docherty, G. J. (eds.). Phonological Knowledge: Conceptual and Empirical Issues. Oxford: Oxford University Press. 67-104.

Carruthers, P., & Veillet, B. (2011). The Case Against Cognitive Phenomenology. Cognitive Phenomenology, 35–56. http://doi.org/10.1093/acprof:oso/9780199579938.003.0002

Carruthers, P. (2006). The architecture of the mind: massive modularity and the flexibility of thought. Oxford: Clarendon Press.

Carruthers, P. (2015). The centered mind: what the science of working memory shows us about the nature of human thought. Oxford: OUP.

Chalmers, D. J. (1996). Does a rock implement every finite-state automaton? Synthese, 108(3), 309–333. http://doi.org/10.1007/bf00413692

Chalmers, D. J. (1994). On implementing a computation. Minds And Machines Mind Mach, 4(4), 391–402. http://doi.org/10.1007/bf00974166

Chalmers, D. J. (2011). A Computational foundation for the study of cognition. Journal of Cognitive Science. (12) pp. 323-357.

Chomsky, N. (1955). The logical structure of linguistic theory. Cambridge, MA: M.I.T. library.

Chomsky, N. (1959) (Review Of) Verbal Behavior, By B.F. Skinner. Reviewed by Noam Chomsky. Language, 35, 26–58.

Chomsky, N. (1975). Reflections on language. New York: Pantheon Books.

Chomsky, N. (2000). New horizons in the study of language and mind. Cambridge: Cambridge University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral And Brain Sciences, 36(03), 181–204. http://doi.org/10.1017/s0140525x12000477

Collins, J. (2004) Faculty disputes. Language and Mind. 19, 503-533.

Collins, J. (2007a). Meta-scientific Eliminativism: A Reconsideration of Chomsky's Review of Skinner's Verbal Behavior. The British Journal For the Philosophy of Science, 58(4), 625–658. http://doi.org/10.1093/bjps/axm041

Collins, J. (2007b) Linguistic competence without knowledge of language. Philosophy Compass. 2, 880-895.

Collins, J. (2008) Knowledge of language redux. Croatian Journal of Philosophy. 8, 3-44.

Collins, J. (2009) The Perils of content. Croatian Journal of Philosophy. 9(27), 259-289.

Cummins, R. (1989). Meaning and mental representation. Cambridge, MA: MIT Press.

Davidson, D. (1987). Knowing One's Own Mind. Proceedings And Addresses of the American Philosophical Association, 60(3), 441. http://doi.org/10.2307/3131782

Dennett, D. C. (1987). The intentional stance. Cambridge, MA: MIT Press.

Devitt, M. (2006a). Ignorance of language. Oxford: Clarendon Press.

Devitt, M. (2006b). Defending Ignorance of Language: Responses to the Dubrovnik papers. Croatian Journal of Philosophy, 6, 571-606.

Devitt, M. (2008) Explanation and reality in linguistic theory. Croatian Journal of Philosophy, 8, 203-231.

DiCarlo, J. M. & Wandell, B. A. (2000) Illuminant estimation: Beyond the bases; Paper presented at IS&T/SID Eighth Color Imaging Conference; Scottsdale, AZ: p. 91-96.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004) Speech Perception. Annual Review Of Psychology, 55(1), 149–179. http://doi.org/10.1146/annurev.psych.55.090902.142028

Dieter, K. C., Hu, B., Knill, D. C., Blake, R., & Tadin, D. (2013). Kinesthesis Can Make an Invisible Hand Visible. Psychological Science, 25(1), 66–75. http://doi.org/10.1177/0956797613497968

Dill, M., Wolf, R., & Heisenberg, M. (1993). Visual pattern recognition in Drosophila involves retinotopic matching. Nature, 365(6448), 751–753. http://doi.org/10.1038/365751a0

Dooling, R. J., Best, C. T., & Brown, S. D. (1995). Discrimination of synthetic full-formant and sinewave /ra–la/ continua by budgerigars (Melopsittacus undulatus) and zebra finches (Taeniopygia guttata) The Journal Of the Acoustical Society of America, 97(3), 1839. http://doi.org/10.1121/1.412058

Dretske, F. I. (1981). Knowledge and the flow of information. Oxford: Blackwell.

