Mental Representation and Closely Conflated Topics

Angela Mendelovici

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Philosophy
Primary advisor: Frank Jackson
Secondary advisor: Gilbert Harman

September, 2010

For my mother, my father, and my grandmother.

Abstract

Part I of the dissertation argues for the **production view** of mental representation, on which that mental representation is a product of the mind rather than a relation to things in the world. I argue that the production view allows us to make best sense of cases of reliable misrepresentation. I also argue that there are various theoretical benefits of distinguishing representation from the tracking and other relations that representations might enter into.

Part II is about the relationship between representational content and phenomenal character, the "what it's like" to be in certain states. I argue for what I call the **phenomenal-intentional identity theory (PIIT)**, the view that phenomenal character is identical with representational content. In the course of arguing for PIIT, I argue that we need to distinguish representational content from what we might call "computational content," the type of content a state might be said to have solely in virtue of its role in a computational system.

Part III of the dissertation presents a view of the structure and content of concepts: the **efficient concept view**. On this view, concepts are structurally simpler and represent less complex contents than is usually thought, but can be unpacked to yield further related contents when needed. We can define various notions of **derived content** that capture these related contents. I argue that these notions of derived content can do much of the work that the notion of content was initially supposed to do. For instance, I claim that the type of content most closely related to folk psychological notions of content is a species of derived content. As a result, when we are interested in the truth-value of our thoughts, what we happen to be interested in is the truth conditions of a type of derived content, not of content proper.

The view that emerges is one on which mental representation does not play all the roles it is often thought to play, such as roles in computational or folk-psychological theories of mind and behavior, or roles in metaphysical theories of truth and reference. Rather, it turns out that these roles are played by other features of mental states, each crucial for understanding the mind and its place in nature, and each importantly related to mental representation.

Acknowledgements

This dissertation has benefited tremendously from countless interactions with friends and colleagues at Princeton and elsewhere. Embryonic versions of various ideas in this dissertation have been presented at various conferences, colloquia, and other forums at Princeton University, the Australasian Association of Philosophy, the Australian National University, the University of Waterloo, the University of Western Ontario, Cornell University, the University of Washington at St. Louis, the University of Texas at Austin, the University of Toronto, the University of Minnesota, and CUNY. I thank the audiences at those talks, and my commentators David Ivy, Mark Herr, Janette Dinishak, and Mike Collins for their probing comments.

Many people have provided detailed and incisive written comments on one or more chapters of this dissertation. I am grateful to Paul Benacerraf, Matt Ishida, David Pitt, Vanessa Schouten, Jack Spencer, Helen Yetter, Joshua Hershey, Corey Maley, Andrew Huddleston, Caleb Cohoe, Jeff Speaks, Uriah Kriegel, Cathal Ó Madagáin and Heather Logue. I am especially indebted to James V. Martin, Philipp Koralus, and Jack Woods, who have provided extensive comments on multiple drafts of multiple chapters and who constantly challenge my views.

I am also grateful for various meetings and discussions that have helped me see many issues in a new light. For these I thank Derek Baker, Mark Budolfson, David Chalmers, Tim Crane, Kati Farkas, Bill Fish, Bas van Fraassen, Tamar Gendler, Josh Knobe, John Maier, Matthew Moss, and Susanna Schellenberg.

My greatest debt is to my advisors, Frank Jackson and Gilbert Harman, for reading multiple versions of many chapters, as well as versions of chapters that never made it to the final copy, for offering guidance, and for fully supporting me, despite not always fully agreeing with my views.

Finally, I thank my friends and family, especially my parents, Sam Baker, Corinne Gartner, Elizabeth Ktorides, James V. Martin, Carla Merino, Vanessa Schouten, and Jack Woods, for their emotional and intellectual support throughout the past few years.

And as usual, all remaining mistakes are solely my own.

Contents

List of Figures

List of Tables

Part I

The Production View

Chapter 1

Introduction

You are presently having visual experiences, auditory experiences, and thoughts. These mental states in some way or other seem to be *about* things. Your visual experiences might be about, present, or represent the color of this sheet of paper, the shapes of the marks on it, and the words that they form. Reading these sentences might also cause thoughts in you that are about this sheet of paper, words, and mental states. *This* is **mental representation**.

Mental representation is supposed to be philosophically problematic in ways that other kinds of representation are not. We can see this by contrasting mental representation to conventional representation and other kinds of non-mental representation. In the case of **conventional representation**, what a representation represents is in a certain important sense up to us. When planning a bank robbery, we can use a salt shaker to represent things involved in our crime, say, an escape car. The salt shaker gets to represent the escape car simply because we stipulate that it does. If we change our minds and decide the salt shaker is to represent the driver of the escape car, then we thereby change the salt shaker's representational content. The salt shaker does not specify its own interpretation. Rather, what it represents is in a certain sense up to us and can change as our uses and intentions change. It is open to reinterpretation and even multiple interpretations (say, if we misunderstand each other and each take the salt shaker to represent different things). In contrast, genuinely mental representational states are not open to interpretation, reinterpretation, or multiple interpretations. What a mental state represents is not up to us. I cannot stipulate that the states of some region of your brain are to represent the movements of the planets and thereby change the mental representational content of those states, even if I find your brain states useful for my reasoning about the movements of the planets. If the brain state you're in corresponds to your thinking about Santa Claus and the movements of his reindeer, then your state mentally represents Santa Claus and the movements of his reindeer, and no amount of stipulation on my part can change that. Mental representations are not open to interpretation in the way that conventional representations are.

This is not to say that mental representations might not also count as other kinds of representations. In the above example, the very same brain state genuinely mentally represents Santa Claus and the movements of his reindeer, but it also conventionally represents the movements of the planets, thanks to my stipulation. And so, rather than distinguishing between mental-representational *states* and other kinds of representational states, it is more precise to distinguish between mental representational *features* and non-mental representational *features*.

There are other types of non-mental representation. We might say that a cloud of smoke represents a fire in virtue of being caused by the fire. Call this type of non-mental representation **causal representation**. Like conventional representation, causal representation is tolerant of multiple interpretations. The smoke can correctly be said to have been caused by various events, such as the tossing of a lit cigarette, the passing by of a negligent smoker, and a failed governmental anti-smoking campaign. But there is no conflict here; in the same sense in which the smoke represents the fire, it can be said to represent all its other causes. In contrast, even if mental representation has something to do with causation, it is importantly different from regular cases of causal representation in that it is strikingly intolerant of multiple interpretations.

Likewise, we might say that one domain represents another in virtue of being **isomorphic** to it, regardless of whether there is a causal mechanism responsible for the isomorphism (perhaps this is the sense in which calculators represent addition). But again, there is no real problem if it turns out that a representing domain is isomorphic to multiple potentially represented domains—it can be said to represent all of them in the same way, even if we choose to focus on some but not others.[1] In contrast, mental representations do not tolerate multiple interpretations. A mental state representing Santa Claus and the movement of his reindeer mentally represents just that—*Santa Claus and the movement of his reindeer*—and nothing else that it might, say, happen to be isomorphic to.

The special difficulty involved in understanding mental representation, then, is this: Non-mental representation is cheap and easy. It tolerates openness to interpretation, reinterpretation, and multiple interpretations. But genuine mental representation is strikingly intolerant of multiple interpretations and reinterpretations. Genuinely mental representational contents seem to bear a more selective connection to the states that represent them than do non-mental representational contents. This makes it very difficult to understand mental representation on the model of non-mental representational relations, such as those involved in conventional representation, causal representation, or relations of isomorphism. Unlike theories of non-mental representation, a theory of mental representation has to attribute unique contents to mental states. It has to rule out possible contents that, when

[1] See Cummins (1994) for discussion.

we try to understand mental representation on the model of non-mental representation, try to rule themselves in. A possible diagnosis of the source of this difficulty is that mental representation involves a more intimate relation to a content than non-mental representation most naturally allows for.

Rather than try to fit mental representation into the mold of various kinds of non-mental representation, I take these and other observations to support the idea that mental representation is fundamentally different from other kinds of representation. The main difference, I will argue, is this: Non-mental representation is (at least usually) relational; the representing domain bears some stipulated, causal, or other kind of relation to the represented domain. But mental representation is not a relation. Rather, it is a product of states or processes in the mind or brain. And this is why mental representation appears to involve such an intimate connection between vehicles of mental representation and their content. I argue for this view of mental representation, which I call the **production view**, in Part I.

I mentioned above that a state can have both mental representational features as well as non-mental representational features. This idea will play a key role in what follows. I will argue that many of the roles we take mental representation to play—roles in helping us get around in the world, detecting biologically significant features of the environment, and even roles in folk psychological theories of mind—are in fact played by non-mental representational features of mental states, and that understanding the relationship between a state's mental representational features and non-mental representational features is key to understanding mental life and behavior. In Chapters 4–6, I argue that although tracking relations do not constitute mental representation, they play an important role in explaining the success of behavior produced by mental representations. In Part III, I argue that thoughts and concepts do not genuinely mentally represent most of what we intuitively take to be their contents, but rather merely derivatively represent those contents, in much the same way as a salt shaker might derivatively represent an escape car in virtue of our uses and intentions.

Part II argues that genuine mental representation is identical to another problematic mental feature: phenomenal consciousness. This is true not only of perceptual states, where the thesis seems most plausible, but also of pain states, which might seem to be phenomenal but not representational, as well as of thoughts, which might seem to be representational but not phenomenal.

In summary, the dissertation argues for the following three views:

1. **The production view**
   Mental representation is a *production* of the mind or brain, as opposed to a *relation* to something else. (Part I)

2. **The phenomenal-intentional identity theory**
   Phenomenal character is identical to representational content. (Part II)

3. **The efficient concept view**
   Concepts have fairly impoverished contents but can be unpacked to yield more complex contents. (Part III)

The end result is not a solution to the problem of mental representation, but rather a limitation and confinement of the problem in several ways. First, on the view I propose, there is less genuine mental representational content than we might have initially thought. For instance, thoughts don't genuinely mentally represent the kinds of contents that can determine what we intuitively take to be their truth-conditions, and non-conscious states do not genuinely mentally represent anything at all. Second, mental states have both genuine mental representational features as well as non-mental representational features, and the non-mental representational features do a lot of the work that it is usually thought the genuine mental representational features do. For instance, mental states bear tracking and other relations to things in the environment and each other, and this partly accounts for the successful behaviors generated by those states. In the case of non-conscious states, these kinds of non-mental representational features are all we need to understand them. In the case of conscious states, the non-mental representational features interact with the genuinely mental representational features to generate successful behavior. Thoughts, too, have non-mental representational contents that they obtain derivatively from genuine mental representation, much like conventional representations get their contents from our uses and intentions. Third, genuine mental representation just amounts to phenomenal consciousness. While this solves neither the problem of mental representation nor the problem of consciousness, it reduces the two problems to just one.

