November 2016

Structuring Thought: Concepts, Computational Syntax, and Cognitive Explanation

Matthew B. Gifford
University of Massachusetts Amherst

STRUCTURING THOUGHT: CONCEPTS, COMPUTATIONAL SYNTAX, AND COGNITIVE EXPLANATION

A Dissertation Presented

by

MATTHEW B. GIFFORD

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2016

Philosophy

STRUCTURING THOUGHT: CONCEPTS, COMPUTATIONAL SYNTAX, AND COGNITIVE EXPLANATION

A Dissertation Presented

by

MATTHEW B. GIFFORD

Approved as to style and content by:

Louise Antony, Chair

Erik Cheries, Member

Joseph Levine, Member

Hilary Kornblith, Member

Joseph Levine, Department Chair
Philosophy

DEDICATION

To my wife, Anne

ACKNOWLEDGMENTS

It is widely agreed among those concerned that this dissertation has been a *long time coming*. I have incurred some serious debts of gratitude along the way.

This project can trace its origins to a brief conversation with Louise Antony. Recently returned from a conference, she remarked in frustration, "everyone is talking about empiricism again!" The subtext being that this situation could not stand, and I should do something about it. No matter how inchoate or confused my ideas, when I brought them to Louise I found them charitably and thoughtfully received. Any clear ideas in this dissertation were ground from rough stone by my discussions with her. Beyond remedying my confusions, our discussions would leave me feeling that my ideas were worth pursuing, and encouraged to do so. More than an advisor, Louise has been a friend and mentor, and I am a better thinker and person for knowing her. Thank you, Louise.

I am grateful to all of my committee members. Hilary Kornblith, Joe Levine, and Erik Cheries have all helped me out of jams both intellectual and practical. This dissertation has been much helped by their ideas, criticism, and suggestions. I want to extend a special thanks to Hilary, who I have tapped for an enormous amount of support, ideas, and advice over the years.

Thank you to Georges Rey, who got me into this whole mess in the first place, and to whom I owe a great intellectual debt. Most of the topics discussed in this dissertation I first encountered during our frequent meetings when I was an undergraduate at the University of Maryland. I would also like to thank Eileen O'Neill and Lynne Baker, who helped me find my philosophical voice early on in my graduate career.

I thank the UMass philosophy department for their generous funding from several sources including two fellowships. Thanks to Beth Grybko, without whom I'm pretty sure we would all fall into disarray.

My parents, Sarah and David, I thank for their unending support and encouragement over the years.

And finally, thanks to the person who copyedited this entire document: my wife, Anne Cecelia Holmes. You, dear reader, can thank her for every typo that is not here. (Any that are, I made after her careful revisions and against advice.) Anne has been my emotional support, best friend, and partner for my entire grad school life, and I genuinely could not have done this without her. I also may never have finished if she hadn't let me know that *it was time to be done*. I agree. Thank you, Anne.

ABSTRACT

STRUCTURING THOUGHT: CONCEPTS, COMPUTATIONAL SYNTAX, AND COGNITIVE EXPLANATION

SEPTEMBER 2016

MATTHEW B. GIFFORD

B.A., UNIVERSITY OF MARYLAND, COLLEGE PARK

M.A., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Louise Antony

The topic of this dissertation is what thought must be like in order for the laws and generalizations of psychology to be true. I address a number of contemporary problems in the philosophy of mind concerning the nature and structure of concepts and the ontological status of mental content. Drawing on the empirical literature, I develop a number of new conceptual tools for theorizing about concepts, including a counterpart model of concepts' role in linguistic communication, and a deflationary theory of concepts' formal features. I also suggest some new answers to old problems, arguing, for example, that content realism is not hostage to a naturalized semantics.

This dissertation can, as a whole, be read as a sympathetic re-evaluation of the *language of thought hypothesis* (LOT), a view most famously associated with the philosopher Jerry Fodor, who has been its most ardent champion for decades. LOT claims that thoughts are sentences in a mental language of computation, and are

composed of meaningful, atomic symbols—concepts—which are individuated entirely by their formal features. Each chapter either defends various components of LOT from recent criticism, or fills in gaps in the theory.

I begin the dissertation by introducing the language of thought project, and motivating and explaining each of its central components. I provide the necessary background for understanding the controversies surrounding LOT by explaining the computational/representational theory of mind. While these are now relatively uncontroversial, LOT's commitments to concept atomism and content realism are the subject of much debate. In chapter 2, I discuss a recent competitor to atomism—neo-empiricism—and argue that it fails to meet several key desiderata on theories of concepts, and defend atomism from similar charges. In chapter 3, I argue against a view common to both philosophy and psychology: that concepts must be *shareable*. If true, atomism is in jeopardy. I find shareability to be unmotivated, and suggest an alternative counterpart model of concepts that does a better job of explaining the things shareability was supposed to explain. Chapter 4 takes up the question of what formal features are and how mental symbols are individuated. I develop a reductive account, arguing that formal features are certain sorts of physical properties and symbol types are sets of these properties. I turn to the topic of content in chapter 5, and defend a very strong version of content realism against recent criticism.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION: THE STRUCTURED MIND

1.1 The armchair reader

Human thought is a very powerful tool. We can think about a great many things—things we perceive, and things we do not; things that have happened, things that will happen, and things that never come to be; things that exist, things that could exist but do not, and things that cannot exist at all. We can reason with thought. Thoughts can influence our behavior. And there seems to be no limit—apart from the constraints of time—to the number of new and different thoughts we can think.

Perhaps our greatest resource for discovering what thought must be like to explain all of these things is the empirical, scientific study of the mind—psychology. Since modern cognitive psychology's birth in the second half of the twentieth century, the field has provided invaluable insight into the rules and laws that govern the mind's operations. This dissertation is not, however, an empirical work. I have conducted no experiments, collected no data. But neither is it armchair psychology or philosophy; it is informed and constrained by empirical findings. It is about is what thought must be like in order for the laws and generalizations of psychology to be true.

This dissertation addresses its topic by providing a sympathetic re-evaluation of the *language of thought hypothesis* (LOT). LOT makes the following claims. Thoughts—mental states that express propositions—are sentences in a mental computational language. They are constructed from concepts, representations that have intentional content (they refer to things) and are individuated by their formal (non-semantic) features. Concepts themselves are structureless atoms. They are compositional—they

can be combined together to form new complex representations, the meaning of which is wholly determined by the meanings of its parts and their rules for combination. The rules governing the mind's operations are defined over only the formal features of mental representations.

As anyone familiar with LOT knows, it is the brainchild of the philosopher Jerry Fodor, who introduced it, developed it, and has been its most ardent champion for decades. Fodor (1975) advanced a seminal defense of the idea that thoughts are computational/representational structures. Fodor (1998) developed the idea of concept *atomism*—that concepts are structureless mental objects. And Fodor in his (1987, 1992a, and 1995) argues that intentional content plays an important role in psychology, and that content determination can be the product of a naturalistic relation between the mind and the world. While Fodor was not the first or only person to defend each of these views (except, perhaps, for atomism), he was the first to unify them in a single theory.

Though widely discussed, LOT has gained only minor acceptance in both philosophy and psychology. This is partly because it is misunderstood.[1] But it also faces a number of very serious, current challenges. Atomism in particular is under constant fire. Psychologists seem to have landed on versions of prototype theories[2], exemplar theories[3], theory-theories[4], or neo-empiricist theories[5] instead. Defenders of these

---

[1]See, for example, chapter 2, §2.4.

[2]Prototype theories hold that concepts are structures of statistically typical features of the categories they represent. c.f. Rosch (1978).

[3]Exemplar theories identify concepts with (collections of) examples of category members (e.g., a *dog* concept is an image of a typical dog). c.f. Medin et al. (1982).

[4]Theory-theories claim that concepts are sets of representations linked together to form a mini-theory of what they represent. So a *water* concept might contain representations of *water being a natural kind*, *water being wet*, etc. c.f. Carey (2009).

[5]According to neo-empiricists, concepts are structures of sensory symbols. c.f. Barsalou (1999) for a defense of this view, and chapter 2 of this thesis for a critical discussion.

views argue that atomism fails to meet some or other desiderata on a theory of concepts. Content realism is perennially called into question. Egan (2014) has recently argued that it is not needed in cognitive science, though content ascription is a necessary heuristic for explanation. Various authors have criticized "formal features" as being an incoherent notion, and that individuating concepts by these features renders LOT unable to explain certain facts about our mental lives. There is also very little said about what formal features actually are, so most discussions of them remain metaphorical, to the detriment of LOT.

My goal is to remind us of LOT's attractions and address recent criticisms. Each chapter addresses a different puzzle or challenge facing LOT. Together, they provide a new defense of some old ideas, and introduce a number of new conceptual tools for those trying to understand the nature of thought.

At the end of this chapter, I will provide an overview of each subsequent chapter's project. First, I figure I ought to say something about why LOT is worth taking seriously. I will introduce and motivate each of its four main components: the representational theory of mind, the computational theory of thought, content realism, and concept atomism.

1.2 Representationalism

The representational theory of mind (or *representationalism*) holds that thoughts and propositional attitudes are relations between people and mental representations. To believe that p, according to representationalism, is to bear the belief relation to a mental representation of p (*mutatis mutandis* for the other propositional attitudes).

There are a number of reasons that we should accept representationalism. One is that it vindicates our ordinary folk psychology. We might, for example, say that Mary went to the water fountain because she desired water and believed that going to the fountain was the best way to get it. These belief/desire attributions have proven

extremely useful for predicting and explaining behavior. This strongly suggests that they are (often) true. If so, then there must *be* beliefs and desires, and they must be about things. Representationalism goes some distance towards explaining what this would mean: to believe that Mary wants water is to bear the belief relation to a mental representation with the content *Mary wants water*.

Representationalism also provides a framework for understanding the causal relationship between a thinker and the objects of thought. My currently unseen, distant cat plays a role in getting me to buy cat food on the way home. The (non-obtaining) state of affairs of *Mary drinking water* plays some role in causing Mary to go to the fountain. An imaginary barber's absurd standards for clientèle famously played a role in getting Russell to write a letter to Frege. How do these things—from the mundane, to the non-actual, to the impossible—influence our thought and action? According to representationalism, they do not. At least, not directly. Instead there are proxies for these things: mental particulars representing them. What enters into the causal chain is a mental representation of my cat, of Mary drinking water, or of an impossible barber.

Representationalism raises a number of questions. If it is true, we should want to know what sorts of things mental representations are, how thinkers are related to the representations, and how representations are related to their referents. The (classical) computational theory of thought (or *computationalism*) is an attempt to answer the first two of these questions.

1.3 Computationalism

Computationalism holds that thought is computation. Mental representations are symbols, and thinking is a process of symbol manipulation. More specifically, it is a process that responds to the symbols' non-semantic features in such a way as to respect their semantic relations.

Unpacking this we can see that computationalism is committed to three main claims: thought is a process of symbol manipulation; mental symbols have semantic and non-semantic features; and thought responds only to the non-semantic features of these symbols.

The idea behind the first claim is simple enough. Token representations can causally interact with the world only if they are concrete entities. And a symbol just is a concrete entity with a semantic interpretation. If we accept that representational states are causally efficacious, then we ought to accept that representations are symbols. So, if the process of thought consists in manipulating representations, then thought consists in manipulating symbols.

Thoughts themselves are symbolic structures. They are like sentences in a computational language. They are composed of primitive computational symbols, analogous to words, which have their own semantic and non-semantic features, and contribute these features to the more complex representations. Primitive symbols are combined according to rules of combination, defined over their non-semantic features. So, for example, the thought that Doug is dreaming might be a complex representation—⟨Doug is dreaming⟩—with primitive symbols ⟨Doug⟩ (meaning *Doug*) and ⟨dreaming⟩ (meaning *dreaming*) as its constituent parts, combined together according to the mind's rules so that one is predicated of the other.

Symbolic structure is a key point of contrast between computationalism (or *classical* computationalism) and non-classically computational models, like connectionist networks. Non-classical models treat thoughts as *sui generis* computational kinds.[6] Computationalists might hold, for example, that a belief that chocolate is tasty is a structured symbol like this: ⟨chocolate is tasty⟩. This, they claim, has meaningful constituent parts: ⟨chocolate⟩ and ⟨tasty⟩. By contrast, a non-classicist might hold

[6]See Smolensky (1988) for a description of a connectionist computational model that works along these lines.

that the belief is itself a symbol, but has no independently meaningful parts. Consider as an analogy the English word 'chocolate'. The word has parts—'hoco', e.g.—but these parts are not constituents—they are not independently meaningful symbols in English which contribute their meaning to the meaning of 'chocolate'. On at least some non-classical views, attitude states are like this too.[7]

This fact about connectionist models has led Fodor and Pylyshyn (1988), Fodor and McLaughlin (1990), and McLaughlin (1993) to object to them on the grounds that human thought exhibits certain systematic regularities that are only predicted and explained by classical computational structures. To borrow an example from McLaughlin (1993), a human with the capacity to believe that the dog is chasing the cat *ipso facto* has the capacity to believe that the cat is chasing the dog. The classical model predicts this. Since both of the relevant beliefs use exactly the same symbols and rules of composition, a classically computational system with the resources to entertain one of these beliefs has all the resources required to entertain the other.[8] By contrast, it is possible for connectionist systems to entertain one but not the other. Since (roughly) the beliefs are taken as unstructured objects, these beliefs need not have any meaningful parts in common, nor are they composed by a system of rules defined over such elements. So, nothing about the capacity of such a system to entertain one implies the capacity to entertain the other. More recent connectionist models, like that of Frank et al. (2009), explicitly attempt to overcome this problem. While these models do respect systematic regularities in certain limited circumstances, unlike classical computationalist models they do not *predict* systematic regularities, but rather must be designed to *accommodate* them.

---

[7]In particular, Smolensky's work in the 1980s and 90s (*op. cit.*). But, as Hadley (2004) argues, the following criticism applies to more recent work as well (e.g., Bodén and Niklasson (2000)).

[8]Though see Hadley (1997) for criticism of the claim that classical systems can, by their nature, capture this type of systematicity. See Aizawa (2003, ch. 5) for a response to Hadley's arguments and (qualified) defense of classicism.

Computationalism's second claim is that mental symbols have both semantic and non-semantic features. As representations, mental symbols must have a semantics. There is debate about just what it means for symbols to "have" a semantics. In chapter 5, I argue for a strong realism about the content of mental representations. For now, it will suffice to say that part of what distinguishes symbols from everything else in the world is that they are representational.

Mental symbols must also have *non*-semantic (/formal) features. Clearly, as concrete entities, symbols have all sorts of non-semantic properties (their location and size, e.g.). The interesting claim is that lexical types of mental symbols are *individuated* by their non-semantic properties. The mind needs to know what sorts of symbols it has—when two symbols are the same lexical type, and when they are distinct. And it can't do this by looking at their semantic properties. So, mental bookkeeping works by keeping track of the formal features of mental symbols. Whether two symbols are the same lexical type depends on whether they share all of their formal features.

Non-semantic symbol individuation leads to computationalism's key insight: that physical systems can respect semantic relations by responding to non-semantic properties.

Our thoughts exhibit semantic regularities. Seeing snow outside causes me to believe that it's snowing and therefore cold, which causes me to desire that I get my jacket before going outside. The relationships between a belief that it's snowing, a belief that it's cold, and a desire for a coat are semantically intelligible. They reflect relations between snow, cold, and jackets.

But meaning *that it's snowing* is not a causally efficacious property. So how do the causal relations between thoughts, *qua* physically realized mental objects, manage to obey semantic relations between what they represent? According to computationalism, thought exploits the non-semantic features of symbolic structures to capture semantic relations. While a physical system might not be able to respond to the

meaning of 'snow', it could respond to the fact that 'snow' is so written. A system could be entirely blind to semantic features, but have a rule that says "if you see the shape 'it's snowing' written here, write the shape 'it's cold' there."

Thought, on the computationalist view, works via the same process. Mental processors obey rules for manipulating symbols that are defined over their non-semantic features, but respect the semantic relations of their representational content. That the mental symbol ⟨snow⟩ means *snow* is not available to the mind. The fact that it "looks like" '⟨snow⟩', however, is. The semantic relations between snow and cold are born out in symbolic relations between ⟨snow⟩ and ⟨cold⟩.

To recap: Thoughts are sentences in a mental language of computation (*mentalese* or *the language of thought*). They are made up of word-like symbols, which have content and contribute this to the meaning of the sentences they compose. The rules that govern how they can combine to form new complex representations are defined over their formal features. Thinking, on this view, is a process of manipulating symbol structures. To infer that *it is cold outside* on the basis of the beliefs that *it is snowing*, and *if it is snowing, then it is cold outside*, is to follow a rule that says you can construct a complex symbol that looks like ⟨cold outside⟩ if you bear the belief relation to symbols that look like ⟨snowing⟩ and ⟨if snowing, then cold outside⟩. These rules are formal, and make no mention of content. One such rule might say, for example, "infer Q from P and *if P, then Q*" where 'P' and 'Q' range over categories of symbols, and stand for formally individuated expressions.

1.4 The structure of concepts

Concepts are the symbols that compose thoughts. They are the computational symbols described above, the analogs to words in mentalese.

Explanatory role of concepts

Concepts are thought to play a number of explanatory roles within cognitive science.[9] There are three demands on concepts that will be at issue in later chapters.

Compositionality: Fodor and his followers have been instrumental in drawing attention to concepts' role in explaining the *productivity* and *systematicity* of thought.[10] 'Productivity' refers to our capacity, in principle, to entertain and understand an unbounded number of novel thoughts. If we can think of a *show* and think of a *pre-show* then we can think of a *pre-pre-show*, a *pre-pre-pre-show*, and so on. We manage to do this despite having finite representational capacities; we cannot store an infinite number of unique representations in our heads. Likewise, we can understand new thoughts. Despite never before having read the sentence "the UMass library is closing due to its books becoming sentient and walking out," you understand it. The systematicity of thought I described briefly above: the fact that our ability to understand and entertain certain thoughts entails the ability to understand and entertain certain other thoughts.

To explain productivity and systematicity, concepts are thought to be *compositional*. They can combine together to form new complex concepts, and what these new concepts mean is determined entirely by the meanings of their constituent concepts and the rules of combination. That means that as long as you understand all of the constituent concepts in a thought, you will be able to understand the thought. And, if the rules of composition are recursive, you can generate an infinite number of well formed expressions in mentalese from a finite base.

Publicity: Almost all theorists believe that concepts can be and are shared between different people. Shared concepts are thought be required in order to enable

---

[9]Different formulations of these same explanatory demands on a theory of concepts can be found in Carey (2009, pp. 489–91), Fodor (1998, ch. 2), Prinz (2002, pp. 3–16), and Rey (1983, p. 241).

[10]c.f. Fodor (1975, 1998); Laurence and Margolis (1999); Rey (1995).

successful linguistic communication—we can communicate about moose by using the word 'moose' because you and I share a concept ⟨moose⟩ and map it onto 'moose'.

We also seem to need shared concepts to capture psychological generalizations. Take for example recent work by Huang and Bargh (2014), who argue that behavior is best explained by modeling action-guiding cognition as a structure of competing goals (mental representations of desired end-states). Their model predicts that individuals' goals "constrain the individual's information processing and behavioral possibilities in a way that encourages achievement of the goal's end-state" (127). As evidence that this does in fact occur, they point to studies showing that when competing groups are given a shared goal, the group's perceptions of the other group are transformed (130). To make such explanations work, it seems that participants in these studies need to share goals—that is, to share representations of end-states. This in turn would seem to require shared concepts.

Because of considerations like these, many theorists endorse what is called the *publicity* or *shareability* constraint on a theory of concepts: whatever concepts are, they must be able to be shared by different people.[11]

Frege cases: Oedipus very much wanted not to marry Mom. Yet, forced by no one, he did. His misfortune was caused by the fact that he was in a *Frege case*. Frege cases, roughly, are situations in which someone fails to realize that two terms co-refer. They are named for the logician and philosopher of language Gottlob Frege, who raised the question of why substitutivity—the principle that co-referring terms can be substituted for one another *salva veritate*—seems to fail in belief reports (Frege (1892)). For example, it seems true to say that "Oedipus wants to marry Jocasta," but not "Oedipus wants to marry Mom" despite 'Jocasta' and 'Mom' referring to the

[11] c.f. Fodor (1998, pp. 28–34); Fodor and Lepore (1992); Laurence and Margolis (1999) (in terms of *stability*); Prinz (2002, pp. 14–16); Rey (1983, p. 282) (in terms of "interpersonal stability functions").

same person. Frege cases are rather common. One might want to buy chickpeas, but fail to do so because they are labeled 'garbanzo beans', for example.

The going thought is that such cases arise because people can have multiple concepts for the same thing, and no way of telling *a priori* that they co-refer. Oedipus is led to his doom because he has a ⟨Mom⟩ concept and a ⟨Jocasta⟩ concept, but nothing he knows clues him into the fact that their referents are one and the same.

The implication, then, for a theory of concepts, is that it must make such situations possible. Concepts need to be individuated more finely than reference alone allows. Just how this is done is a matter of controversy. Co-referring concepts might, some think, still be semantically distinct—there is more to meaning, perhaps, than reference. Other argue that concepts should be individuated by their causal role. If so, co-referring concepts would be distinguished by their taking part in different inferences, beliefs, etc. Still others hold that co-referring concepts are distinguished by their non-semantic features, analogous to the way 'camel' and 'dromedary' are distinguished in written language.

Atomism, LOT's theory of concepts, takes this last road.

Atomism

Atomism is a theory about the structure of concepts—namely, that they do not have any. This view, first developed by Fodor (1998), says that lexical concepts (concepts that, roughly, correspond to words in a natural language) are structureless atoms that carry informational content. They are definitionally and developmentally primitive. They are not made up of other concepts, features, or other semantically evaluable parts. They do not depend on any other concepts for their content.

One of the main appeals of this view is that it is amenable to a causal-informational semantics. Most competing theories hold that concepts' content is somehow determined by their structure. The *prototype theory*, for example, roughly holds that con-

cepts are a list of the features typically had by members of the category the concept represents. So ⟨chicken⟩ might be made up of features like ⟨has feathers⟩, and ⟨has a beak⟩. The problem is that this does not capture the way our semantic judgments work: plucked chickens are still chickens. Further, as Keil (1989) argues, prototype theories fail to predict or explain data concerning category membership under change. We judge, for example, a creature who starts life as a cat but is altered to look and act like a skunk—satisfying all the prototypical features of skunks, and not of cats—to nonetheless be a cat and not a skunk.

Atomism avoids these problems by holding that all there is to being a concept with the content *chicken* is to be causally related to the property of being chicken in the right sort of way.[12]

Atomism is also attractive because it lends itself readily to being compositional—which is necessary for explaining the productivity and systematicity of thought. Since concepts are primitive expressions, and rely on external relations for their content, it is relatively simple to see how atoms can combine to form new, complex, meaningful, intelligible representations. It's roughly the same way that words form sentences or, more aptly, how primitive expressions can be combined to form complex expressions in a formal language of logic.

By contrast with atomism, prototype theories struggle to be compositional. As Fodor (1998) argues, a prototypical pet (e.g. a furry mammal) and the prototypical fish (e.g. something that lives in the ocean) do not combine to generate the features of the prototypical pet fish (e.g. a goldfish).

---

[12]Note that not all of atomism's competitors face this sort of problem. Notably Carey (2009) argues for a theory-theory according to which concepts have a dual-factor semantics, which imbue concepts with two dimensions of meaning. One is an informational semantics, the other is determined by the structure of the concept. This latter factor is supposed to explain things like category judgments and so-called "cognitive content." The problem for this theory is that neither of these are things that concepts need to explain. While concepts must be able to *refer* to categories, they mechanisms that allow us to *recognize* category instances may be external to the concepts for these categories. And, as I will argue in the next chapter, "cognitive content" does not exist.

1.5 Content realism

If the non-semantic features of mental symbols do all of the causal work, as computationalism claims, is there a place for mental content in a scientific theory of the mind? Some cognitive scientists, notably Stich (1983) and Chomsky (1995), argue there is not. Talk of mental content, they argue, can be eliminated from our explanatory vocabulary. I will not address criticisms of realism in this section. In chapter 5 I argue for a strong realism about mental content—that mental representations not only have content, but that their content constitutes part of their type identity within cognitive science. Here I present a *prima facie* motivation for the position: that actual explanations in the cognitive sciences appeal to mental content, and that such appeals are useful.

One example from the developmental literature is Leslie and Keeble's (1987) landmark study using looking time experiments to investigate how infants perceive causal events. They habituated six-month-old infants to two scenarios: one in which a moving object, A, contacted a stationary object, B, which then moved; the other in which A moved, and then B moved, but there was no contact. The researchers then presented the infants with the scenarios played in reverse (B contacts A, which then moves; B moves first and does not contact A, then A moves). They found that the infants looked longer at the reversal of the first scenario than the reversal of the second, indicating that they perceived the former as more novel.

Why would there be a difference in perceived novelty given that just as much has changed visually in both? Leslie and Keeble (1987) argued that we can explain this if the infants took only the first event to show causation. Taken as a causal event, the contact reversal included not only a reversal of event order—which the other scenario also had—but an additional *role reversal*, with A changing from causal agent to causal patient (and *vice versa* for B). Because the infants took there to be more differences

between original and reversed contact events than the reversed non-contact events, the former were more novel for them than the latter.

For this explanation to be true, infants must represent the action in the scenario as *causation* and the objects in the scenario as *causal agents* and *causal patients*. Further, the content of these representations was crucial to Leslie and Keeble's (1987) prediction. They were able to reason that these additional changes would occur in infants' representations of these events because (*pace* Hume) these additional changes *exist* in the reversal of causal events. That is, they reasoned from a relation that holds between the represent*ed* to a relation that holds between the representations.

The developmental literature teems with this sort of explanation. As a another example, take Ganea et al. (2009) who (building on the work of Preissler and Bloom (2007)) show that children learning a new word by associating it with a picture of an object will extend the word to the object itself, even when the object is perceptually distinct from its depiction (e.g., it is a different color). This suggests that young children (15 months) understand pictures symbolically—as depictions of a thing. This ability to distinguish between picture and object, and to understand the picture is *of* an object, seems to require that infants have concepts referring to things they have never directly encountered, and represent pictures as being of those things.