Egan, F. (1992). Individualism, Computation, and Perceptual Content. Mind, 101(403), 443–459. http://doi.org/10.1093/mind/101.403.443

Firestone, C., & Scholl, B. J. (2015). Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. Behavioral And Brain Sciences Behav Brain Sci, 1–77. http://doi.org/10.1017/s0140525x15000965

Fodor, J. A. (1975). The language of thought. New York: Crowell.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. Behavioral And Brain Sciences, 3(01), 63. http://doi.org/10.1017/s0140525x00001771

Fodor, J. A. (1983). The modularity of mind: an essay on faculty psychology. Cambridge, MA: MIT Press.

Fodor, J. A. (1987). Psychosemantics: the problem of meaning in the philosophy of mind. Cambridge, MA: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. Cognition, 28(1-2), 3–71. http://doi.org/10.1016/0010-0277(88)90031-5

Fodor, J. A. (1994). The elm and the expert: mentalese and its semantics. Cambridge, MA: MIT Press.

Fodor, J. A. (1998) Concepts: where cognitive science went wrong. Oxford: Clarendon Press.

Fodor, J. A. (2000) The mind doesn't work that way: the scope and limits of computational psychology. Cambridge, MA: MIT Press.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000) Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. Journal Of Experimental Psychology: Human Perception and Performance, 26(3), 877–888. http://doi.org/10.1037/0096-1523.26.3.877

Fowler, C. A. (1996) Listeners do hear sounds, not tongues. The Journal Of the Acoustical Society of America J. Acoust. Soc. Am., 99(3), 1730.

Gallistel, C. R., & King, A. P. (2009). Memory and the computational brain: why cognitive science will transform neuroscience. Chichester, U.K.: Wiley-Blackwell.

Gallistel, C. R. (1990). The organization of learning. Cambridge, MA: MIT Press.

Ganong, W. F. (1980) Phonetic categorization in auditory word recognition. J Exp Psychol.,6:110–125.

Gigerenzer, G., & Todd, P. M. (1999). Simple heuristics that make us smart. New York: Oxford University Press.

Grah, G. (2005). Path integration in a three-dimensional maze: ground distance estimation keeps desert ants Cataglyphis fortis on course. Journal Of Experimental Biology, 208(21), 4005–4011. http://doi.org/10.1242/jeb.01873

Grah, G., Wehner, R., & Ronacher, B. (2007). Desert ants do not acquire and use a three-dimensional global vector. Front Zool Frontiers In Zoology, 4(1), 12. http://doi.org/10.1186/1742-9994-4-12

Gładziejewski, P. (2015). Predictive coding and representationalism. Synthese. http://doi.org/10.1007/s11229-015-0762-9

Hale, M., & Reiss, C. (2000). Phonology as cognition. In Burton-Roberts, N., Carr, P., & Docherty, G. J. (eds.). Phonological Knowledge: Conceptual and Empirical Issues. Oxford: Oxford University Press. 161-184.

Hale, M., & Reiss, C. (2008) The phonological enterprise. Oxford: Oxford University Press.

Helmholtz, H. von, & Southall, J. P. C. (1874/1962). Treatise on physiological optics: translated from the third German edition. New York, NY: Dover.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. Attention, Perception & Psychophysics, 72(5), 1218–1227. http://doi.org/10.3758/app.72.5.1218

Howe, C. Q., & Purves, D. (2005). The Muller-Lyer illusion explained by the statistics of image-source relationships. Proceedings Of the National Academy of Sciences, 102(4), 1234–1239. http://doi.org/10.1073/pnas.0409314102

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioral And Brain Sciences, 34(04), 169–188. http://doi.org/10.1017/s0140525x10003134

Kenstowicz, M. J. (1994). Phonology in generative grammar. Cambridge, MA: Blackwell.