The main claim of this dissertation can be summarized thus: Genuine mental representation is identical to consciousness, and everything else that appears to be both mental and a matter of representation is not genuine mental representation, but either in some way derived from genuine mental representation, or a clear case of non-mental representation.

Chapter 2

The Problem of Mental Representation

2.1 Mental representation is a phenomenon

You see things, hear things, taste things, smell things, and feel things.[1] Right now you are probably seeing various shapes and colors. You are seeing this paper, perhaps some pens on your desk, or parts of your own body. Likewise, you might be hearing noises, voices, or music. You are having **perceptual experiences**. These experiences **mentally represent** things, or the ways things are or might be. They "say something." They represent this paper, that it is in front of you, and so on. What an experience represents is its **content**.

You are also presently thinking about things and the ways things are or might be. This text is causing you to think about your experiences, or perhaps you are distracted and thinking about something else. Thoughts, beliefs, and other such states also represent. They also "say something," or have content.[2] This is supposed to be obvious and uncontroversial. There are phenomena that we notice in ourselves and readily describe as our seeing things, hearing things, and thinking about things. What is controversial is what these states consist in, how to explain them, and exactly what contents our mental states represent. But what should be obvious and uncontroversial at this point is the existence of the phenomenon itself. I take that as given and as my starting point.[3] This dissertation is about *that*. When I use

[1] I use terms like "see" and "hear" non-factively. Readers who object to this usage may substitute more cumbersome expressions like "have a visual experience."

[2] Perhaps different kinds of states have contents in different kinds of ways, or have different kinds of contents. In using the term "mental representation" to pick out the phenomenon observed both in perceptual states and thoughts, I do not rule out this possibility from the start (see also §2.3.1).

[3] In taking certain intuitive judgments as uncontroversial and obvious, I am not claiming that they are unrevisable. It might turn out upon closer inspection that some of our initial observations were mistaken. Maybe we make certain systematic errors in describing, say, our visual experiences. Then we might have reason to think that we don't really see some

the term "mental representation" I refer to whatever it is that we notice when we notice that we see things, hear things, etc. When I use the term "content" I refer to whatever our visual experiences, auditory experiences, and thoughts are of or about. In other words, I am picking out the phenomenon of mental representation ostensively by pointing to **paradigm cases** of mental representation.[4]

This fairly minimal way of picking out the phenomenon of mental representation is not yet committed to any view of what types of things representational contents are. They could turn out to be ordinary objects and properties, sense-data, ideas in a world of forms, or something else. They could be propositions, or something non-propositional. The phenomenon we've been noticing is our representing particular contents, not the ontological kind of these contents. My starting point also does not make any assumptions about what it is that does the representing, or what the representations themselves are. Representations could be, for example, brain states, functional states, symbols in a language of thought, or states of an immaterial soul. Again, our initial observations do not speak to the question of what are the states doing the representing. They are neutral with respect to these various options.

This way of picking out the phenomenon is also neutral with respect to what mental representation consists in. Although the term "representational" is sometimes contrasted with "presentational," our observations so far are compatible with views on which the phenomenon we observe is a matter of our being directly presented with wordly things and properties, as on direct realist views and at least some versions of disjunctivism. It is also compatible with adverbialism, on which experience amounts to modifications of the subject, rather than relations to things represented. These, together with theories we standardly think of as theories on which mental states represent, are all theories of mental representation, as I am using the term "mental representation."

Before moving on, it is important to distinguish my way of picking out the phenomenon of mental representation from another way that would be less interesting for our purposes. Instead of using our accessible seeings, hearings, and thoughts to draw attention to the phenomenon of mental representation, this less interesting way would be to take mental representations to *just be* those seeings, hearings, and thoughts that are accessible in this way, let us say, **introspectively**. I want to be clear that this is not my suggestion. Rather, representation is a feature that introspectively accessible states have in common apart from their being thus accessible. I don't want to rule out

---

of the things we intuitively think we see. In Part III, I will argue that this is our situation with respect to many thoughts and other states involving concepts.

[4]Although in what follows I assume that there are vehicles of representation, it is possible to reformulate the crucial aspects of the views I will ultimately defend such that it is compatible with views on which there are no vehicles, only contents.

the possibility of there being other states with the same feature that fail to be introspectively accessible.[5]

Rather, the importance of introspective accessibility is this: It provides us with a more or less theory-independent access to some instances of mental representation. Different theories of mental representation disagree on which states have content and what content particular states have. Some of these issues may have to be decided solely or mostly based on theory. This makes it useful to have theory-independent access to at least some of those mental contents. Instances of representation that are readily available to introspection are paradigm cases of mental representation. The hope is that we can use paradigm cases to test the various competing theories, and we can then use our theory to settle some of the issues surrounding non-paradigm cases.[6,7]

2.2 Other ways of approaching the problem of mental representation

Sometimes there is more than one way of picking out the same thing. We can pick out Venus using the description "the evening star" and we can pick out Venus using the description "the morning star." Likewise, there might be other ways to pick out mental representation. The phenomenon I described above is often thought to be related in some way or other to truth and reference, the posits of folk psychology, the posits of cognitive science, and our ability to navigate our way around the world. Instead of picking out mental representation the way I do (i.e. as a phenomenon we observe) one might choose to pick it out by one of its other alleged features. In this section, I will sketch several possible alternative starting points and explain why I prefer my own (excepting the cases where the proposals are more or less equivalent to my own). The reason will be the same in each case: There is a live possibility that mental representation, as picked out in my way, fails to have the features the alternative proposal recommends we use to pick it out.

[5]Ultimately, I will argue that there are no nonconscious representational states, but I will remain neutral on the question of whether there are introspectively inaccessible representational states. My claim here is that a notion of representation that took this to be true as a matter of definition would be a less interesting notion.

[6]Calling these cases "paradigm cases" does not imply that there is something metaphysically special about them, but merely reflects the fact that we have some introspective and relatively theory-independent access to them.

[7]In a similar spirit, Nicholas Georgalis (2006) recommends the following principle of methodological chauvinism: "The foundation for the development of any theory of representation should use a restricted database consisting of elements established from the first-person perspective to be representations or representational states." (Georgalis, 2006, p. 285) My suggestion is slightly weaker. I want to emphasize the role of paradigm cases in picking out the phenomenon to be explained without banning the possible use of non-paradigm cases in aiding the development of our theory.

2.2.1 Folk psychology

We attribute beliefs, desires, and other such states to others, as well as to ourselves. One way to approach the general phenomenon of mental representation is to take content to be that which is believed, desired, seen, etc., according to the folk uses of those terms. For example, when we attribute to Alex the belief that snow is white, we can say that Alex is in a representational state with the content *snow is white*. When we describe Sam as seeing an apple, we can say that Sam is in a representational state whose content is an apple. Call this type of content **folk content**.[8,9,10]

An advantage this sort of approach based on folk psychology might be thought to have is that, depending on how the details of folk theory turn out, it might offer us a richer notion of mental representation to work with. Mental representation is not just a phenomenon we notice when we pay attention to our experience; it has other features as well. Perhaps, according to folk theory, causal potency is a feature of content. Perhaps we think that what we believe, desire, see, and hear at least partly causes us to behave the way we do. This might allow us to make some headway on answering some important questions about mental representation. We could start by looking for the states that have the requisite effect on behavior and go from there.

The potential richness of such notions, however, is my reason for not taking this approach. It could turn out that what we notice when we pay attention to our experiences lacks any of the extra features attributed to it by folk psychology. Then, in the worst-case scenario in which the feature in question is considered crucial by folk psychology, we would be forced to conclude that there are no mental representations.[11] In some better case scenarios, our beginning assumption that contentful states have the features in question might have lead us to unpromising research programs or false theories. Of course, it might ultimately turn out that our folk psychological concepts include only the features the types of states picked out by my favored method actually have. However, given the availability of an independent and less committal way of picking out the phenomenon in question, we need not presuppose this from the start.

---

[8] Names of concepts and other mental representations appear in SMALL CAPS, while names of contents are *slanted*. For simplicity, I also use small caps for variables ranging over representations and slanted text for variables ranging over contents. **Bold** signals important uses of new terms or topics, such as their definition or introduction. *Italics* are used for emphasis and other conventional uses.

[9] David Lewis (1974), David Braddon-Mitchell and Frank Jackson (1996), Fred Dretske (1995) and Jerry Fodor (1987) place some weight on certain commitments from folk psychology, in particular with respect to folk psychological explanations of behavior.

[10] There might be more than one notion of folk content. It will turn out that this possibility is naturally allowed for on the view I will ultimately defend; see Chapter 11.

[11] This is the scenario that Paul and Patricia Churchland take to actually be the case. Notice, however, that they do not think there are no representational states of any sort (Churchland, 1989). It seems they would agree with me that there is a phenomenon of mental representation apart from that presupposed by folk psychology.

At this point, a methodological point about philosophy of mind is order: The assumption crucial to the approach based on folk psychology is that when it comes to mental states, our everyday concepts are a good starting point for inquiry. In some areas of philosophy, arguably such as normative ethics, there may be little empirical evidence that can bear on debates, and it might turn out that all or most of what we have to work with are our concepts and intuitions. Whether or not this yields successful research programs is an open question. Philosophy of mind, however, is not in a similar situation. We have empirical evidence available surrounding the phenomena we study. This is not only empirical evidence about brains and behavior, but also first-person observational evidence of the mental features of mental states, e.g. of their contents. For this reason, we do not need to appeal to folk concepts and intuitions to approach, define, or offer preliminary theories about our topic of study. We have the option to just observe it and go from there. Given that this starting point is less committal, as I've argued above, I claim that it is preferable.

2.2.2 Cognitive science

At least some brands of cognitive science aim to explain mental processes and behavior in terms of computations operating over mental representations. These representations are described as representing something, or having content. Call this type of content attributed to inner states by cognitive science **computational content**. The suggestion for an alternative way of picking out the phenomenon in question, then, is this: Representations are whatever computations are supposed to operate over in cognitive scientific theories, and representational content is computational content.

There are interesting questions in the philosophy of cognitive science surrounding computational content. What is computational content? What role does the notion of computational content play in theories in cognitive science? What do cognitive scientists commit themselves to when they attribute computational content to inner states? Some philosophers explicitly claim to be trying to answer these types of questions and not the types of questions I'm concerned with. Robert Cummins (1994, pp. 278-9), for instance, specifically claims to be describing a notion of representation that is useful for computational theories of cognition, but not necessarily for other kinds of representation that might be of interest to most philosophers.