We can find a different sort of appeal to content in the cognitive neuroscience literature. In one example, Fischer et al. (2011) used an fMRI to study activation patterns in higher-level visual areas of the brain in situations where subjects made errors in reporting an object's location. The researchers identified which patterns of brain activation corresponded with correct reports (e.g., the object was in location 1, and that's where the subject indicated it was). They then presented subjects with rapid changes in object location, which generated errors in reported location (e.g., the object was in location 2, but the subject indicated it was in location 1). What they found was that when subjects incorrectly reported an object as being

in a particular location, the activation patterns of the visual areas were the same as when the subject correctly reported an object as being in that location. So, if the object was in location 2 but the subject indicated that it was in location 1, the subject's visual system would display the activity that corresponded to correct location 1 reports. (They conducted similar tests with similar results comparing the retinal location versus reported location of the object.)

Explaining their findings, Fischer et al. (2011) conclude that "changes in the patterns of activity in each of these higher level areas were more tightly coupled with changes in *perceived position* than in physical position" (p. 120, emphasis added). By saying that these areas are correlated with a perceived position, they are suggesting a correspondence with mental content. The subject represents the object and various locations, and represents the object as being in those locations. If we wanted to predict the activation patterns in these areas, we could not look at the data about where the object was (or its relative retinal location). What we would want to know is where the subject *thought* it was. That is, we need to talk about where the subject represents the object as being—an intentional characterization—to specify the unifying correlate for the observed brain activity.

As a final example, consider a classic case in the psychology of judgment. Using a number of clever experiments, Kahneman and Tversky (1979) show that we are loss *averse*—that the (negative) value we put on losing something is greater than the (positive) value we put on gaining that same thing. So, for example, the price with which I would be willing to part with a coffee mug is higher than the price I would be willing to pay to acquire the coffee mug. Loss aversion interacts with a number of other biases, including what are known as framing effects. In a different study, Tversky and Kahneman (1981) gave subjects a story about a disease that would kill 600 people, and presented them one of two sets of choices. In one set, subjects were asked to choose between a course of action that would save 200 lives, and a course of

action that would have a one-third probability of saving 600 people and a two-thirds probability of saving no one. The other set of choices were between a course of action in which 400 people would die, or one in which there was a one-thirds probability that nobody would die and a two-thirds probability that 600 people would die. Though the dilemma is the same in both cases in terms of objective results, 72% of people presented with the first set chose the first option (saving 200) while 78% of people presented with the second scenario chose the other outcome (one-third chance that no one will die). What this shows is that we are sensitive to how a scenario is framed. In particular, if something is framed as a loss, we will treat it differently than if the very same thing is framed as a gain.

The point I want to make is that loss aversion can only explain the difference in behavior in the framing scenario if we assume that we represent scenarios as *losses* and *gains*. There is not an objective (external) difference in the scenarios. There is only a difference in how we think about them. And to explain the difference in how we think about them, Tversky and Kahneman (1981) subsume us by the content of our mental states.

In all of the above examples, theorists make use of mental content as a theoretical posit. Content is a part of our scientific parlance, not just our unscientific (though useful and predictive) folk psychology. This gives us at least *prima facie* reason to take this talk seriously, to be realists about content. The onus is on anti-realists to show that we can and should engage in the revisionary project of doing away with content in cognitive science.

1.6 The rest of the project

LOT is, at the very least, a promising research program. Whether or not it actually *works* is what I will be exploring in the rest of this dissertation. I want to now briefly say how I will proceed.

Neo-empiricism and reference

Chapter 2 takes up a recent challenge to concept atomism, *neo-empiricism*. Neo-empiricism is the view that concepts are structures of percepts—mental representations in the proprietary vocabulary of our sensory systems. I argue that neo-empiricist theories face a number of problems regarding their approach to reference. Most of the theories do not explain how semantic content is determined, and of those that do, most rely on a naïve "resemblance theory." There is one, however, that clearly states a philosophically sophisticated theory of how concepts get their content. Prinz (2002, 2005) advances a neo-empiricist view, *proxytype theory*, which makes use of a causal-informational theory of content very similar to the one atomism uses. He also levels several direct charges against atomism.

Thorough though it is, I argue that Prinz's view has some glaring holes. The details of the theory do not work well with an information semantics, leading it to make incoherent or errant content attributions. Its referential problems also prevent concepts of the same lexical types to be recognized as such, and allow for indiscernible concepts of different semantic and lexical types. Proxytype theory also fails to be systematic because its theory of composition does not allow *decomposition*—complex concepts cannot be broken into constituent parts, meaning that someone can think of, e.g., a *pet fish* without being able to think of *fish* or *pets*. All of these problems are surmountable, I argue. The problem for Prinz is that the changes he needs to adopt to do so render his theory functionally indistinguishable from atomism.

I also confront Prinz's (2002) criticisms of atomism: that it cannot explain cognitive content, and that "formal features" cannot do all of the work atomists say that they do. Atomism can, I argue, explain all of the things that cognitive content is supposed to explain. It just does so without complicating concepts' semantics. And, indeed, it does so using resources that *every* theory of concepts has available. The

upshot is that "cognitive content" is not something a theory of concepts needs to explain.

Against the shareability of concepts

In chapter 3, I challenge an assumption common to both philosophy and psychology: that concepts, whatever they are, must be shareable. This "shareability" or "publicity" constraint is a problem for atomists. There are very good reasons for thinking that atomic concepts, individuated by their formal features, cannot be shared. Instead of reconciling atomism with the shareability constraint, I argue the constraint should be abandoned. I consider the two broad reasons theorists offer in support of this claim: that shared concepts are needed to explain linguistic communication, and that shared concepts are required for capturing psychological generalizations. I find that neither actually needs shared *concepts*. At most they require shared *content*, which is not the same. Indeed, if shared concepts work the way theorists say they do in explaining communication, we are actually led to a logical contradiction.

Ridding us of shareability does leave a hole in the theory. We still need to say something about concepts' role in communication. I advance a *counterpart model* of concepts, which sets out conditions for distinct concepts to be counterparts to one another using resources implicit in all theories of concepts.

Mental bookkeeping in the language of thought

The previous two chapters make much of the explanatory ability of concepts' *formal features*. In chapter 4, I offer a theory of what formal features are, and how lexical types are individuated in the language of thought. I argue for a reductive view: that formal features are processor-detectable physical properties, and lexical types are sets of such properties. My argument is that the primary job of formal features is to allow the mind to recognize when token symbols are the same lexical

type and when they are not, and physical properties are the only thing that can do the job.

I consider a number of objections to the view, and the only worked-out competitor on the market: Schneider's (2011) theory that formal features are computational roles. I argue that the computational role view has a number of serious problems. Mental processors cannot detect roles, rendering them unable to distinguish and recognize lexical types. The theory is also *ungrounded*—without some prior criteria for symbol individuation, there is nothing that determines which computational roles exist and so, for Schneider, which symbol types exist. Finally, the theory results in false equivocations, saying that symbols of different lexical types are the same.

Content realism and hyper-representationalism

Chapter 5 shifts focus to the semantics of mental symbols. LOT claims that mental representations really do have content—that both their having content, and what that content is are matters of objective fact. Recent work in philosophy suggests that this may not be the case. Egan (2014) takes a look at a number of cognitive explanations in vision science, and argues that while content *attribution* is an important part of these theories, content *possession* is not. Content, she says, plays only a heuristic role in cognitive science, allowing theorists to show how the mathematically characterized mechanisms they propose amount to the intentionally characterized capacity they initially set out to explain. She accuses realists of holding out hope for an unlikely future: one in which a naturalistic theory of semantics is discovered.

I respond to Egan's negative attack on realism by arguing that content realism is not hostage to a naturalized semantics. Realism only requires that content attributions are based on objective, theory independent facts. I point to several examples where this is exactly what happens. Against Egan's positive proposal about the role of content, I argue that she has considered a far too limited set of cases. Look-

ing at examples of cognitive explanation from outside of vision science, I show that theorists use content for far more than matching a mathematical mechanism to a pre-theoretically characterized capacity. Not only do theorists attribute content to capture true generalizations, but they use it to successfully make *predictions* when no capacity is posited, something Egan's view cannot explain.

Wrap up

We have, I think, some very good reasons to take the language of thought hypothesis seriously. It provides a clear account of how thoughts are structured, and what properties they have. As I have said, I intend to show that it can overcome its most pressing contemporary problems. And note that my plan is to do so without adding any extra explanatory posits to the theory. Instead, every chapter shows a better way of working with the resources LOT already has. Preliminaries done, I will now try to make good on my promises.

CHAPTER 2

NEO-EMPIRICISM'S TROUBLES WITH CONTENT

Empiricism is resurgent in philosophy and psychology. This time around, it presents as a theory about the vehicles of thought—of what concepts are and how they are structured. I argue here that empiricist theories yield incoherent accounts of reference, and are not compositional—two key goals for any theory of concepts. Further, the revisions that empiricists would need to adopt in order to overcome these problems would render them virtually indistinguishable from amodal theories.

Like the semantic form of empiricism which Quine (1951) laid to rest, or empirical theories about the origin of our mental capacities which Chomsky helped to bury, the current reanimation of empiricism attempts to explain cognitive features in terms of perceptual elements. Concept empiricism, or neo-empiricism as it is sometimes called, claims that thoughts are composed of (copies of) perceptual representations. To think that the cat is on the mat, for example, is to bring online the copies of the representations that occur in our perceptual system when we see/hear/feel cats and mats.

Defenders of this view have offered both psychological evidence[1] and philosophical argument[2] in its support. The empirical evidence for concept empiricism has faced withering methodological critique.[3] For example, Machery (2007) argues that while neo-empiricists support their view by marshaling experiments demonstrating the use

[1] c.f. Barsalou (1999)

[2] c.f. Prinz (2002, 2005)

[3] c.f. Machery (2006 and 2007)

of perceptual imagery, amodal theorists (those who hold that concepts are elements of distinct, language-like systems of representation) have long held that we use mental imagery for certain cognitive tasks—those where using imagery is the best way to get results. Nonetheless, they claim, not all tasks that require concepts make use of mental imagery, so concepts need not *be* sensory structures. The empiricists' experiments, Machery claims, are those where using mental imagery is the best strategy for solving the cognitive problem at hand. So, they have failed to show a predictive difference between theirs and amodal theories (pp. 32–34).

While arguments like Machery's undermine the evidential support for neo-empiricism, my aim in this chapter is to offer reasons for thinking neo-empiricism is false. I will focus on Prinz's (2002, 2005) *proxytype theory*. It is a particularly clear, and philosophically informed, version of neo-empiricism, and one that is supposed to deal with just the sort of puzzles I raise here.

2.1 Areas of contention

This chapter has a secondary purpose. It aims to defend atomism against Prinz's (2002) direct attacks. In the previous chapter, I set out concepts' explanatory role, and the basic tenets of atomism. The dispute between empiricism and atomism, as it concerns us here, rests on the theories' abilities to satisfy four particular desiderata on theories of concepts. Prinz contends that, among other things, a theory of concepts ought to explain how our thoughts get their *intentional content*, what *cognitive content* is, and how concepts can be *public* or *shareable*. It must also be compatible with a theory of *compositionalality* (Prinz, 2002, pp. 3–16).

Prinz charges that atomism fails both to explain publicity, and to give a plausible account of cognitive content. As stated above, I argue that neo-empiricist theories fail to explain intentional content and composition. Some of these requirements I

discussed in the previous chapter. I will briefly review them here, and explain the additional work Prinz demands of concepts.

Intentional Content

Content is crucial to psychological explanation. For example, to understand Kahneman and Tversky's (1979) theory of loss aversion, it must be true that people have representations with the contents *losses* and *gains*. A theory of concepts must offer a plausible account of how our concepts manage to refer to properties, events, and other things.

Cognitive Content

There are two reasons for thinking that concepts cannot be individuated solely by their referents. One is the existence of Frege cases. Frege cases are those where substitutivity fails—where two co-referring terms cannot be substituted for one another *salva veritate*. For example, someone ignorant of the fact that 'J.K. Rowling' and 'Robert Galbraith' refer to the same person may want to buy everything the author J.K. Rowling has written, yet not buy any books where 'Robert Galbraith' is stated as the author, despite having the opportunity to do so. In such a case, it seems perfectly fine to say "this person wants to buy all of J.K. Rowling's books," yet somehow inapt to say "this person wants to buy all of Robert Galbraith's books," even though (a naïve) substitutivity principle says if one of these sentences is true, the other is too.

One explanation of why these cases occur is that people can have two different concepts of the same thing. The person in this example will have two concepts referring to *J.K. Rowling*: ⟨Rowling⟩, which they associate with 'J.K. Rowling', and ⟨Galbraith⟩, which they associate with 'Robert Galbraith'.

Prinz claims that what makes the ⟨Rowling⟩ and ⟨Galbraith⟩ concepts distinct is that they have some sort of non-referential content that distinguishes them. This

would be something like Frege's theory of *senses*, where senses were some sort of public object mediating the relation between a symbol and its referent, but also part of the symbol's content. Prinz holds that even if Frege's view is not quite right, we still need a semantic distinction between co-referring terms. Whatever this non-referential content is—be it sense, or something else—is what Prinz (2002) refers to as 'cognitive content' (p. 7–8).

Cognitive content is also evoked to explain the similar behavioral dispositions of people who have referentially distinct concepts. A very striking case of this was introduced in Putnam's (1975) famous Twin Earth thought experiment. Twin Earth is exactly like Earth, but instead of H_2O, a substance with the chemical compound XYZ fills its oceans, flows through faucets, etc. The inhabitants of Twin Earth, counterparts to all the people who live on Earth, call XYZ "water." They behave exactly as we do, but wherever we have H_2O, they have XYZ. As a result, my Twin Earth counterpart has a concept, $\langle \text{water}_{twin} \rangle$, which refers to XYZ, and which he associates with 'water'. This concept plays a role in his psychology that is exactly analogous to the role my $\langle \text{water} \rangle$ concept plays in mine. His *XYZ* beliefs, desires, etc. will cause him to act the same way my *water* beliefs cause me to act. Despite the fact that $\langle \text{water}_{twin} \rangle$ and $\langle \text{water} \rangle$ do not co-refer, it seems as if they have something in common. Prinz (2002) says that they have some *semantic* feature, cognitive content, in common.

So, Prinz claims, one job for a theory of concepts is to explain what distinguishes co-referring concepts, and unites referentially distinct concepts. That is, they must offer an account of cognitive content.

Publicity

The going wisdom claims that whatever concepts are, they must be shareable. It must be possible for distinct people to have concepts of exactly the same type. Fodor

and Lepore (1992) develop this point at length, arguing that linguistic communication is impossible if people do not map their words onto the same concepts. If what I mean by 'bird' is different from what you mean by 'bird', then I will be unable to communicate to you what I mean when I say 'that bird is a sparrow'. Fodor (1998) again endorses this, expanding the justification to include giving psychological generalizations. For a generalization like "thirsty people see water" to be true, he argues, it must be the case that the word 'water' maps on the same concepts for you and me; otherwise, the generalization means something different for both of us, and my saying it makes it true only of me (1998, p. 29). Prinz (2002) makes the case by a slightly different route, arguing that because people can act *for the same reasons*, and reasons (beliefs, desires, etc.) are composed of concepts, then people must be able to have the same concepts as one another. So, they agree, a theory of concepts must explain what concepts are such that we can share them. (I disagree, and argue against the publicity requirement (or *shareability* requirement) in chapter 3. But for now, let's grant this.)

Compositionality

We can think a potentially unbounded number of distinct thoughts. This is because our system of representation is *productive*—it allows an infinite number of semantically and syntactically distinct mental structures. Fodor and Lepore (2002) offer a charming example: Our thoughts allow us to think of a missile shield, an anti-missile shield, an anti-anti-missile shield, an anti-anti-anti-missile shield, and so on. (Though constraints on time and memory will stop this at some point, the rules of the system allow it.)

Our thoughts are also *systematic*. The ability to entertain certain thoughts entails the ability to think a variety of related thoughts. Someone who can believe that Jerry bit into the turkey can also believe that the turkey bit into Jerry.

To explain both of these features of thought, the dominant view is that thought is *compositional*. Our system of thought follows recursive rules that allow a finite store of concepts to be combined into an infinity of distinct complex structures. We can understand these thoughts because their meaning is determined entirely by the meanings of their constitutive concepts and the rules governing how they can be put together. Systematicity falls out of this because the systematically related thoughts employ the same primitive elements, and the same rules for composition. Productivity results from meanings being determined in part by recursive rules.

Prinz and Fodor both hold that a theory of concepts must be compatible with concepts' being compositional.

2.2 Prinz's "proxytype" theory

Neo-empiricism holds that concepts are percepts—sensory symbols. On Prinz's (2002, 2005) proxytype theory, this means that concepts are structures of copies of the representations used in an organism's dedicated input systems (2002, p. 115). He construes these systems to include externally directed senses (e.g., vision), internally directed senses (e.g., proprioception), emotions, and motor systems. So "sensory" should be understood rather broadly.

When I see (smell, hear, etc.) a turkey, on Prinz's view, I copy these perceptual representations—in the proprietary vocabulary of each sensory system—into long-term memory. Over time, these perceptual representations get linked together in a network based on patterns of observed co-instantiation. So, seeing a turkey and hearing a gobble links the auditory gobbling percept with the visual percept of turkeys. These linked representations form what Prinz calls a *long-term memory network*. Further instances of turkey percepts are stored in the turkey network on the basis of their similarity with the things already stored, or co-instantiation with things that are similar to stored representations (2002, pp. 144–149).

These networks, Prinz says, refer to particular properties or objects, and act as *detectors* for their referents: when enough of the features contained in the network are detected, the network is triggered. Over time, we adjust the weights of various elements in the system to better detect these objects or properties. If, for example, we find that the best indicator of whether something is a turkey is that it has a distinctive sort of wattle, then the visual representation of the wattle becomes more heavily weighted.

Thinking requires the ability to hold conceptual structures in working memory, which has a relatively limited storage capacity. These long-term memory networks would be too unwieldy to bring into working memory. I could not think, e.g., that *Jerry bit the turkey*, because working memory lacks the capacity to hold all of my perceptual information about turkeys, much less all of my perceptual representations of Jerry or biting. Prinz's solution is to identify concepts not with networks, but with *proxytypes*. Proxytypes are subsets of elements in long-term memory networks that can be called into working memory to stand for whatever the network detects (2002, p. 149). Tokening a proxytype, he says, "is generally tantamount to entering a perceptual state of the kind one would be in if one were to experience the thing it represents" (2002, p. 150). Which element serves as a proxytype depends on context. If I am thinking that I would like a turkey sandwich, I might pull up an image of a cooked turkey or the smell of a roasting turkey. If I think that turkeys have interesting feathers, I may deploy an image of a turkey fanning its tail. If there is no relevant context (e.g., someone just says "think of a turkey!"), Prinz says that we use a default proxytype, which is sort of like a prototype in that it represents some common features of whatever the network represents (2002, pp. 155–156). So, there are many different proxytypes, and thus many different concepts, for whatever a particular network represents.

Prinz claims that the proxytypes themselves constitute cognitive content—the stuff that's supposed to semantically distinguish co-referring concepts (2002, p. 270). My proxytype for J.K. Rowling might contain an image of 'J.K. Rowling' written on a book, while my Robert Galbraith proxytype does not. Or my proxytype of Superman contains an image of a guy flying around in spandex, while my proxytype for Clark Kent contains no such thing. Presumably, such proxytypes wouldn't even be parts of the same long-term memory network, since (being ignorant of these identities) I have no experience that would cause me to associate them with one another. On the other hand, my Twin Earth counterpart and I share cognitive contents because we have proxytypes containing the same types and weightings of *feature representations*. Feature representations are parts of percepts that represent particular distal properties of that percept, and count as the same type if they represent the same perceptually detectable distal properties (e.g. *being red*, *being an edge*).

I am not convinced that possessing a common set of feature representations can capture what Prinz means by having the same cognitive content. Quite a few different proxytypes, with what Prinz would presumably say have different cognitive contents, share feature representations. A proxytype for a square will contain exactly the same primitive features as one of a diamond, for example. Prinz might solve this by requiring that the feature representations occur in the same way in the proxytype. But this would seem to have the result that the only people who share proxytypes are Twin Earth counterparts, since all of us normal folks are likely to have slight differences in our perceptual experiences of things. Maybe that is OK, since it seems to only be Twin Earth examples that motivate the demand for cognitive content. I won't push too hard on this because (a) I don't care about the psychology of science fiction creatures, and (b) I don't think there is any such thing as cognitive content anyway (see §2.4).

Putting cognitive content to the side, a more pressing project is to explain how proxytypes get their intentional content. For this, Prinz (2002) borrows a page from causal-informational semantics:

> X is the *intentional content* of C if
> IC1 Xs nomologically co-vary with tokens of C and
> IC2 an X was the incipient cause of C. (2002, p. 251)

Nomological co-variation is standard fare for informational semantics.[4] The "incipient cause" condition is Prinz's way of handling the disjunction problem—that a concept would refer to the disjunction of everything it happens to covary with. For example, certain small bushes may reliably cause *turkey* representations on dark nights. The IC2 makes it so that $\langle \text{turkey}_{animal} \rangle$ represents *turkeys*, and only *turkeys* if a turkey encounter caused $\langle \text{turkey}_{animal} \rangle$ to come into existence.

This theory faces an immediate problem. If a concept has *turkey* as its intentional content, it must have all and only turkeys in its extension. But this isn't the case with Prinz's proxytypes. My visual representation of a turkey fanning its tail would have been created during an encounter with a living turkey. My olfactory representation of a roasting turkey (let's hope) was not. These proxytypes neither share an incipient cause, nor do they co-vary with the same things. Worse still, this makes a mess of our inferential abilities. Suppose I'm thirsty, and think to myself "I need to get a cup for some water." In doing so, I represent *cup* using the perceptual representations of a conical waxed paper cup. I look around and see styrofoam cups, thinking to myself "there are some cups," but this time my *cup* concept is an image of these cups. I cannot infer that that's the sort of thing I need, since that's not the object I initially pictured. My mental processors have no way of knowing that these are concepts of the same thing, since they look different in working memory.

---

[4] c.f. Dretske (1981)

Prinz (2005) is aware of these problems in his (2002) formulations. Here he offers a different account, which recruits the detection capabilities of long-term memory networks to do the work of conferring intentional content on proxytypes. On this version, to be a proxytype referring to X is to derive from a long-term memory network that nomologically co-varies with X and which had X as its incipient cause. So now both my turkey representations refer to *turkeys* because they come from the *turkey* network—the network whose activation co-varies with turkeys, and was initially created upon a turkey encounter.

This should suffice to give a reasonably clear picture of what Prinz's theory is. In what follows, I will argue that it is unworkable.

2.3 Problems with proxytypes

Prinz's theory faces two main problems: (1) its account of intentional content locks on to the wrong targets, does not actually explain how concepts get their content, and leads to problems for individuating concepts; and (2) it fails to be compositional. These problems are not without solutions. Unfortunately for Prinz, the solutions render his theory functionally indistinguishable from atomism.

Proxytypes and Reference

Prinz's (2005) theory of reference says that concepts, *qua* proxytypes, determinately have X as their intentional content because they derive from networks that act as X detectors, and had X as their incipient cause.

What do the networks detect?

The incipient cause constraint leads to serious problems nailing down the referent of concepts. This is because what causes a network to exist sometimes ends up being different from what the network eventually detects, and because a network's cause is often indeterminate.

Suppose that as a child I watched cartoons featuring cartoon cows. These are my first ever cow-like experiences, and my mind stores some of my visual and auditory representations of cows in long-term memory. Later in life, I see real cows. They look similar enough to the cartoons that I match them to the elements in the network containing these cartoon percepts, and add these percepts them as additional links. Over time I collect more percepts of real cows, and these gain much more weight than the original representations in the network. That is, I treat them as more reliable detectors for whatever the network detects. Whenever I see a cow, this network is activated. While it also gets activated when I see cartoon cows, I may reliably distinguish depictions from the real thing if, upon viewing cartoons, both this network and my cartoon detecting network fire off at the same time. (Well, maybe. It seems as if I would actually be misrepresenting something as both a cartoon and a cow. See my remarks on compositionality below.)

What do proxytypes deriving from this network represent? Intuitively, it represents cows. After all, the network is nomologically activated by cows, most of the percepts it contains were caused by cows, and the cow-caused percepts are the most highly weighted. But the proxytypes cannot represent cows. The network was created by an image of a *cartoon* cow, which is not a cow. It does not matter what caused subsequent percepts to come into being, since intentional content is determined by the network's incipient cause. At best, if Prinz is correct, these proxytypes represent *looking like a cow*; at worst I am constantly misrepresenting all cows as cartoon cows.

Counterpoint: The cartoon cow has a special causal relation to cows (by way of the artist's cow network). So perhaps that is, somehow, enough to fix the incipient cause as cows—by deference to an expert.

Response: You can run the same sort of cases with something that bears no causal relations to what the network seems to detect. Suppose I am at a farm and see an alpaca for the first time. I ask what that is, and a misinformed friend tells me it's

a baby llama. Into long-term memory the percept goes, and I spend the rest of the day looking up llama pictures and videos, and adding them to my long-term memory network. Fascinated, I go on to become a llama expert over the course of many years. Along the way, I acquire a network of exclusively alpaca percepts, and, being an expert, can easily distinguish them from llamas. I never realize that that hazily remembered percept from years and years ago was actually an alpaca. If Prinz's theory of reference is correct, I, the llama expert, do not even have a *llama* concept. Since the network's incipient cause is distinct from what it causally co-varies with, then it doesn't seem to mean anything at all, on Prinz's theory. Or perhaps it means *the sort of thing that looks like so*, which is still not a *llama* concept and would also seem to require that it detect alpacas too.