Kirchner, R. (1997). Contrastiveness and faithfulness. Phonology, 14(1), 83–111. http://doi.org/10.1017/s0952675797003291

Kluender, K., Diehl, R., & Killeen, P. (1987). Japanese quail can learn phonetic categories. Science, 237(4819), 1195–1197. http://doi.org/10.1126/science.3629235

Kuhl, P., & Miller, J. (1975). Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants. Science, 190(4209), 69–72. http://doi.org/10.1126/science.1166301

Kuhl, P. K. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. The Journal Of the Acoustical Society of America, 70(2), 340. http://doi.org/10.1121/1.386782

Kuhl, P. K. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. The Journal Of the Acoustical Society of America, 73(3), 1003. http://doi.org/10.1121/1.389148

Levin, D. T., & Banaji, M. R. (2006). Distortions in the perceived lightness of faces: The role of race categories. Journal Of Experimental Psychology: General, 135(4), 501–512. http://doi.org/10.1037/0096-3445.135.4.501

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. Cognition, 21(1), 1–36. http://doi.org/10.1016/0010-0277(85)90021-6

Lotto, A. J., Kluender, K.R., & Holt, L.L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). The Journal Of the Acoustical Society of America, 102(2), 1134. http://doi.org/10.1121/1.419865

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. Proceedings Of the National Academy of Sciences, 110(35), 14196–14201. http://doi.org/10.1073/pnas.1303312110

Lupyan, G. (2015). Cognitive Penetrability of Perception in the Age of Prediction: Predictive Systems are Penetrable Systems. Review Of Philosophy and Psychology http://doi.org/10.1007/s13164-015-0253-4

Magnuson, J. S., Mcmurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: the ghost of Christmash past. Cognitive Science, 27(2), 285–298. http://doi.org/10.1207/s15516709cog2702_6

Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English 'l' and 'r.' Cognition, 24(3), 169–196. http://doi.org/10.1016/s0010-0277(86)80001-4

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. Perception & Psychophysics, 28(3), 213–228. http://doi.org/10.3758/bf03204377

Marcus, G. F. (2001). The algebraic mind: integrating connectionism and cognitive science. Cambridge, MA: MIT Press.

Marr, D. (1982). Vision: a computational investigation into the human representation and processing of visual information. San Francisco: W.H. Freeman.

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746–748. http://doi.org/10.1038/264746a0

Mendelovici, A. (2012). Reliable misrepresentation and tracking theories of mental representation. Philosophical Studies, 165(2), 421–443. http://doi.org/10.1007/s11098-012-9966-8

Menzel, R., Greggers, U., Smith, A., Berger, S., Brandt, R., Brunke, S., & Watzl, S. (2005). Honey bees navigate according to a map-like spatial memory. Proceedings Of the National Academy of Sciences, 102(8), 3040–3045. http://doi.org/10.1073/pnas.0408550102

Millikan, R. G. (1989). Biosemantics. The Journal Of Philosophy, 86(6), 281. http://doi.org/10.2307/2027123

Millikan, R. G. (1984). Language, thought, and other biological categories: new foundations for realism. Cambridge, MA: MIT Press.

Millikan, R. G. (1986). Thoughts Without Laws; Cognitive Science with Content. The Philosophical Review, 95(1), 47. http://doi.org/10.2307/2185132

Millikan, R. G. (1993). White queen psychology: and other essays for Alice. Cambridge Mass.: MIT Press

Miłkowski, M. (2013). Explaining the computational mind. Cambridge, Mass: MIT Press.