It could turn out that computational content and the representational content that we've been discussing amount to the same thing. Maybe our paradigm cases of content are just special cases of computational content. That would be exciting. But it also might turn out that these are two distinct types of thing. Maybe the best elucidation of the notion of computational content describes a different type of thing from the type of thing we've been observing. One *prima facie* reason to think this might be the case (although perhaps it is controversial once we have been corrupted by theory) is that

it makes sense to ascribe computational contents to states of artifacts that we do not really believe to have genuine representational powers, such as calculators and computers. If calculators and computers can be said to represent, it is at least partly because of how we interpret (at least some of) their states, not because they somehow manage to represent on their own. But it seems they do have *genuine* computational content, whatever that turns out to be. If this is right, then we cannot straightforwardly identify representational content with computational content.

It might still turn out that representational content is a *species* of computational content, or computational content that meets some further conditions, but now we are entering the realm of providing a theory of mental representation, rather than picking out the phenomenon of interest. If such a theory is false and we pick out the phenomenon of interest in this proposed way, we fail to capture the phenomenon that we picked out in my way, the phenomenon we were initially interested in. This is why I do not choose this approach for picking out mental representation.

2.2.3 Getting around in the world

For the most part, we manage to acquire the things we need, avoid the things that are harmful, move around without bumping into too many things, perform sophisticated actions involving many steps, re-identify objects over time, and in general get around in the world fairly successfully. How do we do this? One plausible suggestion is that we do this by means of internal representations of the world. Inspired by this way of thinking about things, we might take mental representations to be explanatory posits in a theory of the generation of successful behavior.[12]

Behavioral phenomena such as the above call out for explanation, and it may very well be that mental representation is a crucial part of such an explanation. What is less clear is how and in virtue of which of their properties mental representations play a role in our getting around in the world. To repeat, mental representations are states having content; they might be brain states, functional states realized by brain states, symbols in a language of thought, or states of an immaterial soul. On most of these views, mental representations end up having properties other than the properties of having particular contents, such as having a particular mass.[13] It could turn out, then, that content is causally impotent and it's these other properties of mental representations, say their syntactic or physical properties, that are responsible for their usefulness in helping us get around in the world.

[12]There could be versions of this approach that are also versions of the folk psychological or cognitive scientific approaches described in §2.2.2 and §2.2.1. Perhaps the commitments to folk psychology in Fodor (1987) and Dretske (1995) are best construed as a commitment to the folk story of the role of contentful states in causing behavior. Fodor's appeal to folk psychology emphasizes its usefulness in explaining behavior.

[13]This is Dretske's distinction between representational facts and mere facts about representations (1995, p. 3).

If this were the case, an explanatory theory of behavior wouldn't need to posit *contentful* representations at all, and so we would fail to capture mental representation as picked out in my favored way.

Relatedly, this proposal seems to rule out, or at least deem initially unlikely, the possibility that there are two or more important phenomena related to mental representation that explain our successful interaction with the environment. For example, we might want to distinguish the **content** of our internal representations, or what they "say," from the properties, objects, etc. that they **track** in the world. It could be that these two features of mental representations, their ability to represent and their ability to track, together conspire to generate successful behaviors. Ultimately, in Chapter 5, I will argue that this is the case. For now, it suffices that we do not start off by conflating the two by understanding mental representation as the features of inner states that account for successful interactions with the world.

To sum up, while it may be fairly certain that something like mental representations play a role in the generation of at least some types of behavior, it is unclear how exactly they do this, and in particular what is the contribution of their representational features, as opposed to their non-representational features. This is why I prefer to pick out mental representation in the way I have suggested *and then* ask the question of whether and how mental representations contribute to the generation of behavior.

2.2.4 Assessability for truth or accuracy

If mental states "say something," then it seems to follow that what they say can be either true or false, accurate or inaccurate, or veridical or non-veridical. If assessability for truth is taken to be a definitive feature of mental representation, then there is the possibility of approaching the problem of mental representation from another direction: from that of a theory of truth.

Here is one reason to resist this approach: Mental representation is in some sense prior to assessability for truth. Mental states are true, veridical, or accurate *because* they represent (as well as because the world is as they represent it to be). Further, if one is at all attracted to a compositional view of mental representation, on which there are simpler representations that can combine in various ways to form more complex representations or representational states, it is reasonable to suppose that at least some of the simple representations are not themselves assessable for truth, giving us another reason to resist the present approach. The perceptual representation BLUE, for instance, might be such a simple representation. Activated in the absence of any other representations, BLUE would not be true or false, veridical or non-veridical, or accurate or inaccurate. It would not manage to say enough to be a candidate for truth or accuracy. However, when combined in the appropriate way with other representations, we can get a representational state that represents a more complex content such as *there*

is a blue object at arm's length in front of me. Only then would we have a state that would be assessable for truth.

Further, we can imagine a creature that cannot form such complex representations. Such a creature can experience blueness, but cannot represent that an object is blue. Such a creature would have states that "say" *blue!* but no states that say that any particular thing is blue. A natural way to describe such a creature is as being capable of representing but not of representing truly or falsely.

One might object that I am misdescribing the case: such a creature would not have a *representation* of blueness, but rather a (let's call it) **schmepresentation** of blueness. More generally, one might claim that non-propositional representational contents do not really count as representational contents. The idea might be, loosely, that when enough schmepresentations get together in the right way, we get representation. This may just be a terminological disagreement, but behind the choice of terminology lurks the assumption that schmepresentations and representations are different types of things, demanding different explanations. This assumption strikes me as false. The two instances of activating the concept CAT by itself and in the context of a propositional thought seem to be relevantly similar. By being activated in a certain context, the concept CAT does not transform from a schmepresentation into a component of an entirely different type of thing, a representation.[14]

Further, it is at least *prima facie* plausible that certain representational states have non-propositional contents, such as desires for ice cream or fears of snakes. These states are relevantly similar to states that have propositional contents, such as the desire that I have ice cream now or the fear that there are snakes nearby. Of course, on some theories, it is false that there can be non-propositional representational states, but our initial approach to the problem should not rule out these possible views.

Finally, here is a point that does not move me very much, but that might move some: we readily extend our everyday notion of representation to representations representing particular objects or other non-propositions. For instance, we say that the word "cat" represents cats. The fact that we do

[14]One might suggest that the important extra ingredient that transforms schmepresentations into representations is the having of a compositional structure. And so, schmepresentations and representations are different types of things because representations, but not schmepresentations, involve compositionality. However, this is not the case. There are compositional schmepresentations as well, such as SMELLY TIGER. Compositionality does not always result in representations with propositional structures that can be assessed for truth and so it cannot be used to draw a representation/schmepresentation distinction.

not use a distinct term, such as "schmepresents" to describe such cases might be taken to suggest that we are dealing with a unified phenomenon.[15,16,17]

Moving on, one might suggest an alternative proposal on which the property required in order to qualify as a mental representation might be assessability for truth, accuracy, *or* **reference**, rather than assessability for truth or accuracy alone. The perceptual representation of blueness and other such representations that fail to be assessable for truth or accuracy must at least *refer* to objects, properties, or some such.

A reason to resist both the original proposal as well as this amended version is that it might turn out that some representations do not manage to determine truth conditions or reference without the help of additional representations. There is a view in the philosophy of language on which the literal meaning of sentences sometimes fails to determine truth conditions and the reference of terms (see Chomsky (2000) and Atlas (2005)). Rather, on this view, truth conditions and reference conditions come in at the level of speaker or hearer meaning, after pragmatic implicatures have been computed. There is an analogous possibility in the philosophy of mind: It could be that some mental states do not determine truth conditions or conditions of reference all by themselves, but require certain further representations to do so. Ultimately, I will argue that something like this is in fact the case for thoughts. But for now, it will suffice that we do not rule out the possibility by definition.

One might suggest a weakened version of either of the proposals: A necessary feature of mental representations is the *potential* to form states that can be true or false, or at least that can refer. The strength of this proposal depends on what else we say may be added to a representation to secure truth or reference. If the constraints on what may be added are suitably relaxed, then the proposal will be fairly unobjectionable. Unfortunately, it will then fail to be a very useful constraint on what may count as a representation. If the constraints on what might be added are too strict, the constraint may fail to be satisfied by anything. Perhaps there is a happy medium, but it is now starting to seem that we would require a substantive theory of mental

[15]This point fails to move me very much because it is not obvious that the everyday notion of representation that we apply both to mental states and linguistic representations such as words and signs picks out a unified kind. Rather, it is quite plausible that mental representation and linguistic representation are different types of phenomena, deserving different explanations. Still, one might agree with this and nonetheless be slightly moved by the consideration under discussion if one thinks that these different phenomena have some features in common, including not requiring representations to have a propositional structure.

[16]There is another type of opponent who will deem assessability for truth or accuracy necessary in order for something to count as a representation. Such an opponent claims that there are two distinct mental phenomena here. There is something like "raw matter" and then there is an interpretation of the raw matter. But the raw matter is not itself representational. It is only after interpretation that we have a representational state. See §2.3.2.

[17]Thanks to James V. Martin for discussion on this and related points.

representation, truth, and reference in order to specify what it is, making such a constraint unhelpful as a starting point. Instead, I leave the question of the relation between representation and truth and reference an open question to ultimately be answered with the help of a theory of mental representation, rather than an initial constraint on what is to count as representation.

2.2.5 Intentionality or directedness

Some philosophers have described mental states as being somehow directed at things in the world. Franz Brentano took this directedness to be the mark of the mental (Brentano, 1874). Perhaps this affords another way to pick out the phenomenon of mental representation: Mental representations are the states that are **directed** in this way.

One way of understanding directedness (though not Brentano's) is as **reference**, potential reference, conditions of reference, or what specifies the conditions of reference. If this is the right interpretation, then we are back at the set of proposals discussed in §2.2.4, and the arguments there apply here as well.

Another way of understanding directedness is as **apparent reference**. It seems that we are in contact with things in the world in some way or other. We notice that we see things, hear things, and think about things, and we take most of these things to be external world things. If this is the way directedness is to be understood, then this picks out more or less the same phenomenon that we started off with in this chapter (perhaps with the additional assumption that we are generally naïve realists about the contents of thought and perception), and there is little for me to disagree with.

2.3 Objections

2.3.1 Perception and thought

One might object that perceptual states are importantly different from "conceptual" or "cognitive" states such as thoughts, beliefs, and desires, and so what I am calling "perceptual content" is not *the same kind of thing* as what I am calling "thought content." If this is the case, then the paradigm cases of mental representation are actually paradigms of two different kinds of phenomenon, which I am conflating.

First, observations concerning thought and observations concerning perception seem relevantly similar in that both thought and perception appear to "say something" or to be *about* objects, properties, states of affairs, or other types of things.[18] So it seems it might be fruitful to at least start off

[18]This is the case even if we think that perception and thought have very different kinds of contents, perhaps as on the view that perception supplies us with objects and thought attributes properties to those objects. Still, on such a view, it is natural to say

thinking about these two phenomena as relevantly similar and deserving of a similar explanation.