In a related worry, the incipient cause constraint does not allow the theory to pick out a determinate referent. Suppose my first encounter with Thin Mint cookies was seeing a sleeve of the cookies on a table. This image gets dumped into memory. I then eat the whole sleeve (minus the wrapper) and gain a slew of new percepts that get linked with my initial image of the cookie sleeve. Now I have a Thin Mint network. Do proxytypes deriving from it represent the cookies themselves, or sleeves of cookies? This network's activation reliably co-varies with both individual cookies and with cookie sleeves. Its incipient cause was likewise both. There is no determinate fact as to whether the proxytypes in this network represent *Thin Mint*, or *Thin Mint sleeves* (misrepresenting individual Thin Mints as sleeves), or *Thin Mints or Thin Mint sleeves*. Clearly the most plausible interpretation is that it represents the individual cookies, but that's a matter of our interpretation as theorists. Nothing in Prinz's semantic theory can distinguish between these possibilities.

Prinz might avail himself of non-semantic resources to try and solve this problem. The most plausible candidates are the different types of links he envisions as holding between items in long-term memory networks, using the following taxonomy: *Trans-*

formation links hold between representations that are permissible transformations of each other (e.g., an image of a sleeve of Thin Mints sitting still on the table, and one of a sleeve rolling across the table). *Hierarchical* links are formed between representations of a thing, and representations that "zoom in" on that thing (e.g., an image of a cookie sleeve, and an image of one small group of cookies in that sleeve). *Binding links* hold between items from different sensory modalities on the basis of their being co-instantiated (e.g., the sound of the cookie crunching and the taste of the cookie). *Situational links* hold between percepts that co-occur but are not physically bound to one another (e.g., someone eating a Thin Mint might cause us to bind other percepts of that person with other Thin Mint percepts). *Predicative links* join similar representations when we are disposed to transfer the features of one percept to another (e.g., though I did not bite into your Thin Mint, I am disposed to transfer the taste features of my previous Thin Mint percepts to yours) (Prinz, 2002, pp. 145-146).

The idea would be that my Thin Mint network could determinately refer to individual Thin Mints and not sleeves because they are not *simply* linked together, but linked together *via a hierarchical link*, so that the sleeve is indicated as a "zoomed out" version of the individual. So, we might think, the network represents the individual cookies because they are the common elements in each representation: a percept of *a cookie*, and a zoomed out image of *a cookie*.

But links are two-way streets. If my cookie sleeve image is hierarchically linked to a single cookie image, the former is a zoomed-out version of the latter, but the cookie is also a zoomed-in version of the sleeve. The same point holds for all of the other link types. For links to help in this scenario, there needs to be some prior way of privileging which item is the object being detected. Neither incipient cause nor co-variation can do the job. So, we've made no progress on the original problems.

Perhaps I am misinterpreting the incipient cause condition. I've been treating the cause of the first percept in a network as the network's incipient cause. But perhaps

Prinz intends to draw a distinction between the incipient cause of the percepts and the incipient cause of the network. The network is formed when the first link is made. Again, however, this doesn't solve the problems above. It does nothing to resolve the cartoon cow case, since the original network was formed when multiple cartoon images became linked. A more general issue is that we run into problems identifying the incipient cause of joined networks. If I have a small network consisting of the crinkly sound a cookie sleeve makes and the image of a cookie sleeve (joined because they were co-instantiated) and another consisting of a cookie taste and cookie-based emotion (also co-instantiated), and then the networks become linked (perhaps through a situational link or transformation link), is the cookie or the sleeve the new network's incipient cause? Prinz does not provide us with the resources to say. A deeper problem still is that the causal origins of networks are very different than the things Prinz takes them to detect. A llama did not cause the link between my *alpaca* representation and subsequent *llama* representations. While llamas caused those subsequent representations, the link formed because these mental images (etc.) were sufficiently similar to kick off the association process. Yet Prinz would not want to say proxytypes from that network thereby represent *that these two mental images are similar*.

All of these problems can be solved if we abandon the incipient cause constraint in favor of some other way of solving the disjunction problem. Fodor's (1992b) *asymmetric dependence* theory would be one option. Fodor's strategy is to say that for a representation ⟨X⟩ to have intentional content Y is for:

1. $\langle X \rangle$ to causally co-vary with Y, and
2. for all Z≠Y that cause $\langle X \rangle$
    - if Y did not cause $\langle X \rangle$ Z wouldn't either, and
    - if Z did not cause $\langle X \rangle$ Y still would.

So, for a representation ⟨cow⟩ to mean *cow* is for cows to cause its tokening, and all other causes of that representation to asymmetrically depend on its causal connection with cows. The fact that horses on dark nights cause ⟨cow⟩ tokens is only true because cows cause ⟨cow⟩ tokens. If they didn't, then dimly lit horses wouldn't either. If, however, horses at night did not cause ⟨cow⟩ tokens, the causal relationship between cows and ⟨cow⟩ would be unaffected (Fodor, 1992b, pp. 90–93).

If Prinz adopted this change, then in the initial cartoon cow case the mature network just means *cows*. If cartoons stopped causing it to activate, cows could still cause it to activate. If cows stopped activating it, cartoons would not either.[5] In the llama case, after I become an expert, the network's activity co-varies with llamas. Any alpacas that happen to cause it (seen from a distance, perhaps) do so because I mistake them for llamas—that is, if llamas did not cause it, the alpaca would not either, but not *vice versa*. Finally, in the Thin Mint case, we might suppose that though the network's activity co-varies with both sleeves and individual cookies, if the causal link between individual cookies and the representation were broken, the one between sleeves and the representation would be too, but not the other way around.

The above criticisms depend on unique features of Prinz's (2002, 2005) version of neo-empiricism. They would not apply to other versions of concept empiricism. This is because proxytype theory depends on neither the "contents" of the percepts themselves nor the similarity relations between empirical concepts and the world to

[5]Alternatively, neither might depend on the other. Cartoons would still cause it if cows didn't, and cows would still cause it if cartoons didn't. In this case, it is an ambiguous network. It could mean the disjunction of *cow or cartoon cow*, or perhaps *looks roughly like this*. Atomism does not face the same sort of problem because the causal relations hold between concepts and what they represent, not between concepts, detection networks, and referent. Because of this, atomic concepts can use many different resources to maintain their asymmetric dependence supporting causal connection to their referent. They may even use one or more Prinzean long-term memory networks to do so. So if one network causes an X concept to co-vary with some non-X things, the fact that the rest of its connections make sure that, should that link be severed, it would still co-vary with X. The reason Prinz cannot use the same sort of strategy because proxytypes have to issue from a single detection network.

determine conceptual content. This is a virtue of his theory, and one of the reasons I chose it as my target. Other neo-empiricist theories either do not offer a semantic theory (e.g., Damasio (1994)), or rely on the similarity, or alleged semantic properties of percepts to play a significant part in content determination. Those that adopt the latter strategy do so to their detriment. Barsalou (1999), for example, claims that percepts themselves are intentional, and that *their* content partially determines the reference of the concepts they constitute:

> There is good reason to believe that perceptual representations can and do have intentionality. Pictures, physical replicas, and movies often refer clearly to specific entities and events in the world. (1999, p. 597)

He then says that the other determiners are things external to the concept. He does not have in mind causal co-variation, but rather things like a concepts' occurring in a particular *definite description*. (This is the only example he gives, so I am not sure what other external factors he thinks determine content.) He considers the sentence, "the computer on my office desk is broken," and says:

> On hearing this sentence, imagine that a listener constructs a perceptual simulation of a computer. The content of this simulation is not sufficient to specify its reference, given that it could refer to many computers. However, when conjoined... with the simulation that represents the region "on my office desk," the complex simulation that results establishes reference successfully, assuming that only one computer resides in this region. (1999, p. 597)

Neither of the above claims withstand philosophical scrutiny. Barsalou cites Goodman (1976) as evidence for the claim that pictures (etc.) have intentionality. But Goodman did *not* claim that pictures (etc.) can have intentional content by themselves. Indeed, he derided the resemblance theory of pictorial representation that this would suggest as the "most naïve view of representation" (1976, 3). Instead, he argued that pictures (etc.) only have content *relative to a conceptual schemata for interpretation* (1976, 7-9). This more informed view cannot account for conceptual content, since it requires conceptual content to do the content determining work. Or, put more simply: images

have content because we interpret them to, but concepts do not. And the definite description account of reference is a non-starter. As Kripke (1980) argues, every definite description we have about someone may be false, yet our concept for them would still refer to that person. (Kripke's argument is about words and names, but the point transfers.) I can think that the computer on my desk is broken, referring to a particular computer, even if, unbeknown to me, someone moved it from the desk.

Information semantics is a much more promising route towards content determination. While the specifics of Prinz's (2005) and his (2002) do not hold up, the revision I suggested above does make it more plausible. But even this is not enough to solve proxytype theory's reference problems.

How do proxytypes get their content?

Let's suppose that causal co-variation accounts of content work. Representations refer to the things they causally co-vary with (plus some account to overcome the disjunction problem). On Prinz's (2005) view, the things that causally co-vary with their putative referents are *networks*, not *proxytypes*. Proxytypes cannot, in fact, co-vary with these things. My roasted turkey imagery will not co-vary with live turkey instances (suppose), and *vice versa*.

Instead, Prinz holds that proxytypes get their putative content by *deriving* from a network that bears the right co-variation relation. But being part of something that co-varies with X is an altogether different relation from co-varying with X. Causal co-variation is a form of informational semantics. Informational semantics is plausible because it is based on a lawful link between a representation and the thing it represents. Prinz has gotten rid of this lawful relation at the level of concepts.

Lawful co-variation is a requirement on carrying information about something. Without it, Prinz does not have an informational semantics for proxytypes. This is

not itself a problem, but we need some reason to think that this new "derived from a reliable detector network" theory is content determining.

Prinz does not provide one. He introduces the idea in his (2005), exchanging the term 'file' for what he previously called 'long-term memory networks' (there seems to be no substantive difference). Here is what he there in defense of network (or file) derivation being a content determining relation:

> There is a reliable causal relationship between encounters with members of a category and representations derived from the file for that category. My Pomeranian representation is not reliably caused by dogs, in general. It might not be activated when I encounter a sheep dog. But my Pomeranian representation is a member of a mental file containing variable dog representations, *and these collectively are reliably caused by dogs*. (2005, p. 6, emphasis added)

Let me pause here to note that it is not true that the members of a long-term memory network are collectively reliably caused by their referents. Or, at least, Prinz is equivocating when saying so. It may be true that for every perceived dog, there exists some part of the network that is activated. And if we consider the network to be activated when any of its parts are, or a sufficient number or weight of its parts are, then we might say that the network reliably co-varies with dogs. But this is placing a technical definition on what counts as a network being activated, and it differs significantly from our ordinary understanding of collective relations. Though any random 10 one-gram weights grabbed from a box of 50 will cause an accurate scale to indicate 10-grams, the entire sack of weights does not *collectively* cause the scale to indicate 10-grams, no matter how many times I repeat the action. Nor would a group of 10 people collectively cause a cake to disappear if only three people ate any. And a group that reliably votes 60/40 for a particular candidate in multiple elections may collectively *elect* the candidate (since "to elect X" is to participate in an election where X wins), but they do not collectively *vote for* the candidate. So, I think Prinz is asserting a less tenuous causal relation between members of a long-term memory network and the category the network detects than he is entitled to. At any

rate, even if I am wrong about the semantics, all this gets us is that the networks have the content *dog*, not the proxytypes. Prinz goes on to say how he thinks content transfers to the proxytypes:

> In other words, dog encounters reliably cause us to access the dog file. Items in the dog file are, in that sense, under the nomological control of dogs. Items in a mental file can be said to refer to the category that the file reliably detects. Thus, the nomological condition on reference can be met, even if the representations we use to categorize dogs are highly variable. (2005, p. 6)

This is no more than hand-waving. In what sense, exactly, are particular items in the dog network under the nomological control of dogs? There is no law-like relation that holds between any given percept and dog instances. And why *should* it be said that these particular items refer to whatever the category detects? There is likewise no nomological relation between the network's activation and a given proxytype's instantiation. Further, while Prinz is trying to make the case that *parts of the network* bear some sort of causal connection to the network's category, he is not telling us how *proxytypes* get this content. Proxytypes are the *copies* of parts of the network tokened in working memory. These bear even less of a nomological relation to their alleged referents since they are not technically part of the network that gets activated, but are rather derived from parts of this network.

The burden of proof in showing that this "derived from" relation is content determining is on Prinz. There are plenty of cases where being part of something that represents X, or derived from part of something that represents X, does not suffice for representing X. If one had letter blocks spelling 'cow', pulling out the 'o' block would not thereby make that single block represent cows. Nor would a photocopy of half of the word 'empiricism' represent empiricism. And the concept ⟨cow⟩ does not mean *brown cow* even though it is a constitutive part of the more complex concept ⟨brown cow⟩.

These problems could be avoided if at least one element of a network is always activated when the relevant category is detected, and that element is present in all proxytypes for that network. This way, at least part of every proxytype would nomologically co-vary with category instances. Let's call this the 'key percept'.

How are proxytypes individuated?

Percepts can belong to many different networks, according to the proxytype model. An image of Daisy the cow may occur in both my cow detector network and to my Daisy detector network. It is also possible, following my emendation above, for the same percept to be the key percept for multiple networks. Daisy may be the very model of a cow for me, with the same image being the key percept in both cow and Daisy networks. Or, more likely, one network's key concept may be a non-key element of another. Since proxytypes can be constructed from any percepts in a network, it is possible for working memory to contain indiscernible proxytypes standing for the same thing. This might happen if, for example, I think "Daisy is a cow." The most relevant percepts in my cow network in this context will be of Daisy, as will the Daisy percepts.

This is a problem. In order for my mind to distinguish between *cow* thoughts and *Daisy* thoughts, there must be something it can use to tell which concept is which. And I need it to distinguish between them so that it can reliably make truth-preserving inferences. If I believe that Daisy only eats oatmeal, and that Bessie is a cow, but my *Daisy* and *cow* concepts are indiscernible, I might falsely conclude that Bessie eats oatmeal:

1. ⟨MOOO only eats oatmeal.⟩
2. ⟨Bessie is a MOOO⟩.
3. ⟨Therefore, Bessie eats oatmeal.⟩

So, not only do these concepts need to be referentially distinct, they need to somehow be distinctly marked.

Prinz's response to this sort of worry is to appeal to the weighting and "detection tendencies" of different proxytypes:

> Two mental representations composed of the same primitives can qualify as distinct proxytypes if those primitives are weighted differently. Two such proxytypes would have different detection tendencies and perhaps different intentional contents. (2002, p. 275)

This does not overcome the conflation problem. For one, proxytypes do not themselves have detection tendencies. Insofar as they might bear some sort of causal relation to their referents, the indiscernible *Daisy* and *cow* proxytypes will bear it to the same thing: Daisy. There is also no reason to think that the proxytypes' weightings need to be different. Their respective *networks'* weightings may be different, but the elements in the proxytype may nonetheless be weighted the same. So, it looks like we need an additional resource to explain how the mind distinguishes symbols from distinct networks.

On the other side of this coin, my mind needs to be able to tell when two symbols come from the *same* network. Suppose ⟨<hoofs, moos, black spots>⟩ is one of my proxytypes for *cow*, and ⟨<moos, brown, wears a bell>⟩ is another (⟨<moos>⟩ being the key percept). If my desire to feed a cow is represented like this:

⟨I want to feed a <hoofs, moos, black spots>⟩

And my belief that Daisy is a cow is represented like this:

⟨Daisy is a <moos, brown, wears a bell>⟩

There is nothing to allow me to infer that Daisy is the sort of thing that I want to feed. There is no way for my mind to recognize that I am deploying two concepts of the same semantic types.

What Prinz needs, then, is something to mark all proxytypes from the same network that is unique to that network. It can't be a percept, because percepts will

not be unique. We would still end up in a Daisy/cow scenario. If the marker is not perceptual, then it is amodal. And since it will occur with all proxytypes for a category, it can do the work the key percept was supposed to do. In turn, this amodal marker will causally co-vary with the category the network detects.

Since the accompanying percepts are not doing any work in terms of standing for that content, or allowing the proxytypes to be recognized and distinguished, they are not necessary for the concept to do its work. So, we could just say that the amodal symbol *is* the concept. It can be accompanied by various perceptual representations, if that is useful for the cognitive task at hand.

Here is the picture we are left with: concepts are amodal symbols that causally co-vary with what that concept represents. The networks still function as detection mechanisms, but their purpose now is to maintain the causal connection between the amodal symbol and its referent.

This modified view overcomes all of the problems I have raised for Prinz. It is not an empiricist theory, but it is compatible with some of the neo-empiricists' ideas (e.g., that we sometimes use mental simulations when thinking about things). It is also, you may have noticed, exactly what atomism claims. To hold on to informational semantics, it seems Prinz is required to give up neo-empiricism and become an atomist.

Compositionality

Prinz (2002) offers a detailed, three-stage model of proxytype combination. In the first stage, the *retrieval* stage, one looks for a single concept or exemplar that corresponds to the complex concept one is trying to construct. So, a *pet fish* concept, Prinz claims, is not constructed by combining *fish* and *pet* concepts. Rather, it either has its own concept, or it is constructed by finding an exemplar that exists in both the *fish* and *pet* networks (2002, pp. 301–302).

If the retrieval stage is unsuccessful—no common exemplar or previously created concept can be found—we move on to the *composition* stage. In this stage, the default proxytypes for a pair of concepts are combined together into a single proxytype using combination rules. One is "aligntegration," Prinz's term for a process in which features of one concept are substituted for the features of the other, and weights are adjusted. So, ⟨striped apple⟩ may be constructed by adding ⟨striped⟩ to the "surface pattern attribute" of ⟨apple⟩ (2002, p. 303). The other combination rule is what Prinz calls "feature pooling." Feature pooling creates a new proxytype, and combines the highest weighted features of the combining proxytypes in it. So, he says, ⟨houseboat⟩ may be constructed by mushing together the most salient features of a *house* concept with those of a *boat* concept in a new proxytype (2002, pp. 304–306).

Finally, background knowledge is brought to bear on the composed concept to make sure that it is coherent. This is the *analysis* stage of composition. This process fills in gaps in the composed concept by drawing inferences from it, and introducing features that were not part of the combined concepts. Prinz gives the examples of ⟨nonmaterialistic⟩ being added to the combined representation ⟨Harvard carpenter⟩, and of the conflict between the features ⟨malodorous⟩ and ⟨pleasant to cuddle⟩ of ⟨pet skunk⟩ being detected and resolved by background knowledge about surgically removing scent glands (Prinz (2002, pp. 306–307)).

Together, retrieval, composition, and analysis form Prinz's "RCA" model for composition. Despite its detail, it faces some very elementary problems.

To what do combined concepts refer?

The RCA strategy comes from Prinz (2002), before he made the shift to causal networks bearing the co-variation relation. In the original reference paradigm, recall, the proxytypes themselves were supposed to bear the right sort of causal relation to their referents. The problem was that the incipient causes of and causal relata to

these proxytypes were never entire categories, but proper subsets of the categories' extensions. So, if, in the combination stage, I were to create a concept for *bespectacled turkey*, the resulting combined concept might only include features from a male turkey and wire-rimmed glasses, resulting in a proxytype that refers to *a male turkey with wire-rimmed glasses*. Close, but not close enough.

Prinz's (2005) revised strategy actually fares worse on this front. Proxytypes are supposed to refer to whatever their originating networks detect. Strictly speaking, the combined concept does not derive from a network, at least not in the same way that simple proxytypes do. We could say the new proxytype originates from two networks, but then its referent would seem to be the union of turkeys and spectacles, and not a bespectacled turkey.

I suspect that Prinz is implicitly relying on a similarity model of reference here, instead of an informational model. The reason he thinks that these combined concepts refer to particular complex categories is not because they are causally related to them, but because they have the features that prototypical members of those categories might have. This leads to all of the classic problems for similarity views, and prototypes as semantic determiners.[6] In brief, many things are similar to the stored bespectacled turkey representation that aren't bespectacled turkeys (e.g., a plush turkey doll with glasses, a monocled turkey, or a particularly large chicken wearing Ray-Bans). And many bespectacled turkeys are not particularly similar (e.g., a dead and plucked turkey wearing glasses, or a turkey wearing glasses with a very convincing pilgrim costume).

Reference problems also arise at the retrieval stage. If the same exemplar exists in my *pet* network and my *fish* network (one depicting a goldfish), and that is supposed to serve as a pet fish concept, it is difficult to see why it would mean *pet fish* instead of

[6]See Fodor and Pylyshyn (2014, ch. 2) for a recent discussion.

pet, or *fish*, or *goldfish*, or *pet or fish or goldfish or pet fish*, etc. Again, Prinz's (2005) model only makes things worse. Here we have distinct yet indiscernible exemplars in different networks. It is no longer a simple matter to say which is *the* exemplar used. It matters which one, because the new proxytype's semantics are supposed to be determined by the originating network. If it is a copy of one particular exemplar, it *can't* mean *pet fish*; it can only mean what the originating network detects. But if it is a copy of neither, then it doesn't mean anything. If it is a copy of both, yielding a proxytype with two goldfish exemplars, and both exemplars retained their progenitors' referents, then perhaps it could refer to the intersection of *fish* and *pets*. But without something like the amodal network-marker I proposed above, there is nothing that pushes for the *pet fish* content instead of *two goldfish*, etc. And with the amodal marker... well, who needs the exemplars?

What happened to systematicity?

Part of what compositionality is supposed to explain is the systematicity of thought. Prinz's RCA model does not. This is because while it provides a theory of composition, it has nothing to say about decomposition—breaking a complex representation into its constituent parts. But decomposition is equally important to systematicity. Anyone who can think of a large fish and a pet bear is able to think of a pet fish and a large bear. Yet someone who has a proxytype for *pet bear* and *large fish* cannot automatically do this. The proxytypes do not contain distinct elements for the terms in the compound. The largeness of the fish is property of the fish, not an independently meaningful element. Even if the percepts standing for the fish's size properties could be transferred to parts of the pet bear proxytype, it would yield a proxytype of something that is large for a fish, yet quite a bit smaller than an average bear. Moreover, a ⟨pet bear⟩ proxytype will not have any of the features that our prototypical pets

have. So it seems unlikely that that representation could be spun off into a reliable pet detection network.

So, the RCA model cannot provide a truly compositional semantics, and it violates the systematicity desiderata that Prinz himself set out. For these reasons, and the fact that it lacks a feasible theory of reference, I believe that proxytype theory is not an adequate theory of concepts. Prinz believes the same of atomism. I want to turn now to his criticisms of this view.

2.4 Prinz's criticisms of atomism

Prinz (2002) offers two general lines of attack against atomism. One is that cognitive content cannot be a purely formal property because the difference between concepts must be epistemically accessible. The other is that the notion of form is borderline incoherent, and no consistent reading can account both for behavioral similarity across Twin Earth cases *and* yield shareable concepts.

Must senses be epistemically accessible?

Atomism says the mind distinguishes concepts purely by their formal (that is, non-semantic) properties. Prinz argues that cognitive contents must be distinguished by epistemically accessible features, and since formal features are not epistemically accessible, they cannot be used to distinguish cognitive contents. He motivates this claim by describing how someone might distinguish their own co-referring but cognitively distinct concepts: Suppose, he says, that Sally has two concepts of Farrakhan. How might she determine that she has two different concepts? By appealing to features like "this concept is about an *orator*, but the other is about a *musician*" (Prinz, 2002, pp. 96–97). Prinz claims that she can only make this distinction because the concepts contain epistemically accessible features.

In what he puts forward as a related line of reasoning, Prinz asks how we might discover that two people share a concept. He gives an example in which Ollie has a concept of Austria that he expresses as 'Austria', and Otto has a concept that he expresses as 'Österreich'. Ollie and Otto come to find that they have the same concept by comparing their beliefs, and finding a common core. In both this and the Farrakhan case, Prinz says, we point to certain beliefs that the people have. In neither case would we point to formal features (Prinz, 2002, pp. 96–97).

It is unclear what Prinz thinks a theory of concepts is supposed to be explaining in these cases. The first case suggests that he thinks figuring out whether two concepts co-refer is discoverable *a priori*. But this is a peculiar demand. It was not in his initial list of desiderata, and we have every reason to think this is not true (the existence of Frege cases would seem to put this idea to rest).

Perhaps he is thinking that it must merely be possible for people to have distinct, co-referring concepts and know it. Fair enough, but nothing atomism says is incompatible with this. Presumably one would not discover this *a priori*, so the concepts themselves would not need to have epistemically accessible features. Instead, one might have epistemically accessible beliefs which turn out to be false. For example, $\langle \text{Farrakhan}_{orator} \text{ is not Farrakhan}_{musician} \rangle$. So it's not the concept's features that need to be epistemically accessible, but rather beliefs. And, a theory of concepts does not need to explain how we know our own beliefs (and thank goodness for that).

In the Ollie/Otto case, the participants are not actually verifying that they share the same concepts. They're verifying that their words refer to the same thing. As Prinz well knows, distinct concepts can refer to the same thing. So trying to verify that two people have the same concepts would require quite a bit more work than just pointing to a map and saying "this is the place I mean by 'Austria'." (In fact, as I will argue in the next chapter, verifying that two people have the same concepts is rather a hopeless endeavor, and not one worth engaging in.) Since all they need to do

is figure out that they're talking about the same thing, they only need to examine and express some of their beliefs (e.g., that Austria is *here* on the map). The structure of concepts simply just doesn't factor in.

Is formal individuation incoherent?

Atomism handles Frege cases by claiming that co-referring concepts can be formally distinct. The reason I can rationally believe that Superman flies, and disbelieve that Clark Kent does, is because the *Superman* concept in the first belief has different non-semantic features (think, roughly, orthographic features) than the concept referring to the same person in the second belief.

Prinz (2002) correctly notes it is not at all clear what *formal features* are. He suggests three different possible accounts (p. 97). Token concepts could have the same form if they share either:

1. physical properties,

2. intentional content, or

3. conceptual role.