Mirman, D., Mcclelland, J., Holt, L., & Magnuson, J. (2008). Effects of Attention on the Strength of Lexical Influences on Speech Perception: Behavioral Experiments and Computational Mechanisms. Cognitive Science: A Multidisciplinary Journal HCOG, 32(2), 398–417. http://doi.org/10.1080/03640210701864063

Neander, K. (2006). Content for Cognitive Science. In Graham MacDonald and David Papineau (eds.). Teleosemantics. Oxford: OUP.

Neander, K. (2012). Toward and informational teleosemantics. In Ryder, D., Kingsbury, J., & Williford, K., (ed.s). Millikan and her critics. Malden, MA: John Wiley & Sons.

Neander, K. (forthcoming) manuscript on teleological semantics.

Newell, A. (1980). Physical symbol systems. Cognitive Science, 4(2), 135–183. http://doi.org/10.1016/s0364-0213(80)80015-2

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. Progress in Brain Research, 155, 23–36. doi:10.1016/S0079-6123(06)55002-2

Orlandi, N. (2014) The innocent eye: Why vision is not a cognitive process. Oxford: Oxford University Press.

Pettit, P. (1986) Broad-minded Explanation and Psychology. In Subject, thought, and context, Philip Pettit, John McDowell, Eds. Clarendon Press.

Piccinini, G. (2006). Computation without Representation. Philosophical Studies, 137(2), 205–241. http://doi.org/10.1007/s11098-005-5385-4

Piccinini, G. (2007). Computing Mechanisms. Philosophy Of Science, 74(4), 501–526. http://doi.org/10.1086/522851

Pietroski, P., (1992). Intentional and Teleological Error. Pacific Philosophical Quarterly, 73: 267–81.

Pietroski, P., & Rey, G. (1995) When Other Things Aren't Equal: Saving Ceteris Paribus Laws from Vacuity. The British Journal For the Philosophy of Science Br J Philos Sci, 46(1), 81–110. http://doi.org/10.1093/bjps/46.1.81

Pinter-Wollman, N., Bala, A., Merrell, A., Queirolo, J., Stumpe, M. C., Holmes, S., & Gordon, D. M. (2013). Harvester ants use interactions to regulate forager activation and availability. Animal Behaviour, 86(1), 197–207. http://doi.org/10.1016/j.anbehav.2013.05.012

Purves, D., Wojtach, W. T., & Lotto, R. B. (2011). Understanding vision in wholly empirical terms. Proceedings Of the National Academy of Sciences, 108(Supplement_3), 15588–15595. http://doi.org/10.1073/pnas.1012178108

Putnam, H. (1988). Representation and reality. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (1980). Cognitive representation and the process-architecture distinction. Behavioral And Brain Sciences, 3(01), 154. http://doi.org/10.1017/s0140525x00002302

Pylyshyn, Z. W. (1984). Computation and cognition: toward a foundation for cognitive science. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (2003). Seeing and visualizing: it's not what you think. Cambridge, MA: MIT Press.

Pylyshyn, Z. (1999). Vision and cognition: How do they connect? Behavioral And Brain Sciences, 22(3), 401–414. http://doi.org/10.1017/s0140525x99622026

Ramsey, W. M. (2007). Representation reconsidered. Cambridge: Cambridge University Press.

Reid, C. R., Latty, T., Dussutour, A., & Beekman, M. (2012). Slime mold uses an externalized spatial "memory" to navigate in complex environments. Proceedings Of the National Academy of Sciences, 109(43), 17490–17494. http://doi.org/10.1073/pnas.1215037109

Rescorla, M. (2013). A theory of computational implementation. Synthese, 191(6), 1277–1307. http://doi.org/10.1007/s11229-013-0324-y

Rescorla, M. (2015a) Review of The Innocent Eye. Notre Dame Philosophical Reviews. Retrieved September 7, 2015, from https://ndpr.nd.edu/news/55073-the-innocent-eye-why-vision-is-not-a-cognitive-process/

Rescorla, M. (2015b) Bayesian perceptual psychology. In M. Matthen (ed.), The Oxford Handbook of Philosophy of Perception., Oxford: Oxford University Press, 694-716.