Second, in folk psychological or commonsense attributions of content, we treat perception and thought as relevantly similar. We say things like "I have to see it to believe it," where both tokens of "it" refer to the very same thing.[19] One might take this to be some evidence that perceptual content and thought content are relevantly similar. Even if one does not take this to be evidence on its own, perhaps because one is weary of such appeals to intuition, it at least lends some support to the first point that observations concerning thought and perception are relevantly similar: this observed similarity might at least partly explain our similar treatment of them in folk psychology.

The above are reasons to think perceptual representation and representation in thought are the same general type of phenomenon. However, even if it turns out that they are distinct phenomena, my starting point will not steer us too far in the wrong direction. One way to put the problem with some of the approaches that impose extra conditions on what is to count as mental representation is that they rule out possible views that should not be ruled out from the start. For example, understanding mental representational contents as explanatory posits in a theory of behavior might rule out the possibility that the representational features of representations are causally impotent. And the problem with approaches that might pick out something else entirely is that they fail to include the target phenomenon. However, the analogous problems do not present themselves here. Perhaps perceptual content and thought content are two entirely different kinds of things. Then we would need two distinct, and perhaps unrelated, theories to explain them. If we start off thinking of perceptual content and thought content as relevantly similar, then it might take longer to reach such a conclusion. However, such a conclusion has not been ruled out from the start. Nothing in the way we picked out mental representation requires that it be a unified phenomenon.

Of course, whatever their apparent similarities, perception and thought also seem quite different in certain respects. For instance, perceptual representation is more vivid, detailed, and closely related to phenomenology, than is representation in thought. Eventually, I will offer a view of mental representation that begins to explain both the similarities and differences between perception and thought.

---

that perception and thought both exemplify aboutness; they just are about different kinds of things.

[19] Of course, there are contents that can be seen but that cannot be believed, e.g. *that tree*. What is relevant to my point here is that, according to folk or commonsense attributions of content, at least some contents can be both seen and believed.

2.3.2 "Raw matter" and interpretation

According to some views, it is a mistake to think of perceptual states, or at least some types of perceptual states, as *representing* the world. For example, take the visual experience one enjoys when one views a red ball. This experience (or some aspect of it) might be involved in a state that represents the ball as red in normal lighting conditions, or it might be involved in a state that represents the ball as white but under red light. The visual color experience (or the aspect in question) does not by itself "say" which of these two cases one is dealing with. It is silent between them. So, one might argue, perhaps it doesn't represent any particular colors at all. Rather, perhaps it is a further state, such as a judgment or interpretation of the initial state, that represents the ball as having a particular color. On this view, there is a distinction between non-representational mental features of some perceptual experiences, which we may call raw matter, and further judgments or states about or based on the raw matter, or interpretations. At least in the case of perception, representation only comes in at the level of interpretation. Raw matter is non-representational. If this view is correct, then my approach is too liberal: it includes both raw matter and interpretations, whereas we should only include interpretations.

However, if the stated view is correct (i.e. if (what we might call) a redish-ballish-raw-matter is not itself representational), then it is not in fact true that I am including too much in my initial characterization of mental representation. My observations have to do with *representational contents*, the things that states "say," not with features of mental states that do not represent. And so, my way of picking out the phenomenon of mental representation isn't meant to and shouldn't include uninterpreted raw matter, if there is such a thing. Instead, assuming the raw matter view is correct, my method should pick out *interpreted* raw matter. Ultimately, a full theory of mental representation should isolate these two factors and distinguish their contributions to the phenomenon of mental representation that we observe and want to explain. Nothing in my methodology rules out such a possibility of there being two distinct contributors to mental representation the start.[20]

[20]One might also argue that it is not entirely clear that the alleged raw matter in the examples offered is not itself representational. There seems to be something representationally similar between a state that represents something as being red and one that represents something as merely looking red but actually being some other color. What these two states have in common is the redness bit. Of course, the proponent of the theory we are discussing would say that the redness bit is not itself content, but rather raw matter. However, I am suggesting that we might alternatively say that the redness bit is a common component of the *content* of the two experiences. Both states involve a common representation of redness but differ in whether the redness is represented as merely apparent or actually a property of the object in question. Thanks to Susanna Schellenberg for discussion on related points that inspired the present points.

2.4 Why we're not just talking past each other

When different theorists pick out their topic of interest in different ways, there is a danger that they will end up talking past each other. Fortunately, regardless of what additional features various theorists attribute to mental representations, they are supposed to be theories of the intuitive *aboutness* that I am taking to be crucial (perhaps among other things). These theories are supposed to apply to the noticeable contentful states, the paradigm contentful states. Theorists taking other approaches often use paradigm mental states to illustrate their points, such as visual representations of color, concepts of farmyard animals, beliefs about the color of snow, and hallucinations of pink rats and daggers. Although there could be non-paradigm representational states with such contents, the examples are usually supposed to be of introspectively accessible states. So if these theories are successful, they should also be theories of content in my minimal sense of the term, and they should also apply to paradigm cases.

2.5 The other putative features of mental representation

I have chosen and argued for a fairly minimal way of picking out mental representation. We notice that we are in states that say something about the world. This saying something about the world that we notice is mental representation. This dissertation argues that (1) mental representation, as picked out in this way, is not a matter of being related to things in the world, but is instead a product, process, or in some sense a production of the mind or brain, (2) mental representation can be identified with phenomenal consciousness, and (3) the first two claims are true not only of perceptual representational states and other states that obviously are phenomenally conscious, but also of thoughts.

In this chapter, I dismissed several other ways of picking out mental representation via some of its alleged additional roles. Thus far, I have argued that mental representation *might not* play these roles. A secondary line of argument in this dissertation argues that mental representation alone in fact *does not* play these roles. Rather, these roles are played by other, sometimes related, phenomena. In Chapters 4 to 6, for example, I argue that a crucial part of a story of how mental representations contribute to successful behavior must invoke non-mental-representational features of representations: tracking relations obtaining between mental representations and things in the world. In Chapter 11, I argue that folk psychological notions of content track something other than but related to occurrent perceptual and thought content. I argue that for the purposes of predicting the behavior of others, this is a good thing for our folk psychological concepts to track. This offers a partial vindication of folk psychology, although with important qualifications.

In Chapter 7, I suggest that computational content is something distinct from representational content. In Chapters 10 and 11, it will turn out that regular thought content does not specify conditions of truth or reference, or at least does not specify the truth and reference conditions we care about. Instead, it is folk psychological notions of content that specify the conditions of truth and reference that we care about. The end result is a picture on which mental representation, as picked out ostensively, is a product of the mind that is identical to phenomenal consciousness, and the various other roles mental representation is sometimes thought to be play are in fact played by different, although sometimes related, phenomena.

Chapter 3

The Metaphysical Problem with Relation Views

Production views of mental representation state that mental representation is a process or product of a subject's mind or brain. The mind-brain identity theory, short-arm functional role semantics, and the phenomenal intentionality theory (see Chapter 7) are examples of production views. **Relation views**, on the other hand, claim that relations to items that serve as a representation's content (usually items in the external world) are *constitutively* involved in mental representation. Direct realist theories of perception and tracking theories of mental representation are types of relation view. On relation views, mental representation is a *relation*. On production views, mental representation is a *non-relational*, or *intrinsic*, property of subjects.[1] Production views and relation views might agree that subjects of mental representation both undergo brain activity and bear relations to the world, but they disagree on which of these two factors is constitutive of mental representation. Production views claim it is the inner workings of the mind that is responsible for mental representation, while relation views claim that it is a relation to represented properties and entities.[2]

The next few chapters provide arguments against relation views. The remainder of the dissertation fleshes out a particular kind of production view, one on which mental representation is identical to phenomenal consciousness in both perceptual and thought. This chapter starts my case against relation views by expressing a metaphysical worry with such views. While this may not by itself convince everyone, I believe that this worry gets at the real problem with relation views. In the next two chapters, I set such "metaphysical" considerations aside and offer empirical evidence against relation views.

According to relation views, a mental state represents just in case and because it bears the appropriate relations to the appropriate things. The

[1] This does not preclude mental representation requiring certain relations to hold between a subject's various states, as on certain versions of functionalism.

[2] Hybrid views are possible as well. I will later suggest that some versions of the sense-data view might count as hybrid views.

problem is that it's hard to see how to reconcile this with the following two claims:

(1) A mental state's representing one thing rather than another (or nothing at all) makes a difference to how things are for the subject of the state.

(2) A representational state is very often located somewhere other than where what it is representing (or what it is allegedly related to in virtue of which it gets to represent) is located.

(1) is obvious to introspection: it makes a difference for me whether I am representing an octopus or the state of the economy. (2) should be accepted by most versions of the relation view on which the relevant relata are worldly relata such as objects, property universals, property instances, or the like.[3]

Let's suppose that a relation view is correct and that (1) and (2) are true. Let's further suppose that I am in some representational state A (perhaps A is a brain state, a functional state, a state of an immaterial substance, etc.). Since A is a representational state, according to the relation view, some relation R must hold between A and some thing or things in the world. Given (1), it is in virtue of this relation R that things are a certain way for me. Perhaps I experience redness, or squareness, or think about an octopus. A represents because it bears relation R to something and A's representing makes a difference for how things are for me. Further, by (2), there needn't be any direct contact between A and the thing in the world it represents (or the thing in the world it is related to in virtue of which it represents what it does). This is surprising! Here I am, minding my own business in my own region of space and time; my mental state bears some relation to things elsewhere, and the bearing of this relation *makes a difference for me* right here, right now. How things are right here, right now, for me, depends on what I bear this relation to. In other words, bearing a relation to something seems to make a difference to what quite plausibly seem to be my **intrinsic** properties, the properties I have in virtue of how I am from the skin in.

What kind of relation can affect me in this way? I suggest that when we really start to think about it, this relation starts to look a little magical. Let's try to see this by comparing it to other, tamer, relations.

Many ordinary relational properties do not make a difference for their relata: Being a sister does not constitutively change your intrinsic properties; being thought to be worth a certain amount does not change the intrinsic properties of a ten dollar bill; being parodied by Tina Faye does not make Sarah Palin any less intelligent. So we find that many ordinary relations don't change the (non-relational) way things are for their relata.

There is a class of exceptions. These are relations that constitutively involve a causal relation. For example, the kicking relation constitutively

[3]The exception is a relation view on which the relevant represented relata are mind-dependent entities such as sense-data. See p. 26 for discussion of this view.

involves a causal relation between its two relata.[4] You can't kick someone without causally interacting with her. That's part of what it is to kick someone. Being kicked by someone usually changes some of one's intrinsic properties, such as the shape or location of the boundaries of one's skin. In general, relations that involve a causal interaction between the two relata can result in an intrinsic difference for them.