Prinz quickly dismisses the first option by saying that it would require the type-identity theory to be true in order for two concept tokens to share cognitive content. The second he dismisses because it would fail to explain Frege cases. The third he dismisses because it runs afoul of the publicity desideratum on a theory of concepts.

Atomists would agree that intentional content cannot individuate concepts. The very definition of formal features, as Fodor (1998) uses the term, is to be non-semantic.

Prinz's blithe dismissal of the first possibility is, however, unwarranted. Recall that cognitive content is supposed to explain two things. One is how Twin Earth counterparts can have analogous behavioral dispositions despite lacking co-referring terms. The other is making sense of Frege cases. Prinz errs in thinking that atomists

recruit formal features to do both, that form is identified with cognitive content. What atomists actually say is that there is no such thing as cognitive content. You don't need semantic similarities to explain either thing that cognitive content is presumed to explain.

If formal features are physical properties, we can explain Frege cases by saying that people have two physically discernible concepts that refer to the same thing. These differences need not be epistemically available for the reasons given above. That's good, because you can't do neurology through introspection.

Prinz levels another objection against identifying form with physical properties. In its entirety, it is that doing so "renders it virtually impossible for the same concept to be tokened twice" (2002, p.97). He gives no argument for this claim, other than saying that its falsity would require the type-identity theory to be true. I agree that it may be virtually impossible, outside of Twin Earth cases, for *two distinct people* to have concepts with the same form. But why should they? All form needs to do is provide a way for a single mind to keep track of its symbols, and allow distinct concepts to refer to the same thing. And within a single mind, the same set of physical properties can be realized many times, in many ways.[7]

Behavioral similarities can be explained by causal role. My counterpart's *XYZ* concept plays exactly the same causal role in his psychology as my water concept does in mine. Because of this, our dispositions to behave will be the same. Prinz's complaint seems to be that for us to explain similar behavior in cases where people lack co-referential terms (e.g. Twin Earth cases), it needs to be possible to share conceptual roles. And this is virtually impossible, since two people will almost never share concepts with exactly the same conceptual role. But Twin Earth counterparts share exactly the same conceptual roles by design, so that's not a problem for the

[7]For a full defense of this see chapter 3 for an argument that form need not be shared, and chapter 4 for an argument that physical properties can—indeed must—individuate lexical symbol types.

atomist. And regular, non-science fiction people may not have the same exact roles, but they can have sufficiently similar roles, affording sufficiently similar behavior.[8] I admit to being puzzled by this criticism. Unless there is some reason to think that we have the *exact same behavioral dispositions* and do not in those cases have the same causal role, I can't see why this doesn't do just what Prinz wanted of cognitive content.

Note also that neither of these complicate the theory. Every theory must allow that concepts have physical properties and causal roles. Atomism simply recruits these extant features to do more work.

2.5 Last gasps

What this chapter has shown is that proxytype theory, seemingly the most philosophically sophisticated version of neo-empiricism, fails on several fronts. Its theory of reference delivers the wrong results in many different ways, and it makes a mess of compositionality. I suggested fixes for each of these problems, but the end result was simply an amodal theory—atomism with idle percepts tacked on. Additionally, Prinz's criticisms of atomism rested on misunderstandings about the role of conceptual structure, and atomism's take on cognitive content (viz. there is none).

There is one final objection Prinz raises that bears further thought. If concepts are individuated by formal features, and formal features are something like physical properties (or are otherwise unique to each individual thinker), then it seems like no one shares concepts (2002, p. 97). That is, no two people have tokens of the same type of concepts. So, atomism fails to meet the publicity requirement after all.

---

[8]N.B. this has nothing to do with *content holism*. The similarity governs behavioral similarity, and nothing semantic.

I think this criticism is exactly right. Atomism is incompatible with the publicity requirement. But I draw a different moral than Prinz. In the next chapter, I will argue that the publicity requirement should be abandoned.

CHAPTER 3

THERE ARE SOME THINGS YOU JUST CAN'T SHARE

In this chapter I want to challenge a doctrine common to both philosophy and psychology: that concepts are the sorts of things people can and do share.[1] One reason offered in its support is that communication is possible only if concepts are shared. Gelman and Kalish (2006) write, for example, that "shared concepts are a prerequisite for communication," and claim that "[i]t is only because two people share the concept *dog* that they can talk about dogs" (p. 719).[2] The other reason given in support of this doctrine is that shared concepts are required to capture intentional laws and generalizations. Fodor (1998) takes this line, arguing that people must be able to share a *water* concept if a generalization like "thirsty people seek water" is to apply to more than one person (p. 29).

Despite this apparent support, the doctrine leads to a puzzle. Consider a case where an individual has multiple concepts referring to the same thing. Someone ignorant of the chemical structure of water may, for example, believe that water is safe to drink but H_2O is not, indicating that they have two distinct concepts referring to water. Call these $\langle W \rangle$ and $\langle H \rangle$. Would such a person share concepts with someone who only has one *water* concept (call it $\langle Z \rangle$)? People like this could, at least in some

---

[1] c.f. Carey (2009); Fodor (1998); Gauker (2011); Gelman and Kalish (2006); Laurence and Margolis (1999) (in terms of "stability"); Prinz (2002); Rey (1983) (in terms of "interpersonal stability functions").

[2] Gelman and Kalish (2006) explicitly note that there is a tension here, though not the one that will concern us in this chapter. In the same sentence they say that concepts will differ in some ways between different people and person stages. The lesson they take from this is that sharing must allow for some conceptual differences. My arguments here contend that any talk of *the* concept of something is mistaken, so there is no sense in which people share that concept.

circumstances, communicate about water, and at least some *water* generalizations would subsume them both. So it seems they must share concepts. But which concept, ⟨W⟩ or ⟨H⟩ do they share? It cannot be both; if ⟨W⟩ is the same type of concept as ⟨Z⟩, then ⟨H⟩ must be a different type of concept from ⟨Z⟩ since ⟨W⟩ is a different type of concept from ⟨H⟩. But neither has more claim than the other on being the *water* concept, since we can imagine each playing the role of facilitating communication about water in different circumstances. So, it seems, some concept must be shared, but neither can be.

I argue that puzzles like the above reveal fatal flaws in the motivations for the shareability thesis. The argument from intentional generalizations rests on an equivocation between shared conceptual content and shared conceptual vehicles. The argument from communication leads to logical contradiction. The shareability thesis is thus unmotivated, and should not constrain theories about the nature of concepts. I propose a *counterpart model*, analogous to Lewis's (1973) counterpart theory for modality, that exchanges cross-personal concept identity for a context-dependent similarity relation between distinct concepts. This model not only avoids the problems I raise for shared concepts, but better explains concepts' role in communication without positing additional theoretical apparatus. Additionally, I argue that the counterpart model may pay dividends in the philosophy of language by pointing the way toward a resolution of several long-standing puzzles about belief reports.

3.1 The shareability thesis

I wish to dispute the doctrine that concepts must be shareable. What defenders of the doctrine mean by "share" is different than the sense in which, say, two cars share a garage, or a library shares its books. In these cases multiple individuals in some sense possess numerically identical objects. What defenders of the doctrine have in mind is instead more precisely characterized as what I will call the *shareability thesis*:

> **Shareability thesis:** Distinct people can, and commonly do, have token concepts of exactly the same type.

The shareability thesis, if true, substantively constrains conceptual theories. Concepts, *qua* explanatory posits in cognitive science, are the constituents of propositional attitude states (beliefs, hopes, fears, etc.). This means that they must be mental particulars. Since people do not have numerically identical mental particulars in common, sharing concepts must be understood as a type/token relationship: to share a concept with someone is to have a token concept of exactly the same type as one of theirs.

If we accept the claim that concepts are shareable, we can use this as a check on theories of concepts. These theories aim, among other things, to explain how concepts are individuated. If the individuation conditions they give make it impossible or improbable that two people have token concepts of the same type, then, because concepts are shareable, that theory should be rejected.

Fodor and Lepore (1992) dismiss theories committed to semantic holism on these grounds. They argue that if concepts are at least partly individuated by their content, and conceptual content is determined by the totality of one's beliefs (as holistic theories claim), then virtually nobody shares concepts, since practically everybody has some belief that is unique to them. Ironically, Prinz (2002) turns this line of reasoning back on Fodor. Fodor (1998, 1995, 2007) holds that distinct concepts can have the same referential content if they are different "formal" types. Prinz (2002) argues that to do this work, formal types either have to be physical types—in which case it is improbable that anyone would share them—or functional roles—in which case it seems as if Fodor is led to holism, and thus to the very problem he identified (pp.95-7).

Though Fodor (1998, 2004) and Prinz (2002) endorse very different theories about the nature of concepts, their reasons for endorsing the shareability thesis are the same. They argue that it is needed to capture explanatory generalizations in psychology, and

to enable linguistic communication.[3] The purpose of this chapter is not to argue that the shareability thesis is false. Rather, I will argue that it is unmotivated. Neither of the two arguments given in support—the argument from intentional generalizations, and the argument from communication—succeed.

3.2 The argument from intentional generalizations

Many explanatory claims in psychology are *intentional*. They provide reasons for someone's behavior, asserting a causal relation between that behavior and some mental states specified by their contents. Consider the following example:

(1) Yvonne went to the library because she wanted a book, and believed the library was the best place to get it.

This simple case explains a behavior (going to the library) in terms of some mental states (a desire and a belief) individuated by their content (that *Yvonne get a book* and that *the library was the best place to get that book*). Prinz (2002) uses an example like this to motivate the shareability thesis. Notice, he says, that these sorts of explanations can apply to multiple people:

(2) Xavier and Yvonne went to the library because they wanted a book, and believed the library was the best place to get it.

This ascribes the same set of attitudes to two people. If they have the same attitudes and attitudes are composed of concepts, Prinz reasons, then they must have the same concepts. So, he argues, the fact that we can subsume multiple people under the same intentional explanation shows that concepts are shareable (2002, pp.14-15).

---

[3]Because they make these arguments explicitly, I will be focusing on Fodor and Prinz, but these same lines of reasoning are implicit in other defenders of the thesis. Carey's (2009) concerns about "disagreement" are a special case of communication, for example (p.497). And Rey's (1993) defense of analytic intuitions presents a special case of psychological generalizations (e.g., everyone with the concept ⟨bachelor⟩ has the concept ⟨unmarried⟩).

Fodor (1998) advances a similar argument. Intentional explanations, he claims, require covering generalizations. According to Fodor, if 'I was thirsty' explains why I got water, then there is a law-like generalization that says that thirsty people tend to seek water. This generalization subsumes people under the same type of attitude: a desire for water. (Or, more precisely, it subsumes people under attitudes with the same content. As we shall see, this is the distinction that makes a difference.) If people share this attitude, then, Fodor reasons, people must share the concept ⟨water⟩ because:

> If everybody else's concept WATER is different from mine, then it is literally true that only I have ever wanted a drink of water, and that the intentional generalization 'Thirsty people seek water' applies only to me. (1998, p.29)

Fodor's reasoning is that without shared concepts, "generalizations" are not general, and without generalizations we do not get intentional explanations. So, he concludes, concepts must be shareable (or "public," as he puts it).

The particular example Fodor uses is not, perhaps, the best case for the shareability thesis. Plausibly, the generalization relevant to explaining my water-seeking behavior is this: If someone wants X, and believes they can get X by doing Y, then, *ceteris paribus* they will get Y. Or, in other words, people tend to try and get what they want. This schematic generalization is not committed to any particular attitudinal content. If so, it would give us the generalization that Fodor says we need without committing us to shared attitudes and concepts.

What Fodor needs is a content dependent generalization—one where the particular content it ranges over plays an important role in the explanation. Here is one such example from the psychological literature on categorization: Diesendruck and Peretz (2013) conducted a series of experiments to discover differences between children's categorization of artifacts and their categorization of natural kinds. In one experiment, they told children a story either about a scientist who liked to create new

kinds of artifacts, or one who liked to create new types of animals. They would then present the children with an object that the scientist created, and ask them to put it with the others of its kind. What they found was that while children consider the creator's intent relevant to determining category membership for artifacts, they do not consider this information relevant when categorizing animals. This information allows us to make a predictive generalization about children's behavior:

(3) If a child believes that X is an artifact, and that X's creator intended X to be an F, then, *ceteris paribus*, they will categorize X as an F.

This is a content dependent generalization. It depends crucially on how the child thinks of the target object. If the child does not represent the target object *as an artifact*, it will not hold.[4]

If this is a genuine explanatory generalization, then it looks like we need to ascribe attitudes with the same content to lots of distinct people. And this, Fodor believes, is enough to justify the shareability thesis.

To recap, both Fodor and Prinz reason as follows: There are true intentional explanations. The same intentional generalization can apply to distinct people. (For the reasons given above, Fodor believes they *must* apply to distinct people.) This requires that distinct people have attitudes with the same content. And this, they conclude, means that distinct people must share concepts.

The crucial inference is the last one: that sharing attitudes with the same content entails sharing concepts. Prinz states this explicitly, writing that "actions can be motivated by the same attitudes only if those attitudes are composed of the same concepts" (2002, 15). Fodor leaves the inference implicit. (He spends the remainder of this section arguing that people had better have the same content, not just similar

[4]There are many other examples like this. For an overview of similar work in developmental psychology, see Carey (2009) ch. 3-6.

content (1998, pp 29-34).) This suggests that once you get attitudes with the same content, you are supposed to get shared concepts for free.

If there is an argument to be found for this claim, I believe it is this: To say that concepts are the constituents of attitudes is to say that concepts are the meaningful parts of the attitudes. Both Fodor and Prinz endorse a compositional semantics for propositional attitudes. That is, they believe that the content of an attitude is wholly determined by the content of its constituent parts and the way those parts are put together. So, if two token attitudes have the same content as one another, we might reason, then they must have the same meaningful parts put together in the same way. Since concepts are the meaningful parts of attitudes, then, it appears token attitudes with the same content will have the same (types of) concepts as one another.

We can reconstruct the full argument as follows:

P1 There are true intentional generalizations.

P2 Intentional explanations subsume multiple people under attitudes with the same content.

P3 If (1) & (2), then distinct people often have token attitudes with the same content.

P4 If two token propositional attitudes have the same content, then they have the same meaningful parts.

P5 If people often have attitudes with the same meaningful parts, then people often have attitudes with token concepts of the same types.

P6 If people often have attitudes with token concepts of the same types, then the shareability thesis is true.

C Therefore, the shareability thesis is true.

I grant premises (P1)-(P3). (P4) and (P5), I will argue, rest on an equivocation. If I am right, then the argument from intentional generalizations fails.

3.3 Decomposition and Frege cases

The argument from intentional generalizations depends on premise (P4) above:

> If two token propositional attitudes have the same content, then they have the same meaningful parts.

At a glance, it may seem that this follows from the fact that thoughts have a compositional semantics. It does not. Hard currency has something analogous to a compositional semantics: the amount of money some token collection of hard currency represents is wholly determined by the amount of money its constituent parts represent. (N.B., constituent parts, not mere parts: half a dollar bill does not represent half a dollar.) Our having the same amount of hard currency, however, does not entail our having parts that represent the same amounts. If I have five $2 bills, and you have two $5 bills, we each have $10 in hard currency, but no constituents of my currency represent $5 and none of yours represent $2. The upshot is that compositionality does not entail (P4). One could consistently reject the latter but accept the former.

Attitudes are, of course, different from currency in any number of ways. One that bears on this premise is that the number of permissible compositions for a given semantic type of attitude is fairly limited. While there are many ways of making a US dollar with hard currency (294, counting a dollar coin and dollar bill), there are not too many ways of breaking a belief that dogs bark into constituent parts. Most intuitively, one would need a concept meaning *dogs* and another meaning *bark*. There could also be a singular concept with no constituent parts that means *dogs bark*. It may also be possible that the concepts meaning *dogs* and *bark* are composed of yet more primitive concepts.

Suppose that each semantic type of attitude can be composed in exactly one way—that composition is *monoeidic*. This would mean that every token attitude of a given semantic type also decomposes in the same way. For every meaningful part of a token attitude, then, there would be a constituent part that means the same thing (has the same content) in every other token attitude of that same semantic type. For example, if the only way to compose a belief with the content *dogs bark* is to combine an element meaning *dogs* with an element meaning *bark*, then every token belief with that content has token constituents that mean *dogs* and *bark*. If we read 'same meaningful parts' as 'constituent parts with the same meaning', then, assuming a single permissible decomposition we get a monoeidic decompositionality, which is just what (P4) asserts.[5]

There is, however, another reading available: that having 'the same meaningful parts' means having meaningful parts *of the same type*. This is a stronger reading. The weaker reading only asserts *content* identity between constituent parts of content-identical attitudes, not type identity. Elements with the same content are not necessarily type identical. Consider the English sentence 'dogs bark' and the German sentence 'Hunde bellen'. Both sentences have the same content (*dogs bark*), and both have constituent parts with the same meaning ('dogs' and 'Hunde' meaning *dogs*, and 'bark' and 'bellen' meaning *bark*). But they do not have the same types of constituent parts: 'dogs' and 'Hunde' are different types of words.

If concepts are individuated in part by non-semantic features, then the weak and strong readings of (P4) come apart. Semantically equivalent attitude tokens

[5]The assumption that each semantic attitude type has a single (de)composition is, perhaps, implausible. It is an assumption we need only if we want to *guarantee* that token attitudes of the same type all have constituent parts that mean the same thing. If there are relatively few permissible (de)compositions, then we would get a weaker version that would say attitudes with the same content *often* have the same meaningful parts. The problem I am going to raise applies to the weaker assumption as well, so for exegetical convenience I will stick with the strong version.

could decompose into parts that have the same meaning, but different non-semantic features.

Recall that the crucial moves in the argument from intentional generalizations were (P4), and this:

P5 If people often have attitudes with the same meaningful parts, then people often have attitudes with token concepts of the same types.

With a strong reading, this premise is almost trivial: Concepts *just are* the meaningful parts of attitudes, so types of concepts are types of meaningful parts. With a weak reading the inference requires an additional assumption: that concepts are individuated purely by their content. If concepts are individuated in part by other features, then—just as we cannot infer that two token sentences with the same content will include the same types of words—we cannot infer that attitudes with parts that mean the same thing include the same types of concepts. Defenders of the argument from intentional generalizations thus must claim either that we have good reason to read (P4) strongly (so that the antecedent of (P5) can be read strongly), or that concepts are individuated purely by their content. Neither is true, as I shall argue presently.

Frege cases

If we take the line that concepts are individuated by content alone, we need to say something about what conceptual content is. The simplest idea is that contents are referents. Famously, however, this leads us to Frege's puzzle about *substitutivity*. Substitutivity says that co-referential terms can be substituted for one another *salva veritate*. So, for example, since 'adrenalin' and 'epinephrine' co-refer, then if 'the needle contains adrenalin' is true, 'the needle contains epinephrine' must also be true. Frege (1892) presented the following puzzle for this intuitive principle (though

not in these terms): Suppose I do not know that 'adrenaline' and 'epinephrine' name the same substance. It would seem possible, then, for 4 to be true, but 5 to be false:

(4) I believe the needle contains adrenaline.

(5) I believe the needle contains epinephrine.

But this would be a failure of substitutivity. Frege's puzzle consists in explaining how, if substitutivity is to be saved, such sentences differ in content (or, at least, why they *seem* to differ in content), or why substitutivity does not apply in these contexts.

Strictly speaking, Frege's puzzle concerns propositional attitude *reports*. My concern here will be with the attitudes themselves. To avoid some complicating details[6] I want to focus on a closely related puzzle that deals with propositional attitudes directly. Consider the following case:

> **Ignorance**: Blair is a hospital intern. Doctor X has asked Blair to clean off a shelf in the supply closet. Blair's instructions are to leave all and only adrenaline shots on the shelf, and discard the rest. A different doctor, Doctor Y, has a standing order to bring any discarded epinephrine shots to her office. Blair finds several boxes labeled 'epinephrine shot' and, despite there being plenty of room for them on the shelf, delivers them to Doctor Y.

How can Blair lack the belief that would cause her to leave the boxes on the shelf, but have the belief that causes her to bring the boxes to Doctor Y? On the one hand, if she believed the boxes contained adrenaline shots, she would have left them on the shelf. Since she did not, she must lack the relevant belief. At the same time, she did bring the boxes to Doctor Y, which she would have done only if she believed the boxes to contain adrenaline shots.

This puzzle is not about ascription. It is about the mental states themselves. We need to explain what it is about Blair that prevents her from drawing the appropriate inference. She has a belief with the propositional content *the boxes contain adrenaline*

[6]Kripke's (1979) puzzles about disquotation, for example, are tied to ascription.

shots, but fails to infer that she should leave them on the shelf (as evidenced by the fact that she does not do this, and the assumption that she intends to cooperate).

The tempting reply is to say that she believes the boxes contain *epinephrine* shots but needs to believe they contain *adrenaline* shots. This, of course, is just the original puzzle over again. Posing the answer in these terms does not help solve the puzzle about attitudes themselves since it is merely a change in ascriptive language. We still need to know how the belief Blair has differs from the belief she needs to make the right choice.

Frege (1892)'s own solution was to posit *senses*, a sort of abstract public object mediating the connection between a representation and what it refers to. While 'the shots contain adrenaline' and 'the shots contain epinephrine' share a referent (a truth value, on Frege's view), he says, they are connected to them by different senses. In ascriptive sentences like 'Blair believes the shots contain adrenaline' and 'Blair believes the shots contain epinephrine', Frege claims the embedded propositions ('the shots contain...') refer to their senses, not their ordinary referents. And this, he says, corresponds to a difference in the sense that Blair "grasps" when thinking one instead of the other.

The problem with Frege's account, as Fodor (1998) points out, is that senses are not doing any of the work. In terms of what's going on inside someone's head, it's the *grasping* of a sense that distinguishes the states. But this metaphor is left unexplained. How does the mental feat of grasping one sense differ psychologically from grasping a different sense? Since explaining the psychological difference between such cases is the very puzzle we are trying to solve, Frege's solution is no help.

Another possible solution is to say that Blair suffers some failure of rationality. She fails to draw the appropriate inference and act accordingly because something has gone wrong with her reasoning. This is a poor response. If she were following some faulty inference rules, we would expect errors to appear in a domain general

fashion, whenever the rule was used. But her pattern of mistakes is not general—it is tied to epinephrine in particular. Nor do we have any reason to suspect she has suffered a cognitive failure of some sort, like forgetting what she is supposed to be doing. Blair's mistake is the product of bad information, not faulty reasoning. Were she to get new information—that 'epinephrine' and 'adrenaline' name the same substance—she would not make the mistake.

A much more promising—and intuitive—explanation is that Blair has two distinct concepts that refer to adrenaline. She does not know that they co-refer—in fact she believes that they do not—and her belief about what she is to leave on the shelf uses one of these concepts, and her belief about what is in the boxes uses the other. So her belief about what is in the boxes does not spur her to leave them on the shelf—it uses the wrong vocabulary, so to speak.

This is possible only if Blair's mental bookkeeping individuates concepts by something besides their content. Her mind must use some non-referential features to distinguish and keep track of concepts with the same content. Call these the concept's *formal* features. Suppose Blair's two *adrenaline* concepts are ⟨adren⟩ and ⟨epi⟩. This allows us to reconstruct Blair's reasoning as follows, where the expressions are to be read as an ersatz mental language with orthography as a proxy for concepts' formal features:

(6) ⟨I should only leave adren shots on the shelf.⟩

(7) ⟨The boxes contain epi shots.⟩

(8) ⟨epi is not adren.⟩

(9) ⟨So, I should not leave the boxes on the shelf.⟩

Formally, this is a consistent, valid argument. Blair is not *irrational* for following this line of reasoning—a virtue of the formal features account.

The reason she gets to a false conclusion is that she is working with bad information. Premise (8) is false. (Note that this premise is *using* the ⟨adren⟩ and ⟨epi⟩ concepts, not *mentioning* them. Blair would need a metalanguage to assert the non-identity of her mental symbols.) This premise asserts that adrenaline is not self-identical. But this does not properly describe her error. Indeed, there is no way to describe her error within one semantically interpreted language. We need to be able to talk about her mental states as an object language. What she does not know is that ⟨adren⟩ and ⟨epi⟩ co-refer. This leads to her using (8) in a practical syllogism.

The second virtue of the "two formally distinct concepts" explanation is that it is both predicted and explained by the dominant computational model of thought. On this model, mental processes are sensitive only to the formal features of mental symbols. The content of a symbol or state is itself causally inert, but the workings of the mind are such that semantic relations are encoded in rules defined over the formal features of symbols.

A system that works like this—one that is blind to semantics—will be susceptible to just the sort of error Blair experiences in our example. When things go right, there is one concept for a particular content. But when multiple concepts refer to the same thing, the mind cannot recognize this *a priori*—it only responds to concepts' formal features. The problematic premise is not a *formal* contradiction. Only under semantic interpretation can we see that it says that adrenaline is not self-identical. So, a computational system that only "sees" formal features will not be able to recognize the error.

Before returning to the morals for the shareability thesis, I need to discuss one more possibility. By dubbing the non-referential features of a concept 'formal features', I am making a tendentious assumption that the difference between co-referring concepts is non-semantic. Another possibility is that there is more to the semantics of singular terms like 'adrenaline' (and corresponding singular concepts) than reference.

On this view, 'epinephrine' and 'adrenaline' (and ⟨epi⟩ and ⟨adren⟩) are distinguished by some extra layer of meaning.

This is not, however, a promising line of reply. It is *ad hoc*, and less parsimonious than the formal-features reply. Computationalism demands that concepts have formal features independently of these Frege cases. So, by relying on a single dimension of meaning and showing how formal features solve the puzzle, we are drawing on resources the theory already requires. Moreover, given computationalist scruples, the extra layer of meaning is otiose. There still needs to be some way for her mind—a physically realized system—to respond to semantic differences. And the only way to do that is if they are encoded in non-semantic features. So, the concepts would still need to be formally distinct.

Formal features and shareability

Let us return to the argument for the shareability thesis. The argument was that intentional generalizations attribute states with the same content to many different people, and:

P4 If two token propositional attitudes have the same content, then they have the same meaningful parts.