Rescorla, M. (forthcoming) Bayesian Sensorimotor Psychology. Mind and Language. http://www.philosophy.ucsb.edu/docs/faculty/michael-rescorla/rescorla_bayesian-sensorimotor-psycholoy.pdf

Rey, G. (1996). Resisting Primitive Compulsions. Philosophy And Phenomenological Research, 56(2), 419. http://doi.org/10.2307/2108533

Rey, G. (2003). Representational content and a Chomskyan linguistics. In A. Barber (ed.). Epistemology of Language. Oxford: Oxford University Press, 140-86

Rey, G. (2006) The non-existence of language-- but not cars. In R. Stainton (ed.). Contemporary Debates in Cognitive Science. Oxford: Blackwell, 237-255.

Rey, G. (2008) In defense of folieism. Croatian Journal of Philosophy, 8, 177-202.

Rey, G. (2009) Concepts, defaults, and internal asymmetric dependencies: Distillations of Fodor and Horwich. In N. Kompa, C. Nimtz, and C. Suhm (eds.). The A Priori and Its Role in Philosophy. Paderborn: Mentis, 185-204.

Rey, G. (2012) Externalism and inexistence in early content. In Schantz, R. (ed.), Prospects for Meaning. Berlin: De Gruyter.

Rock, I. (1983). The logic of perception. Cambridge, MA: MIT Press.

Searle, J.R. (1990). Is the brain a digital computer? Proceedings and Addresses of the American Philosophical Association 64: 21-37

Searle, J. R. (1992). The rediscovery of the mind. Cambridge, MA: MIT Press.

Shagrir, O. (2001). Content, Computation and Externalism. Mind, 110(438), 369–400. http://doi.org/10.1093/mind/110.438.369

Silverberg, A. (2006). Chomsky and Egan on computational theories of vision. Minds And Machines, 16(4), 495–524. http://doi.org/10.1007/s11023-006-9050-2

Steck, K., Wittlinger, M., & Wolf, H. (2009). Estimation of homing distance in desert ants, Cataglyphis fortis, remains unaffected by disturbance of walking behaviour. Journal Of Experimental Biology, 212(18), 2893–2901. http://doi.org/10.1242/jeb.030403

Stich, S. P. (1978). Beliefs and Subdoxastic States. Philosophy Of Science, 45(4), 499. http://doi.org/10.1086/288832

Stich, S. P. (1983). From folk psychology to cognitive science: the case against belief. Cambridge, MA: MIT Press.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. Nature Neuroscience, 5(6), 598–604. http://doi.org/10.1038/nn858

Whiting, J. G., Costello, B. P. D. L., & Adamatzky, A. (2014). Sensory fusion in Physarum polycephalum and implementing multi-sensory functional computation. Biosystems, 119, 45–52. http://doi.org/10.1016/j.biosystems.2014.03.003

Witt, J. K., Proffitt, D. R., & Epstein, W. (2004). Perceiving distance: A role of effort and intent. Perception, 33(5), 577–590. http://doi.org/10.1068/p5090

Wittlinger, M., Wolf, H., & Wehner, R. (2007). Hair plate mechanoreceptors associated with body segments are not necessary for three-dimensional path integration in desert ants, Cataglyphis fortis. Journal Of Experimental Biology, 210(3), 375–382. http://doi.org/10.1242/jeb.02674

Woods, A. J., Philbeck, J. W., & Danoff, J. V. (2009). The various perceptions of distance: An alternative view of how effort affects distance judgments. Journal Of Experimental Psychology: Human Perception and Performance, 35(4), 1104–1117. http://doi.org/10.1037/a0013622

Yang, C. (2002). Knowledge and learning in natural language. Oxford: Oxford University Press.