Could the representation relation be relevantly like the kicking relation? On some views of mental representation, the representation relation is or involves a species of causal relation. But one big difference between kicking and representing is this: Kicking changes my intrinsic properties *only* because of its causal effects on me. As long as I've got the effects, I've got the change in intrinsic properties. The extra fact that it was a kick, as opposed to a punch or something else, that resulted in these effects is irrelevant to how things are for me, as least insofar as my injury is concerned. In contrast, with mental representation, bearing the representation relation to something is supposed to change how things are for me *over and above* any causal effect the things represented have on me. After all, everyone, including production theorists, should agree that the world causes changes in us, including changes in what representational states we're in. The world can cause one to be in one representational state rather than another, and our representational states can influence the behaviors we engage in. And the world can also cause one to form some concepts and not others. In a world without frogs, it's unlikely that anyone would form the concept of a frog. The production theorist claims that the role of the world is limited to these kinds of causal interactions. According to production views, the world changes things for us only through its causal effects on us. Relation views, in combination with (1) and (2) above, maintain that the world plays an additional role in determining how things are for us. What's going on in the world has an effect on us that outruns its causal effect on us. It not only causes things to be different for us, but also *constitutes* things being different for us.

To repeat, the worry here is that this representation relation is unlike ordinary relations. It looks a bit magical. This worry gets a little bit more worrisome if we are also committed to **intentionalism**, the view that **phenomenal character**, or the "what it's like" (Nagel, 1974) of being in a certain state, is reducible to representational content. If we are intentionalists, phenomenal character must be explained in terms of this relation too. In virtue of this relation obtaining between me or my states and some other things, what it's like to be me also changes. This is worrisome, since *what it's like for me* looks like an intrinsic property of me. Again, what kind of relation could behave like this?[5]

[4] If you think causal relations hold only between events, then read that as shorthand for: The kicking relation constitutively involves a causal relation between events involving the two relata.

[5] Some intentionalists accept externalism about phenomenal character, but this is generally considered to be an unfortunate consequence of their views.

I am not just complaining that some relation views are externalist, where **externalism** about mental representation is the view that intrinsic duplicates can differ in their mental contents, where **intrinsic duplicates** are molecule-for-molecule duplicates. This vague metaphysical worry also applies to internalist relation views, where **internalism** is the negation of externalism. Take an internalist view on which the environment determines mental states in exactly the same way for intrinsic duplicates, where **intrinsic duplicates** are subjects who share all their intrinsic properties. It is still just as much a mystery how a relation that affects intrinsic duplicates alike gets to affect them in the first place.[6]

What if we deny one of the two assumptions needed to generate this worry?

Suppose the relation theorist rejects (1) while accepting (2). Then what I'm representing does not make a difference to how things are for me. It does not make a difference for the physical and functional facts obtaining within the boundaries of my skin. And it does not make a difference for my intrinsic mental features. Why do things seem to differ for me when I represent, say, *cat* rather than *octopus* or *blue*? Perhaps this is because there is another phenomenon that we're confusing with mental representation: phenomenal character, or "what it's like" to be in a state. Perhaps phenomenal character is what's driving the intuition that representational differences make a difference for me. Perhaps the difference we notice between representing *cat* versus representing *blue* is a phenomenal difference, not a representational difference. And phenomenal character is determined by what is going on within the boundaries of my skin. This would drive a wedge between mental representation and phenomenal consciousness.

Unfortunately for this kind of response, phenomenal character is already infused with representation. Our experiences do not consist in a bunch of raw feels, occurring one after the other. Rather, they consist in an orderly appearance of colors, shapes, tables, octopi, and other putative objects and their features. In other words, phenomenal character already involves representation in some way or other. Indeed, some of the paradigm cases of representation that we started off with in picking out the phenomenon of mental representation are also paradigm examples of states with phenomenal character.

One might retreat to a weaker position on which the production view is true of many of the paradigm cases of representation, but the relation view is still true of some of them. Perhaps one is taken by the Twin Earth thought

[6]Frank Jackson is an internalist relation theorist of just this sort (see Jackson (2004)). If we instead understand externalism as the view that environmental factors at least partly determine the content of mental states, then Jackson's view is externalistic. It just so happens that the environment partly determines mental content in the same way for intrinsic duplicates. On this notion of externalism, I *am* just restating the fact that relation views are externalist, and I am arguing that this is a problem because the obtaining of the relevant relations make intrinsic differences for their subjects.

experiment (Putnam, 1975). Earthling Oscar is in an environment where the clear potable stuff in the rivers and streams is H_2O, while his intrinsic duplicate Toscar is on Twin Earth where the clear potable stuff is some other substance with a long and complicated chemical composition that can be abbreviated "XYZ." Intuition (supposedly[7]) has it that Oscar's watery-stuff-related thoughts represent H_2O, while Toscar's watery-stuff-related thoughts represent XYZ. Whether it is H_2O or XYZ that Oscar and Toscar represent does not make a non-relational difference for them. After all, they can't tell the difference between H_2O and XYZ, and if we could arrange for Oscar to enjoy one of Toscar's watery-stuff-related thoughts, he would not be able to tell it apart from his own watery-stuff-related thoughts. Doesn't this show that in at least some cases, being in a certain representational state does not make a difference for its subject?

I can't say everything I want to say about the Twin Earth thought experiment here. In Chapter 12, I will render harmless and then accommodate the intuitions it is thought to invoke (and, by the way, I will do so without presupposing the production view, making it non-question-begging to appeal to those arguments to support my present points). However, for now, let me just note that relation theorists pursuing this strategy should consider the position to which they have retreated unsatisfying, since it involves conceding that the production view is true of some instances of mental representation. Quite plausibly, most perceptual representation will be of this sort. Second, it seems to involve denying that representing the contents of Twin Earth-able representations (that is, representations thought to be such that Twin Earth thought experiments can be run on them) makes a difference for subjects. If representing the contents of Twin Earth-able representations does not make a difference for subjects, then not only does representing H_2O versus XYZ fail to make a difference for subjects, but so too does representing H_2O versus *gold* or *aluminum* (other substances thought to be Twin Earth-able). Either this will have to be denied, or it will have to be argued that the difference in representing, say, H_2O versus *gold*, is the result of something other than content.

One might suggest a hybrid view on which Twin Earth-able and perhaps other concepts have two sources of content. A production view is true of part of the content, and a relation view is true of the other part. To evaluate such a view, much more will have to be said about the role of the two sources of content and the relation between them. As long as the part of content that makes a difference to how things are for the subject is the part for which a production view is true, then such a view may not ultimately be in too much disagreement with the view I will argue for in Part III. However, if the part of content that makes a difference to how things are for the subject is claimed to be the part for which a relation view is true, then the hybrid view does not avoid the problems discussed in this chapter.

[7]See Cummins (1998).

What if we accept (1) and instead deny (2), the claim that a representational state is often located somewhere other than where what it is crucially related to is located. One way to deny (2) is to deny that I am located somewhere within the spatial boundaries of my skin. Maybe my mind or self extends out into the world where it can reach the things it represents and they can make a difference for it. A view like this doesn't sit well with the observed fact that destroying parts of the brain results in mental deficits, while destroying parts of the world has not been observed to result in mental deficits, except insofar as the destruction has a causal impact on one's brain. Maybe the relation theorist will insist that destroying parts of the world can sometimes result in changes in my mind, but we just haven't destroyed the right parts yet. Or perhaps the relevant relata are abstract objects, and abstract objects are indestructible. Still, whatever evidence we do have concerning what can be destroyed without affecting mental life sits better with a production view than with a relation view supplemented with either option.

Another worry about this sort of denial of (2) is that it makes **mental unity** more difficult to account for. Right now, you are enjoying visual experiences, auditory experiences, and thoughts. There is a sense in which all these representational contents are unified. They are present together, or "co-present," or in the same "arena of presence."[8] In contrast, my mental states are not unified with yours. Accounting for mental unity is difficult for everyone. It's hard to explain how different mental states can be unified in the observed way. However, a view locating the mind wholly within the head at least allows for the possibility of interaction between different parts of the mind. Maybe mental unity has something to do with the synchronized firing rates of neurons, quantum coherence, or some such. Whatever it is, it should be something that is disrupted or absent in cases where unity is lost or not present, arguably such as in cases of brain bisection. The point right now, though, is that there are some options here, and we might reasonably be optimistic that we'll come up with more when we find out more about how the brain works. But it's a bit more difficult to make sense of how a version of the relation view on which the mind extends beyond the boundaries of the skin can account for mental unity. If my mind is scattered all over the world, near, in, on, or identical to everything I represent, we would need to find a process or thing that can result in unity of a spread out or scattered mind but not unify split brains or distinct individuals that are, say, standing side by side. Perhaps one might suggest that there is a core self to which the rest of the self is in some way related. Such a solution opens up a whole new can of worms. This core self has to be somehow appropriately related to the rest of the mind that is spread out over the world, and that relation looks as magical as the apparently magical representation relation we are trying to tame. For one, the relation should not be merely causal, but rather

[8] I borrow from Mark Johnston's term "arena of presence and action."

should allow its relata to somehow constitutively affect one another. This core self should also be capable of splitting in two in the unlikely event of brain bisection. And for reductionistically-minded relation theorists, all this will have to be reducible to the physical and/or the functional, or at least metaphysically supervene on it with the required strength.

Another, perhaps more plausible, way to deny (2) is to deny that the things I represent are out in the world. Instead, perhaps they are in my head. Maybe they are something like mind-dependent sense-data.[9] We get to reject (2) because sense-data, on this version of the view, are in the same place as the rest of me. However, sense-data views involve a bit of a concession to the production view. The production view has it that something in our brains or minds produces content. It does this independently of special relations to the environment. The sense-data view we are now considering has it that we produce sense-data. It adds that once we produce them, we have to become appropriately related to them (perhaps by what is sometimes called the "awareness" relation). In virtue of this relation, we get to represent. The production view cuts the middleman. We don't have to get related to sense-data in order to have contentful states; we just produce the already contentful states themselves. Instead of producing a red sense-datum and then becoming related to it, we produce the "of-redness," or redness-content, which is not itself red, and which does not require us to be specially related to it in order for it to represent. All else being equal, we should prefer the simpler production view.

Of course, all else isn't equal. Sense-data theories face well-known problems accounting for indeterminate and contradictory representation, and this problem arises directly from their commitment to the middleman that the production view denies. On the sense-data view, having an experience with a certain content is a matter of being appropriately related to something that is just as that experience represents. For example, having an experience of a red circle is a matter of standing in a relation to a red and circular sense-datum. The problem with **indeterminate perception** is that there are no appropriately indeterminate objects to which we can stand in such a relation. An item in the periphery of one's visual field can appear to be of an indeterminate shade of pink, but it's hard to make sense of how anything can be pink without being a particular shade of pink, even if that thing is a sense-datum.[10] The sense-data theory faces a similar problem when it comes to

[9] See e.g. Ayer (1956), Jackson (1977), Robinson (1994), and Foster (2000).