P5 If people often have attitudes with the same meaningful parts, then people often have attitudes with token concepts of the same types.

As I argued above, 'same meaningful parts' is ambiguous. (P5) requires a strong reading according to which it means *meaningful parts of the same type*. And this can happen in two ways. One is that (P4) warrants a strong reading. The other is that (P4) only warrants a weaker reading, according to which 'same meaningful parts' means *parts that mean the same thing*, but concepts are individuated by content alone, so the readings collapse.

As we have seen, accounting for Frege cases requires that concepts be individuated (at least in part) by their formal features. So, the second possibility is a non-starter.

The first possibility is likewise unwarranted. The sort of generalizations that motivated shareability proponents are cast in terms of content alone. The formal features do not play any role. Consider the categorization behavior described by Diesendruck and Peretz (2013): If children think of something as an artifact, they will take into account its creator's intent (e.g., the creator's intending it to be a whatsit is relevant to whether or not it gets put with the whatsits). If this is true, then children have a concept that means *artifact* which plays that role (*inter alia*) in their category judgments. The generalization picks out a concept by content alone, and attributes a functional role to it. Any other features that the mind uses to individuate that concept do not enter into the explanation.

Or consider a recent finding by MacNell et al. (2015) that the perceived gender of an instructor has a significant effect on student evaluations. Students in experimental groups were given online courses which differed only in the instructor's name. Student evaluations for courses where the instructor had a feminine name received worse evaluations than (again, identical) courses where the instructor had a masculine name. This provides evidence for the following generalization about social/psychological gender bias: Students who think their instructors are women will tend to rate them lower than if they thought they were men. This generalization makes a number of assumptions about students' gender concepts: that they associate particular names with particular genders, that these are the concepts that enter into other attitudes toward gender, for example. It does not, however, assume that the relevant concepts meaning *men* and *women* have the same formal features. What matters is that people have these concepts, and that these concepts track gender and play certain other roles in our psychology. What features our minds use to identify, re-identify, and distinguish these concepts are irrelevant to the explanation.

Even generalizations that implicate these differences (e.g., that people who associate distinct concepts with 'epinephrine' and 'adrenaline' will tend to ignore boxes labeled 'epinephrine' when asked for 'adrenaline') do not require shared formal features. They only require that for each individual mind subsumed under the generalization, some formal features or other distinguish these concepts.

If we understand the work that formal features are supposed to do, this result should be unsurprising. Formal features play a role in internal bookkeeping. They are what our mental processors use to distinguish the symbols that mental computations range over. The reason our explanatory framework needs to individuate concepts along this dimension is to capture facts about *intra*personal rationality (as in the case of Blair, above).

This is very different from the role of content in intentional explanations. If we want to explain why women suffer a certain sort of systematic disadvantage, we need to talk about beliefs and biases people have about women. If we want to explain a pattern in how children categorize artifacts versus other types of things, we need to talk about the differences in how they think about artifacts versus other categories. These explanations subsume concepts by content—by the *same* content.

If the students in the MacNell et al. (2015) study meant different things by their gender concepts, then there is no generalization: we could not say that there is a bias against *women*, but rather many different biases (against $women_1$ for this student, $women_2$ for this other one, etc.). And this is Fodor's dreaded result: The researcher's use of 'women' would pick out a meaning unique to them, so there would be no way to state the generalization. Indeed, there would be no generalization to capture. By contrast, if one student represents the instructor as a being a woman using a $\langle \text{woman}_1 \rangle$ concept, and another student represents the same thing using a formally distinct $\langle \text{woman}_2 \rangle$ concept, it would still be true that they exhibit a bias against

women, as such, if they conform to the pattern of behavior MacNell et al. (2015) describe.

I am not claiming that such concepts do not have the same formal features across all people. My contention here is only that intentional generalizations do not *require* that the concepts they range over share formal features. The argument from intentional generalizations claimed that explaining behavior requires the existence of true intentional generalizations. It claimed further that for two people to have the same intentional state, they must have the same types of concepts. So, it concluded, people must have concepts of the same type—they must share concepts. As I have argued, however, concepts are individuated in part by their formal features. The truth of intentional generalizations requires only sameness of content, not formal features. This argument fails to provide reasons for thinking concepts must be shareable.

3.4 Communication

The argument from intentional generalizations is only one of the primary supports for the shareability thesis. The other is the idea that shared concepts are required for successful linguistic communication. Undoubtedly, concepts play a role in communication. The question I am concerned with is whether *shared* concepts play any role in communication. One compelling line of reasoning holds that they do. Successful communication requires, minimally, understanding each other's words. We can only do this, the reasoning goes, if we associate the same concepts with the same words. So, successful communication seems to require shared concepts. Prinz (2002) makes this argument explicitly:

> According to the standard picture, people understand each other's words in virtue of the fact that they associate the same (or quite nearly the same) concepts with those words. If no two people associate the same concepts with their words, then communication is impossible. Therefore, concepts must be shareable. (14)

The key premise is that communication requires that we associate the same concepts with one another's words. Call this the *communication hypothesis*. In this section, I will argue that not only is this hypothesis false, but it leads to an outright logical contradiction. Prinz's hedge, allowing for "very nearly the same" concepts is of little help. The clearest candidate for similarity—having the same content—fails to capture key features of concepts' role in communication.

The problems for this hypothesis emerge in pairs of cases like the following:

> **Success 1**: 'Torosaurus' and 'triceratops' name the same type of dinosaur. My nephew is a dinosaur buff, and knows this. For him, these are two ways of expressing the same concept. I, on the other hand, do not know this. I have two distinct concepts referring to this dinosaur, one of which I associate with the word 'triceratops' and the other with 'torosaurus'. He says to me "I want to go to the museum with the torosaurus skull." I know the museum—the skull is very clearly labeled 'torosaurus'—and I take him there.

According to the communication hypothesis the fact that this communication succeeded means that my nephew and I associate the same concept with 'torosaurus'. Call his univocal concept ⟨T⟩, and the concept I associate with it ⟨torosaurus⟩. So, it seems that ⟨T⟩ and ⟨torosaurus⟩ are the same concept.

> **Success 2**: Looking at my nephew's shelf of dinosaur books, I ask him if he has already read *The Mighty Triceratops* today. He wants to communicate that he has, and forestall additional questioning along that line. He says, "I've read all and only my books about triceratops today." I believe that this book is about triceratops, so I correctly infer that he has already read it.

The concept I associate with 'triceratops' is ⟨triceratops⟩, not ⟨torosaurus⟩. The communication hypothesis says that this communication's success depends on our associating the same concepts with the same words. Since my nephew only has the one ⟨T⟩ concept, that means that it and my ⟨triceratops⟩ concept are the same.

So, we are led to the following problem: ⟨triceratops⟩ and ⟨torosaurus⟩ are distinct types of concepts. ⟨T⟩ is the same type of concept as ⟨torosaurus⟩. ⟨T⟩ is the same

type of concept as ⟨triceratops⟩. So, by transitivity, ⟨torosaurus⟩ and ⟨triceratops⟩ are the same type of concept. So, we have a contradiction.

This problem arises, of course, from requiring that people associate exactly the same types of concepts with the same words as one another. If we relax this requirement, then perhaps the problem can be avoided. Note, however, that doing so undercuts the communication hypothesis as a motivation for shareability. The shareability thesis says that distinct people can, and commonly do, have token concepts of *exactly the same type*. The reason for this is that the shareability thesis is supposed to be a constraint on theories that say what concepts are. Concepts must be individuated so as to be shareable, its proponents claim. By backing away from claiming communication requires shared concept types, we give up shareability as a restriction on individuation.

The most obvious alternative to the strict communication hypothesis is this: successful communication requires associating concepts with the same *content* with the same words. This is compatible with both Success 1 and 2. My two concepts are not *content* distinct. They have the same content as one another and as my nephew's, so we avoid the contradiction.

There are two issues with this move. One is that it gives a very incomplete account of concepts' role in communication. As we can see from the following case, it does not provide us with the resources to explain communication failures:

> **Failure** Looking at his shelf of dinosaur books, I ask my nephew if he has already read *The Mighty Triceratops* today. He wants to communicate that he has, and forestall additional questioning along that line. He says, "I've read all and only my books about torosaurus today." I infer that he has *not* read that book, and hand it to him.

My nephew wanted me to believe that he had read the book, but I concluded that he had not. This happened despite the fact that we associated concepts with the same content with his utterance of 'torosaurus', and the instance of 'triceratops' in

the title of the book. Something has gone wrong, but what? It would seem to be that I am drawing on two different concepts where my nephew draws on one. Looking at content alone, however, this error is invisible. And we cannot simply say that the difference in number causes the problem, since my conceptual repertoire is the same here as in the above successful cases. There is some relation that our concepts need to bear to one another that is present in those cases, but missing here.

Counterpart concepts

Lewis (1973) faced an analogous puzzle about individuation across possible worlds: the truth of counterfactuals requires that some identity like relation holds between individuals in different possible worlds, but it cannot be identity. His solution was to replace transworld identity with a context-dependent *counterpart* relation. Borrowing from Lewis, I want to propose a counterpart model of concepts' role in communication: Successful communication requires not that two people associate the same concepts with the same words, but rather that they associate counterpart concepts with the same words.

In many contexts of communication there are certain beliefs that (1) are relevant to the discourse and (2) the communicator expects to share with their audience. Counterpart concepts are concepts that (a) have the same content and (b) play the same role in all such beliefs. More schematically:

> For any context of communication, C, two concepts, $\langle\phi\rangle$ and $\langle\psi\rangle$ are **counterparts in C** *iff* $\langle\phi\rangle$ and $\langle\psi\rangle$ co-refer, and for any belief Φ in the set of contextually relevant beliefs, $\langle\phi\rangle$ occurs in a token belief meaning Φ *iff* $\langle\psi\rangle$ occurs in the same role in a belief meaning Φ.

Note that beliefs are individuated here by propositional content alone. This allows us to avoid problems of cross personal belief identity that would otherwise parallel the problems of cross personal concept identity.

Like Lewis' model, the counterpart relation for concepts is a context-dependent similarity relation holding between distinct individuals. Context-dependence crops up

frequently in communication. The way we choose to refer to things depends on who we are talking to and what beliefs we share with them. I can indicate, for example, that I think Obama is sharply dressed by saying "the president is sharply dressed" only if you share my belief that Obama is the president. We are usually pretty good at choosing useful ways of referring to things. The conceptual counterpart relation tracks something similar to what we are paying attention to when we make these choices. We have to be thinking about things in relevantly similar ways for communication to succeed. We have to have some of the same beliefs, and the concepts we associate with our words have to play the same role in these beliefs. Which beliefs (if any) we need to share will depend on the context in which we are communicating.

Counterparts in Frege cases

The counterpart model can explain the variations in success and failure to communicate in the above cases. In Success 1, there do not seem to be any relevant background beliefs. What my nephew intends to communicate is that he has a particular desire, and he states it overtly. For him to succeed, I only need to pick up what he says. In a simple case like this, then, the counterpart relation is satisfied by concepts with the same content. So, since we associate concepts with the same content with all the words in his utterance, the communication succeeds.

Success 2 is more interesting. My nephew is expecting that I know what the book is about. I could understand his surface utterance without any background, but that is not all he is trying to say. He is trying to tell me something about the book I am holding when he says "I've read all and only my books about triceratops today," *viz.* that he has read it. To pick up on this I need to believe that this book is about triceratops. This belief is thus part of the set of contextually relevant beliefs. My nephew has this belief in terms of his ⟨T⟩ concept. I have the same belief (individuated by content alone) in terms of my ⟨triceratops⟩ concept. These

concepts are counterparts, and we associate ⟨T⟩ and ⟨triceratops⟩ respectively with his utterance of 'triceratops', so communication succeeds.

In Failure, the belief that the book is about triceratops is still relevant. We each still share this belief, mine in terms of ⟨triceratops⟩ and his in terms of ⟨T⟩. But the concept I associate with 'torosaurus' when he says "I've read all and only my books about torosaurus today" is ⟨torosaurus⟩. This is not a counterpart to his ⟨T⟩ concept since it does not occur in the contextually relevant attitude. So, since we failed to associate counterpart concepts with the same word, communication fails.

Putting counterparts to work in other cases

We know that communication is a messy, context-dependent affair. It is also one at which we excel. The counterpart model helps us understand part of how context comes into our discourse. Consider this case from Kripke (1977):

> Suppose someone at a gathering, glancing in a certain direction, says to his companion, "The man over there drinking champagne is happy tonight." Suppose both the speaker and hearer are under a false impression, and that the man to whom they refer is a teetotaler, drinking sparkling water. (256, example number omitted)

Kripke provides this example to draw out a difference between speaker reference and semantic reference, which are relevant to evaluating whether the assertion is true. I want to focus on a different matter: what makes this communication succeed.

Notice that the communication's success does not depend on the truth or falsity of what is said. The speaker could be understood even if the relevant man is miserable. Nor does the communication depend on both parties coming to the same truth evaluation of the utterance: the hearer could understand the utterance and disagree. The reason the communication succeeds is that the hearer is able to pick out the man to which the speaker refers and predicate *being happy tonight* of him.

The counterpart model can help explain how the hearer manages to pick out the right person. In the example, the false belief that the man's glass contains champagne

is contextually relevant. Because the *champagne* concepts in each party's belief are associated with 'champagne', they are counterparts in this context.

Suppose we change the example so that the hearer does not believe that the glass contains champagne. Does this mean that communication will fail? Not necessarily. Again, what's important is that the hearer is able to pick out the right person. There are a number of different ways this could happen besides sharing that false belief. The speaker may, for example, indicate the drinker by nodding in his direction. In this case, the hearer is able to pick out the right person by this non-linguistic move. The belief that the man is drinking champagne is bypassed entirely—it is not relevant to the discourse. Since the concepts both parties associate with 'champagne' co-refer, that is enough in this scenario for them to be counterparts.

Another possibility is that the hearer looks around the room and, despite knowing that the drinker is a teetotaler, understands that *that* is the indicated man because he is the only one who *looks like* he's drinking champagne. Here the communication does initially fail—the hearer lacks the relevant belief, and thus a counterpart *champagne* concept. Because of this failure, he looks around the room. When he does, he shifts the context slightly, and makes a different belief relevant: that the substance in the glass *looks like* champagne. The first belief, that the substance *is* champagne, is no longer relevant. The parties have counterparts because they share the belief about the appearance, and associate the *champagne* concept occurring there with 'champagne' in the utterance.

Counterparts may also help solve Kripke's (1979/2008) famous puzzle about belief. Kripke presents two independently plausible principles about the relation between language and belief and argues that they are incompatible. One is the *disquotational principle*[7]:

[7]I am not considering the strong disquotational principle, which changes the conditional to a biconditional. The model of thought I have been discussing suggests that this is false. Someone may

> For any sentence ⌜p⌝ of a language L, if an L speaker, on reflection, sincerely assents to ⌜p⌝, then he believes that p.

The other is the *translation principle*:

> If a sentence of one language expresses a truth in that language, then any translation of it into another language also expresses a truth in that other language. (440)

To show that these principles are incompatible Kripke imagines the case of Pierre, a monolingual French speaker, who hears about the beauty of London while living in Paris. Pierre will sincerely and on reflection assent to French language assertions that London is beautiful:

(10) *Londres est jolie.*

By the disquotational principle we would conclude that:

(11) *Pierre croit que Londres est jolie.*

By the translational principle we conclude that:

(12) Pierre believes that London is pretty.

Now, Kripke supposes, Pierre moves to London, but does not know that the city he lives in is the one he calls *'Londres'*. He learns English from the locals without translating to French, and learns that he lives in a place called 'London'. Because his neighborhood is unattractive, he would assent to:

(13) London is not pretty.

So, by strong disquotation, we conclude that:

---

fail to assent to a sentence of a language they understand even if they have a belief with the same propositional content as the sentence. They may do so, for example, if the belief uses concepts they do not associate with the words in the sentence.

(14) Pierre believes that London is not pretty.

Though (14) does not *contradict* (12), there is a puzzle here. Through plausible principles about belief reports, we are led to characterize Pierre as believing that London is and is not pretty. This portrays Pierre's beliefs in a way that does not accurately capture what is going on in his mind.

Thinking in terms of counterparts can help explain why this happens. While both (12) and (14) are *true*, whether or not they are *apt characterizations* of Pierre's beliefs depends on context. In describing Pierre's beliefs we need to establish a counterpart relation between the concepts in Pierre's relevant belief(s) and the concepts our audience uses to understand our description. For (12) to be apt, for example, the concept referring to *London* in Pierre's belief that *London is pretty* needs to be a counterpart to the concept our audience associates with 'London' in the context of communication.

So, (12) may be apt if, say, we want to explain why Pierre has a painting of the Palace of Westminster. A contextually relevant belief, individuated by propositional content, is that the painting depicts London. Our audience believes this, as does Pierre (using the concept he associates with '*Londres*'), and so the word they (our audience) associate with 'London' is a counterpart to the concept referring to London in Pierre's belief that London is pretty.

By contrast, (12) would be inapt if we were reporting on Pierre's beliefs to his neighbors in London. A contextually relevant belief in these contexts would probably include one with the content *Pierre lives in London*. But Pierre's *London* concept in his belief about where he lives is distinct from the *London* concept in his belief that London is pretty. So, to an audience that believes Pierre lives in London in a context where that belief matters, the concept they associate with 'London' is not a counterpart to Pierre's in the belief (12) describes. So, it is inapt.

In the context of this chapter, where we know about Pierre's mental states, the belief that *'Londres'* and 'London' co-refer is contextually relevant. So, our *London* concept is not a counterpart with either of Pierre's. That is why neither (12) nor (14) seems apt, despite both being true.

What I hope I've shown is that the counterpart model is a useful way to think about interpersonal concept relations. It gives us a way to "share concepts" without *sharing concepts*, and so avoids the problems I raise for cross-personal concept identity, while still respecting the idea that there must be something in common between people's concepts if they are going to talk to one another. More than that, its application in Kripke's (1979/2008) Pierre case suggests that it may provide new inroads to some old puzzles about the relation between language and belief.

Conclusion

My main purpose in this chapter has been to erode support for the shareability thesis. I argued that the truth of intentional generalizations does not give us reason to think that concepts are shared, and that the success of linguistic communication *cannot* depend on shared concepts.

This has important consequences for theories of concepts. It allows us to reconsider theories that have been dismissed for rendering concepts unsharable. Prinz (2002), for example, dismisses atomism on these grounds. Atomism holds that concepts are unstructured mental objects that are individuated by their formal features. Prinz argues that there is no way to understand what a formal feature is such that it both accounts for Frege cases, and allows for multiple people to (commonly) have the same concepts (2002, 96-8). If I am right, then there is no reason to do both.

Not all theories that have been dismissed for violating the shareability constraint are vindicated. Fodor and Lepore's (1992) arguments against *semantic holism* still hold. They argue that semantic holism renders *content* unshareable (or at least very

unlikely to be shared). As I have stressed throughout, we need to distinguish sharing *concepts* from sharing *content*. The truth of intentional generalizations requires the latter, as does the counterpart model of concepts' role in communication. Shared content has an important role to play in understanding human behavior; shared concepts do not.

CHAPTER 4

MENTAL BOOKKEEPING: SYMBOL TYPES IN THE LANGUAGE OF THOUGHT

The topic of this chapter is mental bookkeeping. It sounds rather dry, admittedly, but bookkeeping is central to LOT. For example, as we have seen in previous chapters LOT explains Frege puzzles by positing bookkeeping errors—having multiple distinct symbols that refer to the same thing. Or, more centrally, intentional laws are supposed to be underwritten by laws that hold between symbols; so, the fact that such laws exist depends on the mind's ability to keep track of its symbolic vocabulary. I offer an account of how the mind recognizes and distinguishes its basic symbolic elements. More specifically, I develop an account of what sorts of properties lexical elements in the language of thought need to share in order to count as symbols of the same type.

LOT symbol types are often described by analogy to orthographic types in written language. If mentalese (the language of thought itself) were English, for example, LOT would say that the mind distinguishes thoughts about dogs from those about cats by distinguishing 'dogs' from 'cats'. The analog to orthographic properties are the "formal features" of mental symbols—some or other non-semantic properties these symbols possess. The goal of this chapter is to discharge this analogy, and say what formal features are and, by extension, how symbol types are individuated in the mind.

On my view, symbol types are physical types. Formal features are physical properties that mental processors can detect and distinguish, and symbol types are sets of such properties. Call this the *physical properties hypothesis* (PPH). I argue that it offers a simple explanation of what it is to be a mental symbol of a certain type, and

explains how formal features play their distinctive role in LOT: allowing mental processors to follow rules defined over symbol types without the need for interpretation.

In what follows, I set out the view and consider two arguments against it: one from Pessin (1995), who argues that PPH undermines the systematicity of thought, and another that stems from an analogy to written orthography, where the physical diversity of symbol types makes PPH looks particularly implausible. Neither argument succeeds. Pessin's worry stems from an overly narrow conception of physical types, and the orthographic worry ignores a crucial difference between written languages and computational languages: that the former are compiled, and the latter are not. I then consider an alternative proposal, Schneider's (2011) view that symbols are individuated by computational role. I argue that there can't be the sort of roles Schneider appeals to unless something like PPH is true, and that her view legislates where PPH explains. At the very least, then, PPH deserves serious consideration in the literature.

4.1 The explanatory target

I want to offer a theory of what symbol types are in the language of thought. The sparse literature on this topic supplies three general strategies. One, which I will defend in the following section, is to say that LOT symbol types are individuated by the local features of symbols—the sorts of properties that a symbol can instantiate in isolation (having a certain shape, for example). Another, which Pessin (1995) discusses, is to say that formal features are semantic features. As discussed in previous chapters, this is a non-starter (this is the moral of Frege cases). The third, which Schneider (2009, 2011) endorses, is to identify formal features with role. The model for this is logical vocabulary. It seems that we can individuate the logical elements in a formal language by giving a truth table—a specification of their role. Perhaps, the thought goes, all elements in a formal language can be individuated in this way.

The role strategy is untenable, as I shall argue in §4.3 ff. It is tempting, however, because it is easy to lose sight of what we mean by "symbol types." The phrase can refer to many different sorts of linguistic elements, at various levels of abstraction. And once we fix the reference, there is still room for confusion on what a "theory" of them is supposed to do. To avoid confusion, I want to begin by introducing some distinctions.

The target here is *not* what I will call "*grammatical types*": linguistic elements like lexical category (e.g., noun, verb), grammatical class (e.g., sortal noun, mass noun), functional category (e.g., agent, patient), or grammatical relation (e.g., subject, indirect object).[1] Grammatical types are likely individuated by syntactic roles. If so, what makes 'Venus' the subject of 'Venus is uninhabitable' is that it is playing a certain sort of role in the sentence—a role defined by the language's rules of composition.

Again, I am not trying to offer a theory of what grammatical types are, or what makes a symbol fall under a grammatical category.

What I *am* interested in is what distinguishes the things that fall under a category. 'Venus' and 'Hesperus' are both nouns, both names, etc. Yet they are distinct types of symbols. For clarity, let's say they are different *lexical types*.

Lexical types are the sorts of things that are supposed to be distinguished by symbols' formal features. ('Symbol' refers to a particular token instance of a lexical type.) In LOT, symbols with the same formal features are the same lexical type. Formally distinct symbols are different lexical types.

This is not true of grammatical types. Symbols with the same formal features can be different grammatical types, and symbols with different formal features can be the same grammatical type. Supposing for a moment that formal features are ortho-

[1]Category examples from Bickford and Daly (1996, F4, p. 2).

graphic features[2], 'Venus' in 'Venus is uninhabitable' is the subject of the sentence, while a formally indiscernible symbol is the object of 'Callippus observed Venus'. And, of course, 'Hesperus' and 'Venus' are formally distinct though they both fall under the same grammatical type of proper nouns.

The target of this chapter is to say how lexical types are distinguished in the language of thought, and what makes a symbol fall under a certain lexical type. Or, since lexical types are individuated by their formal features, the goal is to say what formal features are.

Again, this is dry. But it is important. The ability to recognize and distinguish lexical types is central to the mind's ability to follow formally defined rules.

Suppose I believe:

1. Venus is uninhabitable.

2. Venus orbits the Sun between Mercury and Earth.

3. Mars orbits the Sun between Earth and Jupiter.

According to LOT, these beliefs exist in my mind as sentence-like structures in a computational language. A constituent part of (1) is a symbol referring to the planet Venus. Let's name that particular symbol token 'Venus$_1$'. The second belief on the list also contains a symbol referring to Venus. Call it 'Venus$_2$'. Suppose that on the basis of these two beliefs, I infer that there is an uninhabitable planet orbiting the Sun between Mercury and Earth. That means that my mind recognizes that the Venus symbol in the second belief is the same type of symbol as the one in the first belief. So, there is some lexical type—call it '⟨VENUS⟩'—of which Venus$_1$ and Venus$_2$ are tokens. I do not infer that there is an uninhabitable planet between Earth and Jupiter on the basis of these beliefs. That means the symbol referring to Mars in (3)

[2]They are not. See §4.2 for why the common analogy between formal features and orthography is inapt.

(call it 'Mars$_1$') is of a different type than these other two (call that type '⟨MARS⟩'). ⟨VENUS⟩ and ⟨MARS⟩ are, however, both members of the same grammatical type. My mental rules of composition would treat them both as valid substitutions for x in constructing a belief of the form ⟨x IS UNINHABITABLE⟩.

In addition to their *being* the same lexical type, Venus$_1$ and Venus$_2$ must be recognizable as such to my mind. According to LOT, mental operations are governed by formally defined rules. For example, suppose my mind follows *modus ponens*, which says to infer Q if (P and *if P, then Q*). 'P' and 'Q' range over symbol categories (or syntactic categories). Following this rule requires recognizing consistent substitutions for these category variables. My mind needs to be able to recognize when the same symbol type occupies both 'P' variable places. If Venus$_1$ and Venus$_2$ were the same type, but my mind could not recognize this, then we could get a consistent substitution for 'P' (e.g. ⟨if Venus$_1$ is uninhabitable, then it isn't worth exploring.⟩&⟨Venus$_2$ is uninhabitable⟩), but the inference would not go through. So, the formal rule would fail to govern my mind's operation.

Indeed, if LOT's explanation of thought as formal rule following is going to work, the mind must be *guaranteed* to treat symbols of the same lexical type as such and to be able to treat distinct types differently. Formal rules range over all possible consistent substitutions for their variables. A system doesn't count as following *modus ponens* if it would systematically fail to derive Q on the basis of P and (*if P, then Q*) for certain values of P. (There could, of course, be flukes: if the system is run over by a steamroller mid computation, that would not count against its following the rule. Hence, only the presence of systematic exceptions undermines the rule.) Likewise, if a system were unable to treat symbols of different lexical types as such, then a formal rule like *modus ponens* could not describe its operations. The rule it actually follows allows a mismatch of symbol types in at least that one instance.

This is not just an epistemic problem of figuring out which rules describe the mind's operations. The possibility of a systematic failure to recognize symbols of the same lexical type as such would mean that no formal rule could truly describe its operation (there would always be the exception). Which means that it would not be following formal rules.

This entails that the features *individuating* lexical types must be, at a minimum, co-extensive with the features that allow the mind to *recognize* symbol types. If the mind—or more precisely, mental processors—could not recognize a difference between distinct lexical types, then it would not be able to treat them differently. And, if they could fail to recognize type identical symbols as such, then they would not be guaranteed to be treated as the same type.

So, formal features must be processor-detectable properties of some sort. Formal features mark symbols as being the same or different type in a way that the processor can recognize. If they don't do this, then formal rule following just doesn't get started.

The goal of a theory of mentalese symbol individuation, then, is to say what sorts of properties these are, and how symbol types are defined over them.

4.2 The physical properties hypothesis

I would like to defend a quasi-reductive account, which I earlier called the *physical properties hypothesis* (PPH). Formal features, on this view, are certain neural/physical properties of symbols, and lexical types are sets of such properties. This is a *quasi*-reduction because while it identifies lexical types with something physical in a sense, properties—even physical properties—are still abstract. Particular symbols are concrete particulars, and they count as tokens of a certain type if they instantiate all the properties that define that type, but the types themselves are still abstract sets of properties.

While all of a symbol's formal features are physical properties, for reasons I will make clear in the next section, not all of a symbol's physical properties are formal features. Only certain processor detectable properties are. These are properties that, if they differ between symbols, make it possible for a processor to treat these symbols as inconsistent substitutions in a formal scheme (or as different elements in a logical vocabulary).

Identifying lexical types with sets of processor detectable physical properties implies a very fine-grained system of symbol individuation. Being an F-type symbol, according to PPH, is a property defined relative to a particular mental processor. It is a certain set of physical properties that symbols in a processor's lexicon can have, such that if two symbols share them, it cannot detect any difference between them. This means that while multiple people can (and surely do) share grammatical types (e.g., subject, object; noun, verb; etc.), their lexicons may contain none of the same lexical types. Indeed, within a mind, different modules with their own processors may have no lexical types in common.

So, for example, while you and I may both have *a* symbol for Venus (that is, a symbol referring to Venus), there is likely no such thing as *the* symbol for Venus. It would be something of an accident if there were. This consequence of PPH coheres with the lessons from chapter 3 about sharing concepts. The possibility of Frege cases means that we should not expect there to be univocal lexical types. Assuming there was such a thing as *the* Venus symbol is exactly what led to the logical problems for shared concepts accounts. We want it to be possible for you to have a symbol that "looks like"[3] 'Venus', and for me to have one that looks like 'Phosphorus' and for these to be different types of symbols. That is—it would be possible for my mind to systematically treat a symbol that looked like 'Venus' differently from one that looked

[3]I am using "looks like" here as a stand-in for instantiating a particular set of physical properties.

like 'Phosphorus' and *vice versa*. PPH achieves this by providing an account which says a particular set of processor detectable physical properties is a/the Venus symbol for a particular processor. Whether or not symbols in different minds or modules count as *counterparts* depends on what criteria are relevant for the comparison. But the notion of "same symbol" is only defined relative to a single computational system.

The same is true for logical vocabulary. This is worth noting since logical vocabulary is typically thought to be defined by role. So we might think that to be *the* conjunction symbol is to play a certain role in a system. But we need to be careful. A theory of lexical types attempts to offer the *individuating conditions* of symbols, not the *constitutive conditions* on being a symbol of a certain grammatical or semantic type.

While individuating conditions and constitutive conditions can sometimes align, they need not. What it is to be a first baseman, for example, is to play a certain role on a team. But how first basemen are individuated—how you tell them apart from one another and from other players—is a matter of what they look like (e.g., the colors of their uniform and the name and number on the back).[4]

Likewise, what it is to be *a* conjunction symbol may well be to play a particular role. So, what makes '&' a conjunction symbol in some system is that it plays that role. By contrast, what individuates the '&' symbol, according to PPH, is that it looks like '&'.

In a different system, a '∧' may play the same role. Are they the same lexical type? According to PPH there is no answer to that, *simpliciter*. The question would need to be whether they are the same lexical type within a given system. And the answer would depend on whether that system's processor could recognize a difference between something that looked like '∧' and something that looked like '&'. They would be the

[4]Thanks to Louise Antony for the example.

same type of symbol *iff* it were impossible (barring architectural changes) for it to systematically treat 'P & Q' differently from 'P ∧ Q'. This would be a rather strange circumstance. There is nothing mandating that a symbol that looks like '&' must be a conjunction symbol, nor one that looks like '∧'. So, unless the difference between them is invisible to a processor, it will be possible for it to co-opt those symbols for distinct roles.[5]

All of this falls out of the fact that while we often talk about symbol types by reference to content (or logical role), that cannot be how symbols themselves are individuated—of what marks a symbol as being of a certain lexical type. Lexical types are the bottom-level distinction that processors can make. And they must be guaranteed to treat symbols of the same lexical type as such. The fact that a certain symbol plays a particular role or has a particular meaning does not mean that it *must*. So, having meaning or role in common is not sufficient for two symbols to be the same type of symbol.

This is not to say that the *semantics* of logical vocabulary could not be fixed by role. Very likely, what it is to be a symbol *for conjunction* is to play a certain role, or be described by a certain truth table. Rather, the claim is that role does not tell us how a particular lexical type of symbol that does stand for conjunction is individuated—how a system recognizes its instances, and distinguishes them from other items in its lexicon.

I linger on this point to make clear just what identity claim PPH is making. It is not saying that, for example, to be a conjunction symbol is to be a physical type. "Conjunction symbol" picks out a category that subsumes a variety of different lexical types. If it is defined by playing a certain sort of role, any symbol type that plays this

[5]This leads to an interesting possibility. If a computational system could have two distinct symbols playing the same logical role, there could be Frege cases for logical vocabulary. Such a system might, for example, believe ¬P and not-Q (two different lexical symbols for *not*), but fail to infer either ¬(P or Q) or not-(P or Q).

role counts as a conjunction symbol. Since, according to PPH, lexical types are sets of physical properties, the "conjunction type" of symbol is therefore multiply realized by a number of different physical types. Likewise for the Venus symbol (if any sense can be made of that category).

Again, we need something finer grained in order to make LOT work. It must be the case that different types of symbols can have the same content, or else we lose the ability to explain Frege cases. The fact that the same distinction between semantic types and lexical types would apply to logical vocabulary is perhaps not necessary to LOT, but neither is it a problem.

PPH claims that *lexical types* are physical types. To be a symbol of a certain lexical type within a system is to instantiate certain physical properties—those detectable by that system's mental processors—which all other symbols of that type also instantiate.

Pessin's argument against PPH

LOT claims that thought exhibits *systematicity*. That is, the ability to think certain thoughts entails the ability to think certain other thoughts. For example, someone who can think that John loves Mary can think that Mary loves John. This, LOT says, is because these thoughts employ the same lexical elements and the same rules of composition.

Pessin (1995) argues that PPH is incompatible with the systematicity of thought. His argument goes roughly like this: Within a mind, either the thoughts that John loves Mary and Mary loves John are physically distinct, or they are not. If not, then, either the mental does not supervene on the physical (in which case, why bother with the reduction?), or *contra* possibility the thought that John loves Mary just *is* the thought that Mary loves John. So, there must be some physical difference between them. This physical difference, Pessin reasons, must be a difference between the symbols that occur in each thought. There must be some physical difference that

marks the Mary symbol as the object in the first thought and the subject in the second, he says. But that means that the symbols standing for Mary in each thought are distinct physical types from one another. And if symbol types just are physical types, that means there are two different types of Mary symbols. Which means that the thoughts don't use the same lexical elements after all. Which means that the ability to think one does not necessitate the ability to think the other. Which means we don't get systematicity (1995, p. 36-7).

Pessin's argument seems to hinge on the idea that we cannot physically realize symbols and syntax. But this can't be right. Symbols and syntax can, of course, be physically realized, and they can be realized in different ways. If they couldn't, then there wouldn't be computers.

The paradigm here is that of formal languages in logic. The recursive rules of a productive and compositional language specify how new, syntactically distinct elements can be formed using the same lexical elements. The lexicon specifies what elements there are, and different lexicons delineate different grammatical categories. The rules specify how elements from the various categories can be combined. So a system might follow a rule allowing it to combine two-place propositions with any two names like so: <proposition><name><name>. If 'F' is an element of the propositional lexicon, and 'a' and 'b' are elements of the name lexicon, this system can generate the syntactically distinct formulas 'Faa', 'Fab', 'Fba', and 'Fbb'. There are certain physical features that allow a system following this rule to recognize the 'a' elements as the same lexical type, and a different set of physical features that allow it to recognize its syntax. It makes one set of distinctions on the basis of its rules, and the other on differences in the lexicon.

So, to Pessin's case, while there must be physical differences between the thoughts that John loves Mary and that Mary loves John, it needn't be a difference between the symbols—it could be a physical difference in syntax.

Pessin's error might be due to thinking that *all* of a symbol's physical properties are individuative of it as a symbol type. If that were true, then the formulas 'Fab' and 'Fba' would contain different symbols because 'a' in 'Fab' bears different physical relations to 'F' and 'b' than the 'a' in 'Fba'. But while one *could* individuate lexical types in this way, there is no reason one should. Some of a symbol's physical properties mark it as being a certain lexical type, others mark it for syntax, and others still are irrelevant to the computational system.

One might object that this renders PPH circular in the following way: PPH says that lexical types are sets of physical properties. Which physical properties? Well, not all of them (the exact number of atoms a symbol contains, perhaps). And not even all of the ones that are relevant to the computation (there are some that mark syntax). Just those that individuate lexical types.

To avoid this, what we want is a principled difference between the physical properties that individuate lexical types, and those that are still relevant to the computation, but not individuative. Here's one: the difference between relational and non-relational properties. What makes a symbol token the type of symbol it is would be its non-relational (with respect to other elements of the system), processor detectable, physical properties. What makes a symbol token a subject or object of a sentence depends on certain relations it bears to other elements in the system.

Pessin dismisses this possibility, stating:

> [I]f the neural state corresponding to "John" is instantiated at the same time as that corresponding to "Mary," how could some other neural state dictate the symbol "John"'s grammatical role? It would seem you'd have to mark that symbol somehow; but while in English we mark the symbol by its location in the sentence, that option just isn't plausibly available where the tokened symbol is identical to a neural state. (1995, 37)

But why shouldn't there be something analogous to word order to mark a symbol's syntax? Term order works just fine for other physically realized formal languages. And if not term order, there are plenty of other physically realizable properties for

marking syntax (e.g., adding an -s or -o to the end of the symbol). Again, the pudding proof is that computers actually do this stuff.

This passage indicates the second possible source of Pessin's error. He seems to be implying that the only neurally realizable syntactic property is concatenation. The mental symbols 'Mary', 'John', and 'loves', *qua* neural states, he assumes, can bear the co-instantiation relation to one another, but nothing else. If that were true, then mental logic would be reduced to saying Mary & John & loves, and we would be in a bind.

There is, of course, no reason to think that the only computationally relevant relation between neural states is co-instantiation. The same lexical type of symbol could be instantiated in different brain regions, or at different times, or in different proximities to other symbolic elements.

Here's a toy example. Suppose that 'A', 'B', and 'C' name physical regions of a computational system. These act as "memory registers." Each consists of two "bits"—two physical switches that hold electrical charge. Let's say a bit is in state '1' when the charge is above some threshold, and '0' when below. The table below represents a possible, synchronic physical state of the system.

A		**B**		**C**	
0	0	0	0	0	0

In this example, every bit in every register is in state 0. The contents of each register, then, instantiate the same type of physical state: having two bits in state 0 (or, each is in state 00). In this system, lexical types are the different physical types of register contents (being in state 00, being in state 01, etc.). Suppose that a symbol's syntax is marked by which register it occupies. The values of register A, whatever they are, are grammatical subjects, and those of register C are objects. So, both lexical types and grammatical types are physically realized, but they are determined by different physical properties.

If 01 stands for Mary, 10 stands for John, and 11 stands for love, it will represent the claim that John loves Mary like this:

A		**B**		**C**	
1	0	1	1	0	1

And the claim that Mary loves John like this:

A		**B**		**C**	
0	1	1	1	1	0

The same lexical types (*qua* physical types) occur in different ways within the system. To be a token of the Mary symbol in this system is to be a register value of 01. This occurs in one way in the first instance (instantiated by register C), and a different way in the second (instantiated by A). The similarities are between lexical elements. The differences mark syntactic differences.

It is clear how systematicity falls out of a system like this. Since all of the registers can instantiate the same symbol types, then they can contain any combination of them that its rules of operation allow. As long as the system has recursive rules that range over grammatical categories, it will be able to follow those rules to construct both variations of the sentence.

Moreover, we know that such a system *would* work because this is how computers *do* work (well, roughly anyway).

Compiled vs non-compiled languages

The second objection to PPH I want to consider stems from the common analogy between orthographic features of written symbols and formal features of computational symbols. While there do seem to be some similarities—orthographic features are non-semantic, for example—there is good reason to think that orthographic symbols should not be identified with *any* of the physical properties their tokens possess.

This is because there might be no single set of physical features all symbols of a type have in common.

We might think that what unifies orthographic types is something about the shape properties of their symbols. But, on reflection, word symbols of the same type can come in all sorts of shapes. In English, for example, 'Venus', '$_{Venus}$', and '*VENUS*' all count as tokens of the same type even though they are physically distinct. More distinct still are the highly stylized symbols used in graffiti or calligraphy.

This looks bad for the physical properties hypothesis. If symbol tokens do not need to share a core set of physical properties (shape properties, in this case), then types of orthographic symbols could not be individuated by such properties. The worry, then, is that PPH isn't true for natural languages because there is no consistent set of physical properties shared by symbols of the same orthographic type. So, if PPH is to work for LOT, we need to have good reason to think that LOT is not subject to this sort of variation.

We do. In fact, we have good reason to think that lexical types in LOT *cannot* be subject to the same sort of physical variation as orthographic symbols in a written language. This type of symbol variation is only possible for a *compiled* language. And, while natural languages are compiled, LOT is not.

Compilers, in the computer science sense, are programs that have rules for transforming statements in a human readable programming language into a machine readable language. The processor never "sees" expressions in the higher level language. All it ever has to read or write are the expressions in its own language. For present purposes, compilers perform two crucial functions: (1) they translate expressions, and (2) they map discernibly different symbols of the same lexical type in the higher level language onto indiscernible symbols in the processor's language.

The translational function allows a computational system to execute commands from various higher level languages. Without compilers, programmers would be lim-

ited to a single computational language for each type of computational architecture—machine language. Compilers, then, allow a processor to (indirectly) use a variety of languages.

Compilers also allow symbol variation *within* languages. For example, the programming language BASIC is a case-insensitive language. The command '*goto*' in BASIC is exactly the same as '*GOTO*'—they are just different ways of writing the same symbol. This is not the case for a language like C, which treats upper- and lowercase letters as different alphabets. Writing a '*goto*' command in C will not generate the same result as a '*GOTO*' command. This is a difference in how these languages are compiled. BASIC compilers translate both '*GOTO*' and '*goto*' into the same element in the machine language. A BASIC compiler hides the differences from the processor. Whether or not the symbol is written in upper- or lowercase in BASIC, the compiler ensures that the processor sees exactly the same symbol (or symbolic function). While there is variation in the higher level language, there is no corresponding variation in the language the processor actually uses. The C compiler, by contrast, encodes the difference in the higher level language into a difference in the lower level language.

The reason, then, that there can be discernible symbols of the same type in a higher level language is because the compiler maps them on to the same symbol type in the processor language. And here's the catch: Since the symbols of a processor language are not compiled onto lower level languages—these are the bedrock languages of computation—there is no lower level of symbol typing to unite discernibly distinct processor symbols.

This is the difference between orthographic types and formal types in LOT. Our minds are compilers for natural language expressions. They take things written or said in a high level language (a natural language), and transform them into a lower level language, the language of thought, which our mental processors can read. LOT

is at the bottom of this process. There is no sub-language of thought. If there were, then *that* would be the language of thought. LOT is whatever is at the bottom; it can't be turtles all the way down.

We can understand the compilation process for orthographic expressions as having two components. One is a process of symbol *recognition*—of categorizing a visual input as being a symbol of a certain type. The other is symbol *comprehension* where that representation of symbol type is associated with our concepts for the symbol's content. So, we see 'Venus', recognize it as a 'Venus' type symbol, and associate it with a concept referring to *Venus*.

What concerns us here is the recognition aspect. Word recognition has been studied extensively in cognitive science.[6] The general picture is that our brain transforms a visual input into a representation of its physical features, matches those physical features to stored representations of abstract letter identities, and matches that pattern of letters to representations of orthographic types in our orthographic lexicon. So, recognizing 'Venus' is a matter of representing the markings out there in the world as containing 'v', 'e', 'n', 'u', 's', in that order, and associating that pattern of representations with our stored representation of the 'Venus' symbol type.

Our letter recognition system can match many different physical patterns to the same mental representation of letter type. We see an 'e' on the page, and the system says "that matches very nicely with the shape pattern I associate with the letter *e*—let's say that's an *e*." We see a colorful, flowing design sprayed on the dumpster out back, and the system says "hmmm, I think these are letters... the best match for this symbol is with the pattern I associate with the letter *e*—so *e* it is." This ability to recognize a variety of patterns as instances of the same letter type in turn

[6]See Rastle (2007) for an overview.

allows a combinatorial explosion of associations between physical patterns and mental representations of orthographic types.

If tokens of orthographic types don't need to have physical shapes in common, what makes them count as instances of the same type? We do. It's the fact that we represent them as containing the letters they do that makes them the same type. Orthographic symbols, therefore, depend not on their physical properties for their identity, but rather on their relation to the minds of the symbol using community. What makes the varying written instances of 'Venus' instances of the 'Venus' *type* is that we associate them with mental representations referring to that orthographic type. That is, to be 'Venus' is to be thought of as 'Venus'.

Our mental processors ignore the physical differences between type identical symbol tokens in a compiled language because *they never actually see them.* The inputs 'Venus' and 'VENUS' are mapped onto the same representation by the compiler, and the processor only ever sees that output representation. (*We* see the differences, but, of course, we are more than just language compilers.) So, a processor reading a compiled language is guaranteed to treat physically distinct, but type identical, symbols in the same way because the rules of its compiler ensure there is no difference for it to see.

Nothing like this is available to symbol types in an uninterpreted language. If symbols are distinct in ways a processor can detect, then there is no reason it *must* treat them as the same. So, for an uninterpreted language, whatever processor detectable features a symbol token possesses must be shared by all other tokens of that lexical type. So, at the very least the property of having that set of features is co-extensive with the property of being a symbol of that type.

This paints a picture very amenable to the physical properties hypothesis. If processor detectable features are just physical features, and if tokens of the same lexical type must possess the same set of processor detectable features, then we have

a a stable physical type for reduction. No other features need to enter into symbol individuation.

The more general upshot is that if PPH is true, we know how the mind recognizes tokens of the same type: because it recognizes physical features, and there are literally no (non-relational) physical differences between the symbols that it can see. And we know why it is guaranteed to be able to distinguish distinct symbol types: because they are physically distinct in exactly the ways that a processor can recognize.

We have good reason, then, to take PPH seriously. Contra Pessin, it does not preclude the systematicity of thought, and, because LOT is not a compiled language, it will exhibit just the sort of symbol uniformity PPH needs.

4.3 Schneider's "total computational role" account

Schneider (2011) advances a competitor to PPH. She argues that lexical types are causal/computational roles. That is, to be a symbol of a certain lexical type, on her view, is to play a certain causal role.

More specifically, she says that mental symbols are defined by *Ramsifying* the rules that describe the mind's operations (2011, 120-3). Ramsification is a method of defining a term by conjoining all of the claims a theory makes about it, replacing the term with a variable, and saying that what the term means is the role played by the variable in the sentence.

Casting the account in terms of Ramsey sentences immediately leads to problems. Ramsification is a process of semantic definition, not symbol individuation. It tells you what a certain term means by the role it plays in the meaningful claims of a theory. If you want to know what an electron is, for example, you yank 'electron' out of all the meaningful claims physics makes about electrons, put in an x, and say that an electron $=_{df}$ the thing that does all the theory says x does. This would not tell you what symbol physicists use to mean *electron*; it tells you what 'electron' means.

Sometimes it seems that this is what Schneider has in mind. In one passage, she states the account by saying that "[a] symbol *is defined* by the role it plays in the algorithm that describes the central system" (2011, 121, emphasis mine). If she *were* offering an account of how symbols are defined—that is, what their semantic content is—she could use Ramsey sentences as she does. (Though it would be a wildly implausible account.) But this is not her project. Instead, she claims she is offering a non-semantic theory that identifies the sorts of things that can serve as neo-Fregean modes of presentation (133). That is, of symbols as formally individuated objects—what I have been calling lexical types.

Let's suppose that there is some sense to be made of Ramsification as a method for lexical symbol individuation. If we are going to evaluate Schneider's account, we need to know what sorts of roles are supposed to individuate lexical types. That is, we need to know just what is being Ramsified. Unfortunately, Schneider is rather elusive on this point. She initially states that the sentence ranges over "all the algorithms that a final cognitive science employs to describe the workings of the central system" (121-2). This suggests that we are looking at a fairly abstract rule set. Describing the workings of central cognition algorithmically would be akin to describing the construction and derivation rules of a formal logic system.

The problem with this interpretation is that these algorithms are cast in terms of grammatical types, not lexical types. They tell you that you can combine a noun phrase with a verb phrase, or that you can derive some proposition Q from P and *if P, then Q*. At best, this would give us a distinction between elements in the logical vocabulary, and would tell us that the language contains things like nouns, verbs, subjects, objects, names, etc. It could not give us a distinction between elements in the non-logical lexical elements: the rules for handling the 'Hesperus' symbol are exactly the same as those for the 'Phosphorus' symbol.

The example Schneider gives of a Ramsey sentence suggests that she intends something else: that the Ramsey sentence range over the actual causal profile of a particular mind. The sentences conjoined in the Ramsey sentence would specify input/output relations between mental states. Here is her example:

> x has [⟨beer⟩] $=_{df}$ ∃P∃Q (input [{ }] causes P, and P causes both Q and output [{ }], and x is in P)
>
> Where P is the predicate variable replacing "has [⟨beer⟩]" in the theory, and Q refers to a different primitive symbolic state. (122, some orthographic substitutions made due to formatting constraints)

I think this example gets the idea across, even if the particulars are a little confusing (e.g., the *sentence* does not say which variable defines ⟨beer⟩, and it is unclear why 'has ⟨beer⟩' would be in the sentence and not just '⟨beer⟩'). Instead of taking the algorithms that describe the central system, we describe a particular mind's inferential dispositions for all possible inputs, and Ramsify over those. This would generate distinctions between lexical types. Different lexical types of symbols are likely to take part in different inferences. An instance of ⟨beer⟩ in one place would probably cause a different output (or mediating state) than ⟨water⟩ in the same place.

This, then, is the best candidate for Schneider. Unfortunately, as we shall see, it cannot work.

Processors cannot respond to roles

Processors need to recognize and distinguish lexical types. They need to be able to look at a symbol in an input string and compare it to items in memory, and in explicitly represented rules. Suppose I follow a rule that says to shush my cat if it is meowing and I already fed it. It might be represented like this:

> If input is ⟨cat meowing⟩, then if ⟨fed the cat⟩ is stored in memory, output ⟨shush cat⟩, otherwise output ⟨feed cat⟩.

To make this work, my mental processor needs to recognize that the ⟨cat⟩ symbol in the input is the same lexical type as the symbol in my memory that says ⟨fed the

cat⟩. If lexical types are roles, it is unclear how my mental processor could make this comparison. Being the same type of symbol is not a locally detectable feature, on this view. To recognize that two symbols are of the same lexical type requires stepping outside the system and seeing if they play the same role.

But even this is a perplexing idea. Symbols—that is, token representational vehicles—do not have roles. Roles describe patterns that hold of a symbol *type*. It is not as if one and the same symbol token occurs in all the places where x occurs in the Ramsey sentence (indeed, this would put us perilously close to Pessin's worry that there is no difference between the thoughts that *John loves Mary* and *Mary loves John*). Rather, the same type of symbol occurs in all those places. So, it wouldn't even be possible to take an external look at a system and see whether two symbols are the same type. You would need to be told what types of roles there are *and which role those token symbols played*. Let's call the symbol that occurs in my input 'cat_1' and the one stored in my memory 'cat_2'. If I want to figure out what my mental processor is going to do when it sees ⟨cat_1 meowing⟩ and ⟨fed the cat_2⟩, it does no good to know that there is a type of symbol such that if they were the same type they would cause the output ⟨shush cat⟩. I need to know that cat_1 and cat_2 *are* instances of that type of symbol.