[10] I borrow this example from Adam Pautz (2007, pp. 510-1). Pautz suggests that this kind of example is a better example than the standard speckled hen example because in the case of the speckled hen, it's not implausible that something different happens in the brain depending on the number of speckles the hen has, and this might make it plausible to say that the sense-datum does in fact have determinate properties. However, the analogous proposal in this example, that something different happens in the brain depending on what specific shade of pink the sense-datum has, is less plausible, and so one route that might be taken by the sense-data theorist to argue that the sense-datum in question is actually determinate loses much of its appeal.

contradictory perceptions, such as in the case of motion illusions, where something both appears to be moving and appears not to be going anywhere at the same time. The problem is the same: It's hard to see how any object, including a sense-datum, can be both moving and not going anywhere at the same time. In short, one main problem with sense-datum views is that accounting for indeterminate or contradictory perception requires positing objects with indeterminate or contradictory properties. And this is a direct consequence of the view's commitment to a middleman between representational states and representing. The production view avoids these problems by denying that anything need have the properties that are represented by experience. Representing something as being an indeterminate shade of pink does not require that there be a thing that actually *is* an indeterminate shade of pink. Similarly, contradictory representation does not require there to be objects with contradictory properties.[11]

To summarize, what we're representing makes a difference for how things are for us. The metaphysical worry with relation views is that it is hard to make sense of a relation that has this kind of a non-causal effect on one of its relata, or the subject bearing the state that is the relatum. The representation relation would have to be unlike any other ordinary relation we know of. And that's a reason to think that mental representation is not in fact a relation.

[11] There are other possible views, such as Sydney Shoemaker's view on which at least some representations represent by being related to the properties of being certain manifested dispositions of objects to cause certain qualia in us (for two slightly different versions of the view, see Shoemaker (1994) and Shoemaker (2003)). On this kind of view, representation involves being related to properties that are partly mind-dependent (the qualia bit) and partly mind-independent (the manifested disposition to cause). Whether or not such a view avoids the problem of the sense-data theory, it is still mysterious for all the reasons mentioned above how the mind-independent part of the relatum can make a difference for subjects in the way required.

Chapter 4

The Possibility of Reliable Misrepresentation

4.1 Reliable misrepresentation

In statistics, there is a distinction between a test's validity and a test's reliability. A **valid** test is one that fairly accurately detects what it is intended to detect. A **reliable** test is one that yields more or less the same results each time it is administered, whether or not it detects what it is intended to detect. For example, the Standard Aptitude Test (SAT) is arguably quite bad at predicting what it is supposed to predict, namely success in college, and so it is not a valid test. However, the test is reliable in that it yields more or less the same results when administered to the same subjects on distinct occasions.[1]

I claim that we need a similar distinction when it comes to mental representation, in this case between **reliability** and **veridicality**. We need this distinction because there are representations that **reliably misrepresent**. They are reliable in that they respond similarly in similar circumstances, but they are not veridical, since the world isn't really as they represent it to be. Loosely, reliable misrepresentation is getting things wrong in the same way most of the time.

In Chapters 5, 6 and 7, I will argue that color-representations, hotness-representations, sweetness-representations, and pain-representations reliably misrepresent. In this chapter, I will only argue that relation views, and in particular tracking versions of relation views, would have trouble accounting for cases of reliable misrepresentation, if there were any, and that this is a serious problem for those views.

Let me begin by sketching what an account of a representational state in terms of reliable misrepresentation might look like. Suppose color-experiences reliably misrepresent. Color-experiences are very good at tracking surface reflectance properties of objects, but what they represent doesn't quite match

[1] Thanks to Joshua Knobe for suggesting this analogy.

what they're tracking. Instead, they represent *sui generis* color properties that objects don't really have, like redness and blueness. On this story, color-experiences reliably misrepresent colors. They misrepresent, but they misrepresent in the same way most of the time.

As should already be clear, my interest is not in hallucination. Hallucinations are fringe cases analogous to occasional glitches in the SAT grading process due to, say, a power outage. Such glitches might alter the results of a few tests on a few occasions, but their occurrence is compatible with the overall validity of the test. Similarly, occasional hallucinations are compatible with the veridicality of most activations of the relevant representations. I am also not interested in illusions. Although they arise more regularly and predictably than hallucinations, they can be understood as unintended side-effects of otherwise veridical systems. They are analogous to the systematic distortion of the scholastic aptitude of test-takers for whom English is a second language. This kind of distortion is possible even for a generally valid test, one that gets most cases right most of the time, and even for an optimal test, one that is as good as it can be given its constraints (for instance, a constraint on the SAT is that it must be delivered in a language and not be too costly to administer). Likewise, illusions are compatible with the overall veridicality of a well-functioning system. So hallucination and illusion are not the kinds of cases of error I am after.

Rather, my focus is on cases of reliable misrepresentation: A representation R **reliably misrepresents** some content P when (1) R non-veridically represents P, and (2) R does or would non-veridically represent P in the same types of circumstances on separate occasions. In cases of reliable misrepresentation, there is always some property, no matter how gruesome, that is successfully tracked (either by being causally related to the representation, or in some other way). Since reliable misrepresentation is misrepresentation of the same sort in certain circumstances C, C is automatically tracked, and other features might be tracked as well. Compare this to the SAT, which, let us assume, is reliable but invalid. The SAT tracks *something*—that's why test-takers' grades are more or less constant across multiple applications of the test—but what it tracks is not scholastic aptitude. Likewise, a reliable misrepresentation tracks *something*, which is what makes it reliable, but that is not what it represents.

I choose to focus on reliable misrepresentation, rather than hallucination or other cases of non-veridical representation because, for reasons that will soon become clear, I think it is the most difficult type of case for the relation view to account for. However, my argument can be thought of an extension of the following argument schema that takes as its inspiration cases of hallucination, illusion, and other forms of error:

(P1) Relations have existing relata.

(P2) Mental representation R represents P.

(P3) P does not exist.

(C1) Therefore, mental representation is not a relation.

We can complete the argument by filling in R with a representation of a non-existent. The Meinongian rejects premise (P1), claiming that there are non-existent things that can serve as a relation's relata. The sense-data theorist denies (P3), claiming that P exists (it is a sense-datum).[2] This argument form against the relation view is clearly and convincingly developed in Kriegel (2007).[3]

Recent relation views can provide promising responses to this argument when P is a case of illusion or hallucination. A common strategy is to say that P is an abstract property that we can be related to even in the absence of its instances. This allows us to deny (P3) without adopting a sense-data theory. So far so good.[4] The challenge for the view that P is an abstract property is to explain how we get to be related to it. If the case in question is one of illusion or hallucination, then we may have had contact with *instances* of the

[2]The argument schema bears some resemblance to the argument from illusion. The argument from illusion assumes that mental representation is a relation, that relations have relata, and that there are cases where what is represented does not exist in the external world. The conclusion drawn is that what is represented exists in the mind. If the claims in the next few chapters are right, then, as adverbialists have claimed, the culprit in the argument from illusion is the premise that mental representation is a relation.

[3]Kriegel concludes that the production view is true, and in particular, that a phenomenal intentionality version of the production view is true (see Chapter 7). (Kriegel calls the resulting view "adverbialism"; I prefer "production view" to avoid unintended associations with the adverbialist views of Ducasse (1942) and Chisholm (1957), which differ in motivation and overall perspective from the type of production view I think we should adopt.)

[4]Kriegel (2007) argues against the view that P is an abstract property on various grounds. For one, sometimes what is represented is a concrete entity, not an abstract entity (at least in certain cases, such as the case of representing Bigfoot) (see p. 310–311). However, in response, the relation theorist appealing to abstract properties might suggest that concreteness is just another abstract property that is represented in our thoughts and experiences. In other words, all the relation theorist needs in order to account for the representational phenomena is that we represent things *as* concrete, not that we are related to concrete things (except insofar as this is required in order for us to represent things as concrete). Kriegel's other worries with the abstract property view are that it casts a veil of appearances over our knowledge of concreta and that it commits us to abstract properties. However, it's unclear how much of a problematic commitment a commitment to abstracta really is, and it's also unclear whether a veil of appearances is as problematic as it is sometimes made out to be. In any case, gappy content views such as those defended in Schellenberg (2010) and Tye (2009) seem to get the best of both worlds for those concerned with veil-free representational contact with concreta: In the "good" cases, we represent a proposition composed of abstract properties as well as concrete objects and/or property instances, while in the "bad" cases such as those of hallucination and illusion, we represent a proposition composed of abstract properties and a gap where a concrete object or property instance would go were we in the "good" case. I do not endorse the gappy content view—mostly because I don't think the gap or the entities that sometimes fill it actually do any interesting work—but I mention this option to suggest that the abstract property view may be viable even for those concerned with veils of appearances.

represented property on other occasions, and this might form the basis of our relation to it. But in the case of reliable misrepresentation, we may never have had the relevant contact with instances of the represented property, so, as I will argue, this strategy cannot work and the above-mentioned challenge of explaining how we get to be related to a reliably misrepresented property cannot be met on the relation view. In the case of what is perhaps the most prominent family of views of mental representation, causal or tracking theories, not only is there a failure to offer a positive story of how we get to be related to reliably misrepresented properties, but features of the type of view itself block the possibility of offering such a story. §4.2 describes tracking theories of mental representation, while §4.3 argues that they cannot account for reliable misrepresentation. §4.4 argues that reliable misrepresentation is also problematic for non-tracking relation views.

4.2 Tracking theories of mental representation

The level of mercury in a thermometer correlates with temperature. It is natural to describe the thermometer as *representing* temperature. The thought behind **tracking theories of mental representation** is that we are just very sophisticated thermometers. We represent by having states that detect, carry information about, or otherwise correlate with states of the environment. For example, what it takes for the mental representation TIGER to represent tigers is for its activation, or use, to more or less correlate with the presence of tigers.

The correlation between mental states and states of the world might be the result of a **causal relation** between a representation and what it tracks. On such versions of the tracking theory, TIGER gets to represent *tiger* because tigers cause the activation of TIGER. Of course, this causal relation between tigers and TIGER might be mediated by other causal relations, for example, a causal relation between tiger stripes and certain states of the retina, but the relevant feature of this type of tracking view is that TIGER and tigers are linked by some causal chain. Proponents of causal tracking theories include Fred Dretske (1995) and Michael Tye (2000).

Our mental representations can also track things that do not cause their activations. Suppose we want to track F and F is correlated with G (perhaps F causes G, or some third thing H causes both F and G). Then another way to track F is by tracking G. Depending on what F and G are, it might be cheaper or easier to develop a tracking mechanism for one rather than another. For example, migratory birds track certain geographic locations themselves, but they do this by having states that are causally related not to the locations, but rather to magnetic fields that are correlated with those locations. And so, we might say that the birds' states non-causally track geographic locations. The basic, but vague, idea behind tracking theories

is that representations keep track of states or features of the world in a non-accidental way.