This is, of course, moot because processors can't take an external view to figure out what role-type a symbol instantiates. If lexical types are role types, the processor needs some way of encoding type identity—some way of marking symbols to indicate what role they play. But then this mark would do all the work of formal features. It would be the feature that a processor uses to recognize and distinguish symbol types. Its rules would be defined over symbols typed by having that marker. So, while symbol types certainly *have* roles, these roles cannot be what defines them. You can't know what role a symbol plays unless you know what type of symbol it is.

So, Schneider's account fails to explain how a processor can recognize symbols as being tokens of the same type. It is worse, then, than any "detectable features" account—that is, any account which identifies symbol types with sets of processor detectable features. The physical properties hypothesis is one such view, claiming that all the detectable features are physical properties.

TCR is ungrounded

There is a deeper problem about the individuation of roles themselves. There can't be any facts about what roles exist if there is no role-independent criteria for symbol identity. If there is no fact about which symbol is which, then any way of carving up roles is equally good. You cannot construct a Ramsey sentence in a principled way without knowing which symbols the variable is supposed to replace.

Schneider (2011) is aware of a problem in this vicinity. She says that this is an *epistemological* puzzle about symbol individuation, stating the worry as follows:

> [T]o grasp which laws figure in the Ramsification that serves to type individuate the LOT primitives, we must borrow from an antecedent knowledge of symbol types, one that we do not have prior to their definition, for we must already know what the property types are to know which laws figure in the Ramsification. (122)

Her response is that epistemology should not drive our metaphysics. A role (or the property of playing a particular role) can exist even if there is no "epistemological route" into discovering that role. An epistemological worry like this one, about knowing the laws, does not tell us anything about the metaphysics of symbol individuation. So, since she is trying to offer just such a metaphysical theory, Schneider claims the worry does not apply (123-4).

But there is a metaphysical problem here. Suppose I have two mental rules:

> If input is ⟨cat meowing⟩, then if ⟨fed the cat⟩ is stored in memory, output ⟨shush cat⟩, otherwise output ⟨feed cat⟩, and if input is ⟨dog barking⟩, then if ⟨fed the dog⟩ is stored in memory, output ⟨let dog outside⟩, otherwise output ⟨feed dog⟩.

To distinguish between the orthographic properties used in text to distinguish symbols, and that role properties that Schneider claims individuate mental symbols, let's say that all instances of 'cat' are C-type symbols. We could define a role like this:

> C-type symbol $=_{df}$ the x such that input ⟨x meowing⟩ causes output ⟨shush x⟩ if ⟨fed the x⟩ is stored in memory, and output ⟨feed x⟩ otherwise, and if input is ⟨dog barking⟩, then if ⟨fed the dog⟩ is stored in memory, output ⟨let dog outside⟩, otherwise output ⟨feed dog⟩.

The problem, however, is that I simply stipulated that sentence describes a role. Without stipulation or some prior criteria for lexical individuation, nothing makes it true that *this* describes a role, but not a sentence that substituted x for some some 'cat' instances and some 'dog' instances, or any other haphazard substitution scheme. For example:

> C-type symbol $=_{df}$ the x such that input ⟨x meowing⟩ causes output ⟨shush cat⟩ if ⟨fed the cat⟩ is stored in memory, and output ⟨feed x⟩ otherwise, and if input is ⟨dog barking⟩, then if ⟨fed the x⟩ is stored in memory, output ⟨let x outside⟩, otherwise output ⟨feed dog⟩.

This has just as much claim on being a genuine role as the first Ramsey sentence. If there were some other principled way of individuating lexical types (like all 'cat' looking instances are the same lexical type), then we could point to a principled difference between the two. But this way out is unavailable to Schneider. On her view, what it is to *be* a lexical type is to play a particular role.

Unless role types are simply stipulated (as, for example, the rules of chess stipulate what types of pieces there are), the facts about which symbol roles exist are grounded by facts about what lexical types exist. And what grounds facts about what lexical types exist would seem to be facts about what properties symbol tokens have in common. But, if what symbol types have in common is just that they play a particular role, then we have a tight circle, and there seems to be nothing about the world that makes one type of role exist and not another.

TCR leads to false equivocations

Perhaps there is some other principled difference between roles that exist and those that don't that relies neither on fiat nor prior criteria for symbol type individuation. The onus is on the defender of role individuation to say what that would be. Even if one were to be found, however, role based views still fair worse than detectable features views. As I argued above, they cannot explain how processors recognize symbol types, nor why they must treat type identical symbol tokens as the same. They also provide counter-intuitive predictions about symbol identity—saying that two token mental symbols are the same type when, in fact, they are not.

Suppose there is a vending machine that works like this: when you push the button labeled 'Soda' it sends a signal to a computer, generating the input ⟨SODA⟩. This input corresponds to a certain pattern of electrical activity in the computer's memory register. The computer has a rule that says if it detects ⟨SODA⟩ input it should first check to see if there is any soda left in the soda slot (slot A), then, if there is, dispense a soda by releasing the door to slot A. The rule is written like this:

> **Rule 1:** ⟨If input=SODA, then if EMPTY(SLOT A)=false, DISPENSE(SLOT A), otherwise REFUND⟩.

The output ⟨DISPENSE(SLOT A)⟩ sends a signal to slot A, causing it to release the door. The result is that when you push 'Soda', you get a soda.

The machine also has a button labeled 'Tea'. Tea is in slot B, and the internal computer has the following rule:

> **Rule 2:** ⟨If input=TEA, then if EMPTY(SLOT B)=false, DISPENSE(SLOT B), otherwise REFUND⟩.

⟨SODA⟩ and ⟨TEA⟩ are different symbol types in this system. And Schneider's view captures this. The ⟨TEA⟩ input causes a different output than the ⟨SODA⟩ input. We can define ⟨SODA⟩ in a Ramsey sentence like this:

⟨SODA⟩$=_{df}$ $\imath$x such that:

1. If input=x, then output=⟨...SLOT A...⟩, and
2. If input=⟨TEA⟩, then output=⟨...SLOT B...⟩

So far, so good. But now imagine the machine is mis-programmed. The person writing the code for that particular vending machine makes a typo, inserting 'A' where ever there was supposed to be a 'B' in the rules. So instead of Rule 2, it has this:

Rule 2*: ⟨If input=TEA, then if EMPTY(SLOT A)=false, DISPENSE(SLOT A), otherwise REFUND⟩.

Now, if you push 'soda', you get a soda. But if you push 'tea', you also get a soda.

On Schneider's view, this programming error reduces the number of symbol types in the machine's language. Remember, the labels '⟨SODA⟩' and '⟨TEA⟩' are cheats—they stand for symbol types, which Schneider claims are roles. If she is right, then because there is only one role (causing the machine to check slot A, and dispense a soda from that slot), then there is only one symbol type.

The intuitive move here is to say that there are still two different types of symbols because there are two different patterns of electrical activity in the machine's memory register depending on which button is pushed. But, if we go Schneider's route, what reason do we have for saying that these patterns are different types of states *for the machine*? We are supposed to individuate symbolic states by their role, and the role of each pattern is the same: causing slot A to dispense. The appealing move tacitly individuates symbolic states by their physical properties.

We also cannot just bite the bullet, holding that machines following rules 1 and 2* have fewer symbol types than those following rules 1 and 2. Doing so causes us to miss important counterfactual generalizations about those machines: if their *rules* were different, the different electrical patterns in the register would generate different outputs.

A better view would say that the machine still has two different symbol types, but in the mis-programming case, it happens to treat them as the same. This is,

of course, just what the physical properties hypothesis says. The two symbols are physically distinct in a way the processor can detect.

Conclusion

The physical properties hypothesis provides a simple, yet powerful account of how mental symbols are individuated. Even without knowing which particular physical properties delineate symbol types, it gives us insight into how mental processors recognize lexical types. They match and distinguish physical/neural properties of symbols in their input, rules, and background knowledge. This suggests that the next step towards a more complete LOT is finding what sorts of physical properties these processors are sensitive to. And this, unlike the preceding, is a wholly empirical question. Which is to say, it is someone else's job.

CHAPTER 5

A DEFENSE OF HYPER-REPRESENTATIONALISM

Content attribution is ubiquitous in psychology. Vision scientists, for example, interpret states of the visual system as representing *edges*, developmental psychologists attribute representations of *causation* to infants, and cognitive ethologists attribute *distance* representations to certain species of ants. Yet many theorists are hesitant to accept content realism, the view that *having content* is an objective feature of the world. They hold that when we interpret a state as representing P (as having P content), we are not committed to there being a property of *having P content* that the state possesses independently of us and our theoretical goals. The aim of this chapter is to show that this skepticism is unwarranted.

One reason for doubting the reality of content is that while content attribution is undeniably useful, it is unclear exactly why. Perhaps, as Chomsky (1995) seems to suggest, it merely plays a convenient expositional role, with content standing in for some difficult-to-characterize feature of cognition. If so, content attribution would be an eliminable part of our scientific understanding of the mind, at least in principle. Or perhaps, as Egan (2014) argues, it plays the ineliminable role of showing how a mathematically characterized device amounts to performing an intentionally characterized cognitive capacity. This suggests that having content is not an objective feature of the world, but rather a product of our theorizing.

Another reason to doubt the realist ontology is that we lack an (uncontroversial) naturalistic theory of content determination. Egan (2014) claims that without a naturalistic relation that explains how R determinately represents P, any need we

have to attribute *P*-content to R must follow from something about us and our theoretical goals, not the objective facts about the phenomena. So, *having content*, on her view, is not an objective feature of the world.

Such doubts are misplaced. In what follows, I will argue that realism is not hostage to a theory of content. Mental content is a theoretical posit in a science that attempts to describe objective features of the world. We can, and should, be realists if we have good empirical reasons for attributing content. That is, realism requires a naturalistic basis for content attribution, not a naturalistic relation for content determination. Looking at the empirical literature, I argue further that we have such a basis, and that content attribution plays ineliminable roles that neither Chomsky (1995) nor Egan (2014) can account for. I conclude by defending a particular realist view that Egan (2014) terms "hyper-representationalism," which claims mental representations have their content *essentially*—that their content is part of what makes them the type of representation they are.

5.1 Realism and eliminativism

The debate over the ontological status of mental content is a debate over what properties *mental representations* must have to perform their explanatory role in psychology. Both what mental representations are and what role they play are matters of considerable controversy. As relatively untendentious characterization, mental representations are physically realized mental structures that can be interpreted as representing (referring to, standing for, or being about) some feature of the world. The physical state or symbol itself is what is called a *representational vehicle*. When we talk about mental *content*, we are talking about what these vehicles represent.

Because they are physical objects, representational vehicles are responsible for the causal (behavior producing) effects of mental representations. What, then, is content good for? *Realists* argue that content plays some unique, non-causal, and

theoretically important role in our explanations of behavior, and claim that this gives us good reason to believe that having content is an objective feature of the world. *Eliminativists* deny that content plays any theoretically important role, claiming that content attribution can, in principle, be replaced with something else.

Let's give some structure to this debate. Realism is typically defended along the following general lines:

1. Some successful explanations and predictions in psychology attribute content to mental representations. (Empirical premise)

2. Content attribution is an ineliminable part of these explanations/predictions. (Ineliminability premise)

3. If content attribution is an ineliminable part of successful explanations, then the mental representations they range over really have the content ascribed. (Realist premise)

Elminativists challenge realists to justify the ineliminability premise. Realists have obliged. Pylyshyn (1984), Block (1986), and Fodor (1995) argue that the laws of psychology are supposed to generalize over a potentially heterogeneous class of representational vehicles. If that were the case, then eliminating talk of content would leave us with a disjunctive class of states with no account of what they have in common.

In taking this line, realists are making an empirical bet that minds and mental states subsumed under the laws of psychology are often homogeneous only at the level of content. Fodor (1995), for example, writes that he can imagine a world in which:

> the laws that a computational psychology implements might be *intractably and ineliminably* intentional *because* they are laws about [...] a kind of content heterogeneous minds can share in virtue of similarities in their *extrinsic* relations. (53, emphasis original)

But, he continues, he has no *a priori* argument showing that this is actually the case. Our world may be one where psychological laws cover minds that are heterogeneous at the syntactic level and homogeneous at the content level, but "whether it is, is *strictly* an empirical issue" (53). This bet is backed by the claim that nothing in our current understanding of the mind requires homogeneity at other levels of explanation. If it turns out that minds are in fact homogeneous at some other level, it would be a surprise, and something of an accident since our current understanding is compatible with it not being the case.

For their part, eliminativists take the realists' bet. They argue that psychologists' talk of content does not express any deep theoretical commitment, but is a product of an incomplete understanding of the mind and a need for convenient ways of referring to complex states. Chomsky (1995), for example, argues that talk of content within science is merely a convenient, informal shorthand for an internal specification of a computational role. As an example, he considers Ullman's (1979) computational specification of the visual detection of motion. Though Ullman used a video of a rotating cube to test his model and described subjects as representing the cube, Chomsky contends that the theory itself was really about the internal relations of symbols. It describes a mathematical relation between states. What they were about, Chomsky claims, does not figure into the theory itself. So, he writes:

> There is no meaningful question about the "content" of the internal representations of a person seeing a cube under the conditions of the experiments, or if the retina is stimulated by a rotating cube, or by a video of a rotating cube... No notion like "content", or "representation of", figures within the theory... (1995, 52).

What content attribution buys us, according to Chomsky, is a convenient way of labeling mental states. On his view, psychologists use content in the way someone teaching logic might use semantic interpretations of logical symbols. When an instructor provides an interpretation of their logical vocabulary—$\forall x$ becoming 'everybody', $\exists x$ becoming 'somebody', and $\forall x \exists y Lxy$ becoming 'everybody loves somebody', for

example—this provides a quick, intuitive way to talk about the symbols. But the subject matter of the lesson is not everybody, somebody, or love—it is the rules governing the symbolic relations. The interpretation is not part of the theory. Likewise, on Chomsky's view, when Ullman talks about the representation of a cube's corner, it is only to give a convenient name to a symbol; what matters is the symbol's internal role in the visual system.

It is not enough for eliminativists to show that any particular explanation might do without content. They need to give some reason to think that every explanation could, at least in principle, eliminate talk of content. Chomsky justifies this by using a Twin Earth thought experiment in which a person, Jones, is replaced overnight by his twin, *J*—a person whose internal states are exactly like Jones', but have different contents due to *J's* living on a planet with distinct but perceptually indistinguishable elements. Chomsky claims that Earth's psychology will predict *J's* behavior just as well as it would have predicted Jones'. So, Chomsky claims, from the point of view of working psychologists "little seems at stake in these debates" over content (1995, 53).

5.2 Egan's anti-realism

Egan (2014) offers a distinct third alternative to both eliminativism and realism. Unlike the eliminativist, she accepts the realist's ineliminability premise. Our cognitive theories often aim at explaining some competence or capacity, she claims, and since these are described in intentional terms (the capacity to detect edges, for example) we need to attribute content to cognitive mechanisms to show how it relates to the phenomena being explained.

This is not a vindication of realism, she argues. Egan denies instead the realist premise, arguing that the ineliminability of content from explanations in psychology does not give us good reason to believe that mental representations *really* have the

content ascribed to them. She views content as a *product* of explanation. On her view, content is ineliminable in explanations of cognitive capacities, but there are no objective, evaluator-independent facts about what content a representation has.

Egan's (2014) position is that content attribution plays a heuristic role in psychology. We ascribe content to make it clear to us how the operations of some computational mechanism amount to performing a pre-theoretically characterized task. Content ascription, she says, "secures the connection" between these mechanisms and the capacity/competence that is the explanatory target (130). Since explanations are for our benefit—they tell us, the theorists, how something works—Egan views content ascription as a theoretical artifact, not to be understood as making a claim about properties the mechanism possesses independently of us and our theoretical aims. Had we characterized the capacity differently, or evoked the very same mechanism to explain a different capacity, we might ascribe different content. Absent our theoretical concerns, there simply is no fact of the matter about what content the states of some mechanism have. Nonetheless, because psychology is in the business of giving explanations of capacities, and explanations are, again, for our benefit, content ascription is ineliminable from our explanations.

We can think of Egan's (2014) views on what it is to have content along the lines of what Dretske (1993) calls having an "assigned function." He gives an example of a precise, highly sensitive scale. A theorist can place a small item on the scale and interpret its readout as giving the weight of the item. Or, since weight is a function of height above sea level, a theorist could put something on the scale and treat the scale as an altimeter. If the scale previously displayed '1g' at sea level and now, weighing the same object, displays '.98g', the theorist could take that as meaning they are at an altitude of 40,000ft above sea level (Dretske 1993, 301-2). What the display means is a product of the function we assign to it. Dretske writes:

> We get a new functional meaning because our altered background knowledge (normally a result of different intentions and purposes) changes

> what the pointer's behavior [means]. With *assigned* functions, the [meanings] change as our purposes change. (1993, 302)[1]

So it is, Egan (2014) claims, with mental representations in psychology. Theorists posit a mechanism to explain some cognitive capacity, and interpret its states in light of how this capacity is characterized. Should the theorists change their intentions or purpose they could, Egan (2014) contends, interpret the same mechanism differently, and thereby change its states' representational content.

5.3 What is content attribution good for, anyway?

To get clearer on Egan's conception of content's role in psychology, consider the well studied example of the desert ant *Cataglyphis*'s navigational capacities. *Cataglyphis* is able to navigate back to its nest in a straight line after walking hundreds of meters in a labyrinthine path to find food.[2] Unlike other species of ants, the desert ant cannot navigate by a pheromone trail. Instead, theorists argue that it navigates (in part) by using a system of *path integration*.[3] Path integration can be modeled in a number of ways.[4] Maurer and Seguinot (1995), for example, describe one method like this (summarizing Wehner and Wehner's (1990) model):

> Let φ_n be the (recursively computed) compass direction from the starting point to the animal; let l_n be the (recursively computed) beeline distance between the starting point and the animal; let δ be the angle by which the $(n+1)$th step deviates from φ_n, and hence $\varphi_n + \delta$ the direction,

---

[1]For ease of exposition I have omitted Dretske's notation distinguishing what he calls *functional meaning* from *natural meaning*.

[2]This discussion of desert ant navigation draws from Wehner and Srinivasan (1981), Wehner and Wehner (1990), and Burge (2010).

[3]Path integration is also known as *dead reckoning* or *vector navigation*. In addition to path integration, *Cataglyphis* also makes use of *visual piloting*, in which the ant uses a landmark map to guide it through familiar territory, and *systematic search*, in which the ant uses an algorithm that generates a spiral-based pattern for exploring new territory and returning to its search origin.

[4]See Maurer and Seguinot (1995) for an overview of various models, and a discussion of their purposes, advantages, and disadvantages.

> as computed by the animal, in which it proceeds for another step (angles are expressed in degrees; every step is of arbitrary length 1). Then φ_{n+1} and l_{n+1} denote the direction and distance, respectively, after the $(n+1)$-th step:
>
> $\varphi_{n+1} = \frac{l_n + \varphi_n + \delta}{l_n + 1} = \frac{\varphi_n \cdot (l_n + 1) + \delta}{l_n + 1}$
>
> $l_{n+1} = l_n + 1 - \frac{\delta}{90^\circ}$

The computational *model* consists in the mathematical equations at the bottom of the passage. The *explanation* of how the ant finds its way home also includes Maurer and Seguinot's (1995) interpretation of the symbols as angles, directions, vectors, and steps. Egan (2014) argues that this interpretation is purely for our benefit. Without it, we would be unable to see how the model amounts to the capacity to find a straight vector home. So, it is an ineliminable part of the explanation of the capacity. But, she claims, this does not show that the ant *really* represents these things—that, as a matter of objective, evaluator-neutral fact, its states have these contents. What is objectively true is that the ant's navigational system computes the mathematically described function.

There is, I think, much to be said in favor of Egan's analysis of content's role in explanations like the above. Here it does seem that content is playing a heuristic purpose—making the workings easier for the researchers to talk about, and showing how a specific mathematical formula relates to a cognitive task.

As a general model of content's role in psychology, however, it is far too limited. Psychology is not only in the business of explaining how some or other mathematical function could amount to performing some or other capacity. It tries to predict and explain behavior more generally. Content ascription is useful for this purpose, even in the absence of a worked out computational/mathematical model.

Consider again the example from Diesendruck and Peretz (2013), that children's categorization behavior is governed by the following rule:

(15) If X is an artifact, and X's creator intended X to be an F, then, *ceteris paribus*, categorize X as an F.

This implicates a representation with the content *artifact*. When children are in a state that predicates *being an artifact* on some object, their behavior will be as Diesendruck and Peretz (2013) describe. (Children did not follow an analogous rule when categorizing non-artifacts like animals.)

Here, content ascription does not seem to play the role of showing how a non-intentional mechanism amounts to performing a pre-theoretically characterized task. The characterization of the mechanism itself is intentional—there is no explicit computational model that needs to be interpreted. So, content ascription is not *only* useful for "securing the connection" between a computational model and a pre-theoretically characterized capacity.

I want to press a more general point: Content attribution allows us to exploit our understanding of P—some property, object, or whatever in the world—to make specific predictions about the behavior of a system that represents P, without needing to know any details about how that system works. That is, we can infer from properties and relations of the thing represented to behavioral effects of the representing states.

This is common practice in folk psychology. Because we know what it means to be poisonous, for example, we can predict that if Schobert represents the mushrooms as *being poisonous* he will not eat them. The fact that he does eat the mushrooms we take as evidence that he did not represent them as poisonous.

This sort of prediction is also a common practice in scientific psychology. We use what we know about P to make specific predictions about how a system will behave if it represents P. Take for example Leslie and Keeble's (1987) landmark study of infants' perception of causal events. They predicted a behavioral difference in how long six month olds would look at a certain sequence of events based on whether they represent it as *causal* rather than as *non-causal sequences*. The setup worked like this: One group of infants was habituated to a display where a moving object,

A, contacted a stationary object, B, which then moved.[5] They were then shown the same display, played in reverse (B moves, hits A, then A moves). The other group of infants was habituated to a display where A moves, stops short of contacting B, then B moves. After habituation, they were shown that display in reverse.

Leslie and Keeble (1987) predicted that infants in the first group would dishabituate more strongly to the reversed scenario (i.e., look longer at it) than those in the second group. Why should this be? After all, just as much has visually changed in each display. They reasoned as follows: Though they may be identical in appearance, there is an ontological difference between causal events and non-causal events. Causal events have distinct roles for the *agent of causation* and *patient of causation*. When you reverse a causal event, there is an additional change in role (B changes from patient to agent, *vice versa* for A). Since no such roles exist in non-causal events, then there are more changes in causal reversal scenarios than in non-causal reversals. So, Leslie and Keeble (1987) hypothesized that if infants represented the contact scenario as *causation* but not the non-contact scenario, they would perceive contact-reversal as more novel than non-contact reversal. And this is exactly what they found: infants dishabituated more strongly to the contact-reversal scenario than the non-contact scenario.

The moral is that researchers were able to make a prediction about behavior based only on ascribing a particular representational content to infants, and knowing what that content means. Content's role here is to license the move from a thing in the world having certain non-observable properties to a prediction about how mental representations will affect behavior. The realist thus has an advantage over the anti-

[5]Looking time studies are based on the idea that infants look longer at things that are novel. They are first *habituated* to a stimulus. That is, they are repeatedly shown some target scenario until they grow bored of seeing it—where they spend no more time looking at the display than anything else in their environment. Then, they are presented with one of several different test scenarios. Researchers measure how long they look at the various new scenarios, and compare the degree to which they dishabituate. The longer they look, the more novel they perceive it.

realist: they can explain why predictions like Leslie and Keeble's (1987) work, and the anti-realist cannot. Unless these ascriptions are picking out some real (objective, evaluator-independent) feature of the representations, it is difficult to see why the predictions should succeed. The predicted behavior is intimately connected with the nature of the content ascribed. And this match does not seem to be in any way a result of our theoretical goals and aims. Indeed, there was no pre-theoretically characterized capacity being explained. The researchers made a prediction of what behavior the infants would exhibit based on the meaning of hypothetical content.

One possible anti-realist reply is that if we knew how the computational mechanisms underpinning causal perception worked, we could make the same behavioral predictions without needing to appeal to content. If we had such a mechanism, the thought goes, content would be reduced to playing the heuristic role that Egan (2014) envisions. This is an unsatisfying reply for two reasons. First, it leaves it mysterious why these sorts of predictions should be so successful in the absence of a worked out computational model. If infants don't really represent causation, it's something of an accident that ascribing *causation* representations to them should make the very subtle and precise predictions that it does. Second, it does not actually matter to Leslie and Keeble's (1987) theory how exactly the computational mechanism works. Any system that represents causation *as such* is likely to exhibit a sensitivity to agent/patient reversals because of the nature of causation. So, their prediction could in principle subsume many different types of computational mechanisms. Infants, adults, camels, and Martians may all deploy different computational mechanisms, but if they represent causal events *as such*, we would predict that they exhibit different sensitivities to causal and non-causal reversals. And if they do not, that would count as evidence that they do not.

5.4 Does realism require a naturalized semantics?

Realists are typically committed to some version of semantic naturalism. They claim that mental states have determinate content, and that this content is determined by some naturalistic relation. Millikan (2002) and Neander (2013), for example, advance theories that claim, roughly, that a representation's content is determined by its biological function. Fodor (1992a) and Rupert (1999) offer theories claiming that having some content, *P*, is a matter of standing in the right sort of causal relation to P. Egan (2014) aims to erode support for realism, and thereby motivate anti-realism as an alternative, by casting doubt on the plausibility of these types of naturalization projects. If no naturalistic property or relation yields determinate content, she reasons, then eliminativism or anti-realism are our only options. Since content is useful and ineliminable, then it looks like we ought to be anti-realists.

Her argument, in broad strokes, is that there are certain general problems about determinacy facing any naturalized theory of content, and no one has yet proposed a naturalistic relation that uncontroversially overcomes them, despite being aware of these problems for decades. So, she reasons, it is unlikely that there is such a naturalistic relation to be found. This is what Neander (2015) calls an "argument from despair."