There are many tracking relations obtaining between mental states and the world. Relative to some tracking relations, certain mental states in migratory birds track magnetic fields, while relative to others, those same mental states track particular geographical locations. However, there is a determinate fact of the matter as to what mental states represent. This is obvious to introspection. We can tell that a state representing a particular shade of blue doesn't also represent *blue or green*, *cats*, or other things (see also Chapter 1).[5] And so, although a representation can bear various tracking relations to various things, the tracking theorist maintains that only one of those relations is the representation relation. Part of the debate surrounding tracking theories of mental representation, then, is over which tracking relation has this special status.

Many tracking relations are poor candidates for identification with the representation relation because their extension does not match what we take to be the extension of the representation relation. One manifestation of this problem is the **disjunction problem**. Take a simple tracking theory on which a representation represents everything that activates it. If very persuasive tiger statues as well as real tigers activates my tiger representation, then my concept TIGER has the disjunctive content *tiger or very persuasive tiger statue*, and that looks like the wrong answer. Sometimes the disjunction problem is put in terms of misrepresentation: It looks like on the simple tracking theory under consideration I can never misrepresent something that isn't a tiger as a tiger, since anything that activates my TIGER representation becomes another disjunct in the concept's disjunctive content.[6] What we want to say is that some of these activations of TIGER are appropriate, while others are inappropriate. It is only the appropriate activations that fix the content of TIGER. A tracking theory must then distinguish the appropriate from the inappropriate activations. In this way, avoidance of the disjunction problem provides one criterion that we can use to select among the vast number of tracking relations.

There are various options when it comes to distinguishing the appropriate from the inappropriate activations. **Teleological tracking theories** of mental representation take the appropriate activations to be those that occur in design conditions, where **design conditions** are the conditions in which the triggering of the representation helped our ancestors survive and reproduce. So if food triggered R in our ancestors and this helped them

[5]Of course, we don't occurrently think all that when we represent that shade of blue; these are the contents of possible further thoughts. What is obviously to introspection, rather, is the determinateness of the content that is represented, not all the contents that one is *not* representing. Thanks to Anthony Appiah for objections leading to this clarification.

[6]See Fodor (1987, Ch. 4) for the disjunction problem.

survive and reproduce, then R represents *food* (see e.g. Millikan (1989); Dretske (1995)).[7]

Optimal-functioning or well-functioning tracking theories of mental representation take the appropriate activations to be of the type that occur during conditions of optimal functioning or well-functioning, that is, the conditions in which the mental state in question now helps (or would help) its current bearer survive or flourish (see, e.g. Tye (2000)).

Another approach is Jerry Fodor's **asymmetric dependence** theory (Fodor, 1987). A representation represents whatever it is causally related to such that for anything else that it is causally related to, the latter causal connection is dependent on the former and the former causal connection is not dependent on the latter. Dependence is cashed out counterfactually: R represents *food* just in case food is causally related to R and for anything else, T, that is causally related to R, T would not trigger R unless food did, whereas food would continue to trigger R even if T did not.

Although different tracking theorists disagree on what the special content-determining conditions are, they agree that there are such conditions. Call the "right" conditions, whatever they are, the **ideal** conditions.

4.3 The problem for tracking theories

One thing all tracking theories have in common is that they deem it impossible to misrepresent in ideal conditions. This is because activations of a representation in ideal conditions *set* the content of that representation. This makes it very difficult for tracking theories to allow for reliable misrepresentation, since in cases of reliable representation, every actual activation of the representation in question is a misrepresentation. The only way for a tracking theory to allow for reliable misrepresentation is to maintain that ideal conditions do not ever actually obtain for the representation in question. To illustrate this general point, suppose again, as I will argue in Chapter 5, that color-representations reliably misrepresent. They track surface reflectance properties of objects but represent uninstantiated *sui generis* color properties, and color-representation evolved to be the way it is because (say) this is the most efficient way to discriminate and reidentify objects of the same luminance. On an optimal-functioning theory on which ideal conditions are the conditions in which a representation's activation helps its possessor survive or flourish, a representation can only reliably misrepresent if *none* of its actual activations help its possessor survive or flourish. If ideal conditions are design

[7]Fred Dretske (1995) holds a teleological tracking theory on which the tracking relation is a causal relation. Ruth Millikan (1989) holds a teleological tracking theory on which the tracking relation is not a causal relation. On her view, representations represent whatever conditions obtained in an organism's ancestors' environment that allowed for the use of representations to be helpful to survival. In the case of food, these conditions might be the nutritional content of food. So, on this view, R represents *good source of nutrients*, which has no causal effect on the representation FOOD.

conditions, then for a representation to reliably misrepresent *none* of its activations can have helped its possessor's ancestors survive and reproduce. But it seems possible to have cases of reliable misrepresentation in ideal conditions of this sort. If color-representations are triggered by surface reflectance properties that they do not represent in us, then it is quite plausible that this is useful and helps us survive and flourish, and that this connection between color-representations and surface reflectance properties was also useful to our ancestors. Perhaps it helped them reidentify objects over time or discriminate ripe from unripe fruit. Thus, it is implausible to maintain that the relevant sort of ideal conditions never obtain with respect to color vision, as would be required in order for color-representation to be a case of reliable misrepresentation on teleological tracking theories. Another way to put the problem is that on optimal functioning and teleological versions of the tracking theory, it is implausible to maintain that *our* color-experiences misrepresent while at the same time maintaining that they do, did, or would veridically represent in ideal conditions.

Let's turn to the asymmetric dependence theory. For this view to allow the case of color-experience to be one of reliable misrepresentation it must be the case that the surface-reflectance-property-to-color-representation connection is asymmetrically dependent on the color-to-color-representation connection. That means that in the nearest possible world in which the color-to-color-representation connection is broken, the surface-reflectance-property-to-color-representation connection is broken as well. But the nearest possible world in which the color-to-color-representation connection is broken is *our* world (since, by hypothesis, color properties are not instantiated in the relevant areas), and the surface-reflectance-property-to-color-representation connection remains. And so, the surface-reflectance-property-to-color-representation connection is *not* asymmetrically dependent on the color-to-color-representation connection, and the asymmetric dependence view cannot allow for reliable misrepresentation.

But this cannot be the way Fodor intends us to evaluate the relevant counterfactuals, since he claims that mental representations can bear the relevant kind of robust causal connections to properties that are not instantiated (Fodor, 1987, pp. 163–164). The idea behind the counterfactual analysis of asymmetric dependence is that the relation between a representation and what it represents is more robust than the relation between the representation and something other than what it represents. It does not matter whether one of those relationships is not in fact instantiated in the actual world. So, perhaps there is another way to evaluate the counterfactuals that can capture the idea that the surface-reflectance-property-to-color-representation is more robust than the color-to-color-representation connection. Unfortunately, the success of this kind of strategy seems unlikely. The problem is that reliable misrepresentation is *reliable*, and thus that the surface-reflectance-property-to-color-representation connection is *robust* (it's at least as robust as the horse-to-HORSE connection, which is supposed to be a paradigm example of a

robust connection). In cases of reliable misrepresentation, misrepresentation occurs as a matter of course, in normal, adaptive, or otherwise ideal circumstances. Misrepresentations are not isolated instances of falsity hijacking a robust truth-defining connection. Not only is the surface-reflectance-property-to-color-representation connection robust, but it is also not clear that any color-to-color-representation connection, say, holding in some other world, would likewise be robust. Our color-representations are hooked up with our other states and states of the environment such that they are good detectors of surface reflectance properties. It is doubtful that in an alternate possible world in which color properties are instantiated these very same color-representations would detect them instead or as well. In sum, the problem is that reliability usually implies a relatively robust causal relation, but on the asymmetric dependence theory, misrepresentation requires a relative failure of robustness. On the asymmetric dependence theory, you can't have both *reliability* and *misrepresentation*.

I've presented these arguments using the case of color, but the case is only meant to be an illustrative example. Most cases of reliable misrepresentation will be structurally like the case of color as I've described it. The connection between the representation in question and what it tracks will be robust and activations of the representation will occur in ideal circumstances. And that is enough to give rise to problems for tracking theories.

This concludes my argument for the claim that tracking theories of mental representation make it practically impossible to allow for reliable misrepresentation. I now turn to a diagnosis of the problem: On tracking theories, nonveridicality is always accompanied by a **non-semantic defect**, a defect apart from the mere fact that a representation is nonveridical. Non-semantic defects include being in less than optimal situations for recognition (e.g. the lights are off, or the animals to be recognized are overweight), the subject's actions being in some way maladaptive or not beneficial (e.g. performing lethal actions), and other deviations from ideal conditions. The idea is that when there is semantic error, something non-semantic has gone wrong as well (although not all cases of non-semantic defects result in misrepresentation). So we can define the veridical cases, and hence the content of a representation, by appeal as those cases not involving non-semantic defects. This makes tracking theories fairly adept at handling misrepresentation in cases that are non-semantically defective. They can deal with hallucinations as a result of drug use, the one-off misidentification of a crumbled paper bag in the dark as a cat, repeated misidentifications of fat horses as cows, and illusions that occur in circumstances that a representational system did not specifically evolve to handle. These are all cases where quite plausibly something non-semantic has gone wrong, and so, they can be correctly classified as misrepresentations by comparing their triggers to the triggers of the same representations in ideal conditions.

But with reliable misrepresentation, falsity is often the only defect. There are no non-semantic defects we can point to in order to classify the case

as one of misrepresentation. Rather, in cases of reliable misrepresentation, misrepresentation occurs as a matter of course, robustly, and in circumstances that are ideal by any of the criteria described above.

And so, tracking theories have great difficulty allowing for reliable misrepresentation, and this is related to their reliance on a relationship between non-semantic and semantic defects. Of course, I have not yet argued that there are any cases of reliable misrepresentation. So why is it such a big deal that tracking theories are ill-suited to account for them? The reason it is a big deal is that whether or not there are such cases, it would be inappropriate to conclude that there aren't such cases *on the basis of a metaphysical theory of mental representation*. By a **metaphysical theory of mental representation**, I mean a theory that aims to tell us what mental representation *really is*, as opposed to a theory that offers us a model of mental representation or various facts about it, its instances, how it behaves, etc. A metaphysical theory of mental representation should not rule out the mere possibility of reliable misrepresentation. Given the properties we observe ourselves representing (colors, shapes, cathood), our metaphysical theory should leave it an open empirical question whether any of them are reliably misrepresented. In any case, if none of this moves you, the next chapter argues that there are actual cases of reliable misrepresentation.[8]

4.4 The problem for non-tracking relation views

According to relation views, representation is fundamentally a *relation*. But relations have relata. If reliable misrepresentation involves the absence of the relevant relata, then relation views cannot allow for reliable misrepresentation.