The determinacy challenges that Egan (2014) has in mind take this general form: Inevitably, any naturalistic relation that some state, S, bears to some candidate content *P*, it will also bear to some other non-Ps. The details depend on the type of theory. Causal theories of mental content, for example, hold that having the content *P* is a matter of being causally connected to P in the right sort of way. But there are some *non*-P properties such that P is instantiated *iff* those things are instantiated too. This may be accidental—the property of *being a fly* may be instantiated *iff* the property *ambient black nuisances* is instantiated, if flies are the only ambient black nuisances that exist. Some properties, however, are necessarily co-instantiated—*being*

a fly, *being an undetached proper fly part*, and *being a temporal stage of a fly*, to take Quine's (1960) examples. The problem, then, is that if S is causally connected to *being a fly*, it is inevitably causally connected to these other properties as well, so it is difficult (to understate it) to see how S can determinately mean *fly*. (Similar worries can be raised for teleological theories of content.)

One could respond to the argument from despair by defending a particular naturalistic semantic theory, and demanding that Egan debate it on its particular merits. Or, one could say that it is too soon to despair, and the fact that we have made *some* progress warrants a more hopeful outlook.

Neither response, however, is necessary. Egan's (2014) argument from despair does not give us any reason to take the anti-realist stance. Anti-realism, recall, claims that we need to attribute determinate content to representational states, and that our theoretical goals determine what this content is. Given her reasons for despair over finding a naturalized semantics, she cannot justify both claims.

First consider why a theorist would be attributing determinate content. Either they have a good empirical basis for attributing determinate content, or they do not. If they do, then it's not their theoretical goals that determine what content they need to attribute, but the empirical facts. Consider again Leslie and Keeble's (1987) work on infants' causal perception. Their reasons for ascribing determinate *causation* content to infants was that the infants behaved in a way we would expect only if they represented certain situations as *causation*. Their attribution was based on objective facts: how long infants looked at different types of stimuli. So, it seems that facts about infants' behavior are best explained by attributing *causation* content. If we need content ascription to explain these sorts of facts—objective (evaluator independent), empirically based facts—then we have reason to believe that the objective world instantiates these properties. So, if content attribution is based on facts like these,

then the anti-realists' second claim is unmotivated. The world, not our theoretical goals, determines what content we ascribe to a representational system.

Egan's worries about determinacy give us some reason to believe we cannot have this sort of basis. She claims that there are always alternative candidate contents that are equally compatible with all of the observable regularities concerning a system. If the system exhibited some behavioral regularity that privileged, say, a *fly* attribution over a *temporal fly stage* attribution, we would be in the above scenario. But since flies and temporal fly stages are empirically inseparable (you can't have one without the other), it seems we could not find any such regularity. So, if we attribute determinate *fly* content to the system, this can only be because we made some choice based on extra-empirical factors, like "ease of explanation" (Egan's (2014) example, p. 125).

Now, however, the question is why we *need* to attribute determinate content. Why would explanations in psychology need to ascribe content that is determinate beyond the point of empirical decidability? They might do it purely for convenience—it would be a pain to list all of the Quinean competitors for some representation's content. But this would not justify anti-realism. It would justify a limited form of eliminativism. Determinate content would be, in principle, eliminable from explanations. It could be replaced by a (very long) list of competing content. Determinate content ascription would play no role in psychology beyond providing convenient shorthand for such a list. (This falls short of full-fledged eliminativism because *indeterminate* content would still play some theoretically useful purpose.)

Egan's (2014) move to say content is required to "secure the connection" between a mechanism and a pre-theoretically characterized capacity only pushes this problem back a step. Suppose a capacity is characterized in a way that requires us to distinguish among these empirically indiscernible competitors. In order to see how a computational system amounts to, say, the cognitive capacity to visually detect prey, we may need to describe it as representing *flies* instead of *undetached fly parts*. But

now we want to know what our our basis is for characterizing the *capacity* in this way. If it is objective, empirically observable facts about the system, then the determinate content attribution is not theory driven. The characterization of the capacity is a product of the world, not our theoretical goals, and the need to attribute determinate content results from that characterization. If, on the other hand, our characterization of the capacity is based on some extra-empirical facts, then explaining that capacity, so characterized, is not a theoretically sound goal. This characterization is best viewed as convenient shorthand for something indeterminate. So, again, a qualified eliminativism is what is warranted, not anti-realism.

This may be the appropriate analysis of the desert ant's navigational system. We have no good reason to think that it represents *steps taken* instead of, say, *distance walked*. And for a computational theory like Wehner and Wehner's (1990) to work, we do not need to treat their determinate interpretation as anything more than convenient shorthand for a number of empirically indiscernible candidates. Of course, if we learn more about how the ant works, and we come to have a reason to discriminate between these alternatives, we would then start taking these attributions as substantive. But it would be the objective facts about the ant, not the theorist's goals, that demand determinacy.

The anti-realist is thus presented with a dilemma: For any explanation that attributes determinate content to a system, either there is a good naturalistic basis for this attribution or there is not. If there is, realism is justified. If not then psychology, *qua* empirically driven enterprise, does not need to attribute determinate representations, and so does not need the anti-realist's analysis of how content attributions get their determinacy.

The error in Egan's (2014) argument was in thinking that realism requires a naturalistic account of content determination. It does not. What it requires is that there be a purely *naturalistic basis* for determinate content *attribution*. Generally

speaking, we can know that something has a certain property without knowing how it managed to get that property. If our reasons for attributing that property are based on objective, empirically backed facts, and if attributing that property conforms to good scientific practice—if, for example, it is not *ad hoc*—then we have reason to believe that this property exists in the world, independently of our theorizing about it.

My proposal, then, is simply to treat determinate attributions of content as normal sorts of empirical hypotheses. Theorists predict that a system will behave a certain way if it has some or other determinate content, then observe how it behaves, and confirm that it has this content. These hypotheses are, of course, susceptible to empirical uncertainty. It is up to psychologists to do the work of figuring out which content applies by considering and rejecting plausible rival hypotheses.

Realism about mental content is not exceptional in scientific discourse. While it may be true that some hypotheses—e.g., those featuring Quinean competitors—may never be evicted, the inability to dismiss empirically indiscernible rival hypotheses is hardly a problem unique to psychology. The reasons to resist realism here are just the same reasons to resist scientific realism generally.

If someone could show that it was impossible for representations to get content in a naturalistically respectable way, then all bets are off. But no one has, of course, shown any such thing. And the fact that at least some determinate attributions of content have a good naturalistic basis give us very good reason to doubt that anyone could.

5.5 Hyper-representationalism

I want to turn now from realism to a stronger view: hyper-representationalism. "Hyper-representationalism" is Egan's (2014) name for a collection of three interrelated claims, which she states as follows:

> (1) that mental representations have their contents essentially, (2) that misrepresentation is possible, and (3) that such content is determined by a privileged naturalistic property or relation. (117-8)

I am not going to argue for (2) and (3). These have been extensively discussed in the literature.[6] Moreover, Egan (2014) does not deny (2) (on her view misrepresentation is possible, but only relative to an explanatory project), and her argument against (3), as discussed above, is little more than despair.

My focus, then, is on (1), content essentialism. This is a claim about the type identity of mental representations within cognitive psychology. It says that psychology's explanations range over representations that are characterized in part by their content, such that representations with distinct content are *ipso facto* different types of representations.

Representations, on this view, are something like words in a language. Words are n-tuples of vehicles (various orthographic and phonetic types) and content. Both types of elements—vehicle and content—are part of what makes a word the type of word it is. So, <peculate> and <embezzle> are different words, despite having identical content.[7] And <sink$_1$> (meaning a small artificial basin) and <sink$_2$> (meaning to become submerged) are distinct words despite sharing a representational vehicle, 'sink'. Likewise, hyper-representationalism claims that representations with different content are *ipso facto* distinct representational types. A mental symbol with the content *cow*, on this view, is a distinct mental representation from one with the content *horse*, despite any other similarities between them.

The reasons for being a hyper-representationalist are related to the reasons for being a realist. Some of the laws and generalizations of cognitive science are couched in intentional terms. Realists claim that the regularities such laws/generalizations

---

[6] See Adams and Aizawa (2010) and Neander (2012) for overviews.

[7] I am using '<>' to indicate *words*, as opposed to the orthographic types typically indicated by single quotes.

capture only exist at the level of content. Pylyshyn (1984) offers the following example: we can predict and explain people's behavior in light of their believing *that they are in the presence of an emergency.* People who believe they are in the presence of an emergency will try to get help—a generalization that ranges over people who represent their situation *as an emergency* and their actions *as attempts to get help.* Nothing besides being believed to be an emergency unifies, say, a car crash, a child being trapped down a well, or a knife wielding person lurking outside the window. And nothing besides being believed to be an attempt to get help plausibly unifies the actions of, say, flagging down a police officer, dialing 911 on a cell phone, dialing 911 from a pay phone, shouting 'help!', etc. So, this generalization provides a useful way of saying what otherwise distinct states have in common that explains their participating in a certain sort of regularity. And, since no other property can stand in for *believing that X is an emergency* or *believing that Y is an attempt to get help*, this property is ineliminable (Pylyshyn, 1984, p.5–12).

Where realism claims that certain regularities are only expressible in intentional terms, content essentialism says that those regularities are only expressible in *those* intentional terms. If content essentialism is true, then talking about representations with P content picks out *a different class* of mental representations than those with Q ($\neq P$) content. So, if we have good reason to state a law or generalization in determinate intentional terms, then any law/generalization subsuming representations under *different* intentional terms is *ipso facto* capturing a different regularity because it subsumes a different set of representations.

Hyper-representationalism claims that cognitive science is committed to an ontology of representations typed by content. There are generalizations that hold of all and only P representers, and different generalizations will hold of Q ($\neq$ P) representers. By attributing the determinate content P to a system, you thereby lump it in with

all the other P representers, and thus become committed to all the laws that are true of things that represent P as such.

As evidence for this, we would look to see if theorists treat generalizations cast in terms of different content as different generalizations. This is, I suspect, taken for granted in folk psychology. The fact that so-and-so believes that *there is an emergency* licenses a different swath of predictions about their behavior than, say, their believing that *it looks like there is an emergency*, or *there is a traffic accident*, or *there was a loud noise*, or whatever.

The same is true, I contend, of cognitive science proper. We have already seen suggestive examples from the developmental literature. Leslie and Keeble (1987) make different predictions about looking time behavior in infants based on the putative content of their perceptions. Had the content been anything but *causation*, they would not have made the same prediction because the predicted behavior was tied up with facts about causation *as such*. Causation representers, they reasoned, will obey certain rules not true of non-causal sequence representers.

Diesendruck and Peretz (2013) explain differences in children's categorization behavior depending on whether the child represents an object as an *artifact* (versus a natural kind). It would be a difference in theoretical substance to explain their behavior in terms of their representing objects as, say, *human-made items*. Possibly not all artifacts are human-made items, so we can imagine something that responds selectively to only the latter. By saying that infants represent objects as *artifacts* we are grouping them with artifact representers, and not human-made items representers.

This is not conclusive evidence by any means. I offer only *prima facie* evidence for content essentialism: Psychologists use content attribution to buy into certain general predictions based on the properties of what is being represented. They implicitly assume that there are laws/generalizations that apply to P representers as such.

5.6 Egan's anti-essentialist argument

I want to turn now to Egan's (2014) argument against content essentialism. Her argument goes like this: Mental representations are states of computationally characterized devices. Theorists may assign different content to states of computationally characterized devices, depending on things like its normal environment, how it is embedded in an organism, and various pragmatic considerations. That is, were these factors to change, the device would retain its identity *qua* computational device, yet its content would be different. So, states of computationally characterized devices do not have their content essentially. So, mental representations do not have their content essentially.

Because Egan's (2014) argument is brief, I want to quote it at length:

> *The structures posited by the computational theory, what we are calling the 'representational vehicles', do not have their cognitive contents essentially.* If the mechanism characterized in mathematical terms by the theory were embedded differently in the organism, perhaps allowing it to sub-serve a different cognitive capacity, then the structures would be assigned *different* cognitive contents. If the subject's normal environment were different (as, for example, in an Ames room), so that the use of these structures by the device in this environment did not facilitate the execution of the specified cognitive task, then the structures might be assigned no cognitive contents at all. (127, emphasis original)

One thing to note is the slide to *representational vehicles*. Hyper-representationalism, as Egan herself characterizes it, says that mental representations, not representational vehicles, have their content essentially. If psychology treats representations as n-tuples of vehicles and content, then showing that types of vehicles can realize different types of content is no mark against the view. To equivocate between representation and vehicle begs the question.

I take it, however, that her point is this: from the perspective of computational cognitive science, there is no difference between, say, *edge* and *shadow* representations if they are realized by the same type of computational state. Every behavioral

regularity that holds of one type holds of the "other" type since the computational features, not the content, govern the behavior of a device.

Egan (2014) has not, however, shown that states of the same type of device—in the sense that interests psychologists—*can* be attributed different interpretations. We might think that psychology is interested in devices that are partly characterized by factors like how they are embedded in an organism, and what their normal environment is. Suppose we have reason to attribute *edge* to human infants' visual systems. The fact that a Martian cat uses the same computational process to audibly detect *fog horns* would not itself give us any reason to hedge our original attribution. Unless human infants happen to be sensitive to fog horns in the same way, we can't make any useful predictions about their behavior in those terms. And, if that were the case, we would want to distinguish humans' visual representations from Martian cats' auditory ones, meaning that these factors are relevant to how we characterize representations after all.

For Egan's examples to tell against content essentialism, she would need evidence that there are devices sharing these sorts of features—being embedded in a certain sort of organism in a certain sort of way, e.g.—that are attributed different contents in different explanatory contexts. And she has not offered any.

If you want to know what properties are essential to some higher-level type, you need to look at what sorts of property changes or differences would cause something of that type to be categorized differently. If you want to know what properties are essential to being a cow within zoology, figure out what you would have to change about a cow to make zoologists call it something else (give it different parents? different genes?).

What Egan is doing instead is showing that the realizers of a higher level type can also realize other types of things. But the fact that B realizes P in one context, and

Q in another does not mean that P and Q are not real, or not genuine explanatory categories.

In neurochemistry, for example, the organic chemical *acetylcholine* is used as a neuromodulator in the brain—altering the way brain structures process information. At neuromuscular junctions, where nerves meet muscles, acetylcholine is used as a neurotransmitter. Neuromodulators and neurotransmitters are distinct higher level types. Different generalizations and laws subsume them. So, a law about how neuromodulators work would subsume some, but not all instances of acetylcholine. The fact that the exact same chemical can realize one type when it is embedded one way in an organism and the other when it is embedded in a different way does not tell against the reality of those types. They earn their keep by being explanatorily useful and irreducible to other types. And whatever properties are individuative of them as a type, they have essentially, by definition.

The same can be said for the relationship between mental representations and computational/mathematical mechanisms. *Being a mental representation* is a higher level property. If it is useful to type mental representations by the content they have, and we subsume representations (or representing systems) under different generalizations based on differences in content, then the fact that the same type of computational/mathematical mechanism can realize different contents is irrelevant to thinking about how the mental representations they realize should be individuated.

A useful distinction here is between what Shoemaker (1981, 2007) calls *core realizers* and *total realizers*, which he draws as follows:

> A total realizer of a property will be a property whose instantiation is sufficient for the instantiation of that property. A core realizer will be a property whose instantiation is a salient part of a total instantiation of it. (2007, 21)

Being acetylcholine is not a total realizer of *being a neuromodulator* or of *being a neurotransmitter*. The fact that these are different types of properties, and that

acetylcholine can be used as either, is evidence of this. It is, however, a core realizer of each. *Being acetylcholine*, together with certain other properties (e.g., *being used at a neuromuscular junction*) realize the property *being a neurotransmitter*.

What Egan (2014) has in effect shown is that *being a certain type of computational mechanism* is a core realizer, but not a total realizer of *being a mental representation of a certain type*. It is part of a total realizer that may include properties like *being embedded in this organism like so* and *being related to the environment like such*.

This is a point that hyper-representationalism can readily take on board. If representational types are n-tuples of vehicles and contents, and content is partly determined by things like how a representation is related to the environment, then knowing only that a representation is realized by some particular mathematical/computational mechanism won't tell us what type of representation it is. So, lacking this information, we will only know the computational/mathematical generalizations, missing the intentional generalizations. The reason we type representations by content, as I argued above, is to buy into these generalizations. So, on the hyper-representationalist view, *being a certain type of mechanism* is not sufficient for realizing the property of *being a certain type of representation*. So, showing that the same mechanism can realize different contents does not show us that mental representations do not have their content essentially.

Conclusion

In reading the literature on content realism, one can't help but come away with the impression that to theorists of a certain mindset, mental content is like underwear: useful, but not something to talk about in good company. My hope is that this chapter goes some way towards fixing content's image problem. There is nothing metaphysically disreputable about attributing determinate content to mental states when the evidence warrants it, and there is no need to explain away its appearance. In

defending realism and essentialism, I am trying to suggest nothing more than taking cognitive psychology seriously as a science. It offers empirically driven explanations of objective features of the world. So, its explanatory vocabulary, if useful and ineliminable for predictions and explanations, should also be seen as picking out objective features of the world.

There is, however, still reason to be cautious. I have suggested that we should take determinate content attribution seriously if, and only if, there is good empirical reason for doing so. That is not a trivial matter. Figuring out how to test hypotheses that offer competing content attributions requires serious empirical work. But neither is this impossible. When there are differences between two competing contents, it may be possible to find ways of testing sensitivity to these differences, providing evidence about what is represented. Saying what content a representation represents is thus a matter for empirical psychology. Saying how this could possibly be so is still a matter for philosophy, but not one that needs to be resolved before psychology can go about its business.

BIBLIOGRAPHY

Adams, F. and Aizawa, K. (2010). Causal theories of mental content. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Aizawa, K. (2003). *The systematicity arguments*, volume 83. Springer.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and brain sciences*, 22(04):637–660.

Bickford, J. and Daly, J. (1996). *A course in basic grammatical analysis*. Summer Institute of Linguistics, Dallas, TX, 3rd preliminary edition.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, 10:615–678.

Bodén, M. and Niklasson, L. (2000). Semantic systematicity and context in connectionist networks. *Connection Science*, 12(2):111–142.

Burge, T. (2010). *Origins of Objectivity*. Oxford University Press, Oxford.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press, New York.

Chomsky, N. (1995). Language and nature. *Mind*, 104(413):1–61.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam, New York.

Diesendruck, G. and Peretz, S. (2013). Domain differences in the weights of perceptual and conceptual information in children's categorization. *Developmental Psychology*, 49(12):2383–95.

Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press, Cambridge, MA.

Dretske, F. (1993). Misrepresentation. In Goldman, A., editor, *Readings in philosophy and cognitive science*, chapter 14, pages 298–314. Oxford University Press.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, 170(1):115–135.

Fischer, J., Spotswood, N., and Whitney, D. (2011). The emergence of perceived position in the visual system. *Journal of Cognitive Neuroscience*, 23(1):119–136.

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

Fodor, J. A. (1987). *Psychosemantics*. MIT Press, Cabridge, MA.

Fodor, J. A. (1992a). *A Theory of Content and Other Essays*. MIT Press, Cambridge, MA.

Fodor, J. A. (1992b). A theory of content II. In *A Theory of Content and Other Essays*, pages 89–136. MIT Press.

Fodor, J. A. (1995). *The Elm and the Expert*. MIT Press, Cambridge, MA.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong*. MIT Press, Cambridge, MA.

Fodor, J. A. (2007). *LOT 2*. Oxford University Press, Oxford.

Fodor, J. A. and Lepore, E. (1992). *Holism: A Shopper's Guide*. Blackwell, Oxford.

Fodor, J. A. and Lepore, E. (2002). *The Compositionality Papers*. Oxford University Press, Oxford.

Fodor, J. A. and McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why smolensky's solution doesn't work. *Cognition*, 35(2):183–204.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71.

Fodor, J. A. and Pylyshyn, Z. W. (2014). *Minds without meanings: An essay on the content of concepts*. MIT Press, Cambridge, MA.

Frank, S. L., Haselager, W. F., and van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition*, 110(3):358 – 379.

Frege, G. (1892). On sense and reference. In Geach, P. and Black, M., editors, *Translations from the Philosophical Writings of Gottlob Frege*, pages 56–78. Blackwell.

Ganea, P. A., Allen, M. L., Butler, L., Carey, S., and DeLoache, J. S. (2009). Toddlers referential understanding of pictures. *Journal of Experimental Child Psychology*, 104(3):283 – 295.

Gauker, C. (2011). *Words and Images: An Essay on the Origin of Ideas*. Oxford University Press, Oxford.

Gelman, S. A. and Kalish, C. W. (2006). Conceptual development. In Kuhn, D. and Siegler, R., editors, *Handbook of Child Psychology*. John Wiley & Sons, Hoboken, NJ.

Goodman, N. (1976). *Languages of Art: An Approach to a Theory of Symbols*. Hackett Publishing Company, Indianapolis, 2nd edition.

Hadley, R. F. (1997). Cognition, systematicity and nomic necessity. *Mind & language*, 12(2):137–153.

Hadley, R. F. (2004). On the proper treatment of semantic systematicity. *Minds and Machines*, 14(2):145–172.

Huang, J. Y. and Bargh, J. A. (2014). The selfish goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences*, 37(02):121–135.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. In *Econometrica*. Citeseer.

Keil, F. C. (1989). *Concepts, Kinds and Development*. Mit Press.

Kripke, S. (1977). Speaker's reference and semantic reference. *Midwest Studies in Philosophy*, 2(1):255–276.

Kripke, S. (1979). A puzzle about belief. In Martinich, A. P., editor, *The Philosophy of Language: Fifth Edition*, pages 433–459. Oxford University Press, Oxford. Reprinted in collection in 2008.

Kripke, S. A. (1980). *Naming and Necessity*. Harvard University Press, Cambridge, MA.

Laurence, S. and Margolis, E. (1999). Concepts and cognitive science. *Concepts: Core Readings*, pages 3–81.

Leslie, A. M. and Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3):265–288.

Lewis, D. (1973). *Counterfactuals*. Basil Blackwell, Oxford.

Machery, E. (2006). Two dogmas of neo-empiricism. *Philosophy Compass*, 1(4):398–412.

Machery, E. (2007). Concept empiricism: A methodological critique. *Cognition*, 104(1):19–46.

MacNell, L., Driscoll, A., and Hunt, A. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4):291–303.

Maurer, R. and Seguinot, V. (1995). What is modelling for? a critical review of the models of path integration. *Journal of Theoretical Biology*, 175:457–475.

McLaughlin, B. P. (1993). Systematicity, conceptual truth, and evolution. *Royal Institute of Philosophy Supplement*, 34:217–234.

Medin, D. L., Altom, M. W., Edelson, S. M., and Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(1):37.

Millikan, R. G. (2002). Biofunctions: Two paradigms. In Ariew, A., editor, *Functions*, pages 113–143. Oxford University Press.

Neander, K. (2012). Teleological theories of mental content. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Neander, K. (2013). Toward an informational teleosemantics. In Ryder, D., Kingsbury, J., and Williford, K., editors, *Millikan and Her Critics*. Wiley-Blackwell.

Neander, K. (2015). Why I'm not a content pragmatist [online conference paper]. *Minds Online Conference*. Retrieved from http://mindsonline.philosophyofbrains.com/2015/session4/why-im-not-a-content-pragmatist/.

Pessin, A. (1995). Mentalese syntax: Between a rock and two hard places. *Philosophical Studies*, 78(1):33–53.

Preissler, M. A. and Bloom, P. (2007). Two-year-olds appreciate the dual nature of pictures. *Psychological Science*, 18(1):1–2.

Prinz, J. J. (2002). *Furnishing the Mind: Concepts and their Perceptual Basis.* MIT Press, Cambridge, MA.

Prinz, J. J. (2005). The return of concept empiricism. In Cohen, H. and Lefebvre, C., editors, *Handbook of categorization in cognitive science*, pages 679–695. Elsevier, New York.

Putnam, H. (1975). *Mind, Language, and Reality*. Cambridge University Press, Cambridge, UK.

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge Univ Press.

Quine, W. V. (1951). Two dogmas of empiricism. *The Philosophical Review*, 60(1):20–43.

Quine, W. V. (1960). *Word and Object*. The MIT Press, Cambridge, MA.

Rastle, K. (2007). Visual word recognition. In Gaskell, M. G., editor, *The Oxford Handbook of Psycholinguistics*, chapter 5, pages 71–87. Oxford University Press, Oxford.

Rey, G. (1983). Concepts and stereotypes. *Cognition*, 15(1):237–262.

Rey, G. (1993). The unavailability of what we mean. *Grazer Philosophische Studien*, 46:61–101.

Rey, G. (1995). A not "merely empirical" argument for a language of thought. *Philosophical Perspectives*, pages 201–222.

Rosch, E. (1978). Principles of categorization. In Margolis, E. and Laurence, S., editors, *Concepts: Core Readings*, pages 189–206. MIT Press, Cambridge, MA.

Rupert, R. D. (1999). The best test theory of extension: First principles. *Mind and Language*, 14(3):321–355.

Schneider, S. (2009). LOT, CTM, and the elephant in the room. *Synthese*, 170(2):235–250.

Schneider, S. (2011). *The Language of Thought: A New Philosophical Direction*. MIT Press, Cambridge, MA.

Shoemaker, S. (1981). Some varieties of functionalism. *Philosophical Topics*, 12(1):93–119.

Shoemaker, S. (2007). *Physical Realization*. Oxford University Press.

Smolensky, P. (1988). The proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:174.

Stich, S. P. (1983). *From folk psychology to cognitive science: The case against belief*. the MIT press.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):pp. 453–458.

Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London B: Biological Sciences*, 203(1153):405–426.

Wehner, R. and Srinivasan, M. V. (1981). Searching behaviour of desert ants, genus cataglyphis (formicidae, hymenoptera). *Journal of comparative physiology*, 142:315–338.

Wehner, R. and Wehner, S. (1990). Insect navigation: use of maps or ariadne's thread? *Ethology Ecology & Evolution*, 2(1):27–48.