There is a distinction between two types of relation view that is relevant for our purposes: There are relation views that claim that the representation relation relates mental states to particular entities, such as objects, property instances, events, or states of affairs; call these relation views **particularist**. There are also relation views that claim representation is a relation between mental states and general properties, which might be abstracta or universals of some sort; call these relation views **non-particularist**. There are also hybrid views that claim we are sometimes related to particulars and sometimes to general properties, perhaps even in the same experience,[9] so we can also speak of relation views that are particularist or non-particularist about some contents. Particularist and non-particularist views about the content

[8] Holman (2002) argues in the opposite direction: Tracking theories have trouble accounting for reliable misrepresentation about color, so we should be color realists. While Holman and I mostly agree about how the various views in question hang together, I think he is arguing in the wrong direction. It is a defect of a theory of mental representation if it places constraints on the non-mental world that are as serious as color realism.

[9] Many versions of disjunctivism are hybrid views. In the veridical case, we are related to objects and property instances, but in the case of hallucination, we are related to abstract universals. See, e.g. Johnston (2004).

of reliable misrepresentations require separate treatment, since reliable misrepresentation of a property P implies that there are no instances of P that we are relevantly related to, not that there is no general property P that we are relevantly related to.

Let's start with a relation view that is particularist about colors. Suppose again that we reliably misrepresent colors. On particularist relation views, the representation of colors requires a relation to color instances. But the *mis*representation of colors requires the lack of such a relation to color instances. The problem is that you can't have both. In other words, if the relevant particulars do not exist, then it's not a case of representation. If the relevant particulars do exist, then it's not a case of *mis*representation. For similar reasons, particularist relation views face problems accounting for hallucination: If the relevant particulars do not exist, then it's not a case of representation. If the relevant particulars do exist, then it's not a case of hallucination.

On relation views that are non-particularist about colors, color-representations represent colors by being related to general color properties.[10] But this type of view finds itself having to answer the awkward question of how we get to be related to one uninstantiated property (or property that we not are relevantly related to) rather than another.

To see why this is an awkward question, compare the case of reliable misrepresentation to the case of hallucination and illusion. Take a disjunctivist relation view on which in veridical cases we are related to particular instances of properties, while in nonveridical cases we are related to abstract properties. On such a view, in veridical cases we are actually related to instances of the abstract properties that are the contents of our hallucinations and perhaps illusions. It's those relations to those instances that somehow make it the case that in the cases of hallucination and illusion we are related to one abstract property rather than another. How exactly being related to a particular instance of property P in the veridical case allows us to be related to the abstract property P in the hallucinatory or illusory case is an open question. The point is that there is some potential for offering a story here, since we have some kind of contact to instances of the general property in question in the "good" cases.

A non-disjunctivist view that holds that we're always related to abstract properties might make a similar move. We get to be related to one abstract property rather than another in virtue of our being related in one way or another to *instances* of those properties. For example, our squareness-representations get to represent the abstract property of squareness in virtue of being related to instances of squareness. That's why when we hallucinate a square, the content is *squareness*, not some other abstract property like *roundness*.

[10] See Adam Pautz (2007) for a non-tracking relation view of this sort. On Pautz's view, mental representation is a primitive or *sui generis* relation between mental states and abstract properties.

In the case of both strategies, more will have to be said in order to make the accounts convincing. But the point I want to make here is that the situation is worse when it comes to reliable misrepresentation, since the properties represented are uninstantiated (or at least not instantiated in the relevant vicinity). There are no color instances that we are appropriately related to, say in some "good" cases, that can somehow anchor our systematically mistaken color-representations to one abstract color property rather than another. That's because there are no veridical cases that can serve to differentially relate us to one abstract property rather than another, like there are in the cases of hallucination and illusion.

Again, the problem here is that reliable misrepresentation is *reliable*. As with tracking theories, the most plausible way for non-tracking relation views to account for misrepresentation is by treating it as a special exception to an otherwise veridical representational system. We get to be related to one content rather than another because at least sometimes we get to be related to or have some kind of contact with instances of the represented property. Just as on a tracking view, falsity is parasitic on truth in certain ideal conditions. But, as with the case of the tracking theory, this only works with hallucinations, misidentifications, and possibly illusions. The treatment doesn't extend to reliable misrepresentation.

Of course, with enough magic, we can make almost any non-contradictory theory work. We could posit some brute metaphysical semantic facts about the representation relation, such as that representation RED_{682} gets to be related by the representation relation to property red_{682}, and so on for each reliably misrepresented property. I don't think there is anything else that can be said against this sort of view, except that it is quite unattractive.

4.5 Conclusion

In this chapter, I have argued that relation views have difficulty allowing for reliable misrepresentation. I haven't yet argued that there are actual cases of reliable misrepresentation. That is what I will do in the next chapter. However, whether or not there are actual cases of reliable misrepresentation, the fact that tracking theories cannot allow for the mere *possibility* of reliable misrepresentation is a serious strike against them. Whether or not there are cases of reliable misrepresentation is an empirical question. It would be premature to rule out the possibility on the basis of a metaphysical theory of mental representation.

Chapter 5

Actual Cases of Reliable Misrepresentation

To argue that there are actual cases of reliable misrepresentation, I will argue that there are cases of representations that on any plausible tracking theory track one thing but represent another (mismatch cases) and, further, that the properties that they represent are not actually instantiated (and, thus, that **anti-realism** is true of them). The existence of mismatch cases already allows us to form an argument against tracking theories (§5.1.5), and from anti-realism, we can conclude that we are dealing with actual cases of reliable misrepresentation, and thus argue against all relation views. In short, in this chapter I will argue that the things required in order to account for the contents of our experiences on a relation view simply do not exist, or if they do exist, they require an implausible metaphysical story for us to get appropriately related to them.

The chapter is divided into two sections. In §5.1, I argue that there are mismatch cases. In §5.2, I argue that at least sometimes, we can argue from a mismatch case to anti-realism about the represented property.

5.1 Mismatch cases

Mismatch cases are cases of mental representations that track something other than what they represent. From the existence of mismatch cases, we can conclude that tracking is not representation, and hence that tracking theories of mental representation are false. §5.2 further argues that from some mismatch cases, we can at least sometimes argue for anti-realism about the represented property, and hence that the mismatch case in question is an instance of reliable misrepresentation.

5.1.1 How to find out what representations represent

To establish the existence of a mismatch case, we need to be able to tell both what a representation tracks, and what it represents. The latter task is more difficult. In this section, I describe two relatively theory-independent methods for telling what a representation represents.

The first is an inoffensive form of **introspection** that I have been appealing to all along in describing paradigm cases of mental representation. This type of introspection should not be confused with other types of introspection that have been misused in the past. For instance, it is by now relatively uncontroversial that we have very little reliable introspective access to our mental *processes*. We are bad at telling whether a certain state was caused by another state. This is demonstrated in countless studies. For example, subjects asked to memorize a list of word pairs including "Ocean-Moon" were more likely to prefer Tide to other brands of detergent, but they do not realize that this preference is at least partly caused by their exposure to "Ocean-Moon" (Nisbett and Wilson, 1977).

While introspection of mental *processes* is unreliable, certain kinds of introspection of mental *contents* does not share in this disrepute, and is assumed by a dominant experimental paradigm in perceptual psychology and cognitive neuroscience associated with Anne Treisman and Garry Gelade. In this paradigm, visual experience is thought to be the result of primitive visual features that are integrated. Vision represents features (or properties, in the most non-committal sense of "property") and these features are then integrated such that they are represented as applying to particular (apparent) objects in the visual scene. One assumption of this paradigm is that we can report on the presence or absence of a feature. It is often thought that if we are able to find a target defined by a certain visual feature very efficiently among a sea of distractor items, then that feature is primitive (Treisman and Gelade, 1980).The point for now is that it is normally accepted that subjects can accurately report on the presence or absence of visual features in their experience. So we can report that we are seeing green, or that we are seeing a "T." Another assumption of this paradigm is that we can report on how features are bound, or integrated. In other words, we can tell *which objects* certain features go with. When we view the following picture, we can tell that the yellowness visually appears to go with the "X" and not with the "T":

We can tell this experience apart from one where the blueness appears to go with the "X" and the yellowness appears to go with the "T":

This means that we can distinguish between different ways represented properties are bound. Many of my arguments will depend on these kinds of judgments about the presence or absence of a feature or which features are bound to which objects. Others will depend on assumptions I take to be similar in kind and thus similarly unobjectionable.[1,2]

The second way we can find out about the contents of our mental states is by examining the states they are inferentially, evidentially, or otherwise related to. This thought is related to the more general thought that contents are **psychologically involved**. They play a role in the mental economy. For instance, an experience with content c at least sometimes to produces a belief with content c (or that *there is a* c), absent defeaters. At the very least, there should be some overlap between the content of the experience and the content of such a resulting belief. So if we know the content of the beliefs produced by some experience, we can learn something about that experience's content. For example, if you visually perceive a red square in front of you, then you are likely to believe that there is a red square in front of you (unless you have a defeating belief, such as the belief that you are hallucinating, or the belief that there are no colors).

Of course, in many cases, the content of the experience will be just as easy to discern as the content of the beliefs it causes, if not easier. Still, it is useful to have both methods of discerning the content of mental states for cases where the content of an experience is controversial.

5.1.2 Three mismatch cases

This subsection presents three mismatch cases: hot- and cold-experiences, color-experiences, and heaviness-experiences.

Hotness and coldness

You take a tomato out of the refrigerator and it feels cold. What is the content of that experience? First, coldness-experiences represent coldness

[1]Thanks to Philipp Koralus for helpful discussion.

[2]It is important to emphasize that I am not claiming that we are infallible when it comes to reporting the contents of our experiences, nor that they can always be fully verbally described. For example, experiences induced by stimuli presented quickly can be difficult to accurately report upon, and it's difficult to express certain perceptual contents in words, e.g. the olfactory content of the experience of smelling a rose. All I need is that we are pretty good at telling what we are *not* representing in most normal conditions. There are also cases where we inaccurately remember our past experiences (e.g. cases of change blindness) or fail to notice stimuli in our field of vision (e.g. inattentional blindness). These are cases where we either fail to remember a recent experience or we are focusing our attention somewhere else and fail to respond to a stimulus. These phenomena will not affect my arguments, however. The introspective observations I will rely on can be obtained while the experience is being undergone and while the subject is paying full attention to the experience (or the content of the experience), which effectively rules out change blindness and inattentional blindness, respectively, as alternative explanations of the observations I describe.

as a feature of the putative cold object. When you touch the tomato, you represent the tomato itself as cold. The same goes for hotness. When you touch a steaming cup of tea, the cup itself feels hot. Hotness and coldness are represented as features of the putative hot or cold things alone.[3,4]

There are two general options when it comes to deciding what hot- and cold-experiences track: We might take what these experiences track to be some physical property of objects, or we might take it to be some response-dependent property, something having to do with the relations between putative hot and cold objects and us. However, I will argue that while these types of properties are good candidates for what hot- and cold-experiences track, they are poor candidates for what they represent.

The general problem with taking hot- and cold-experiences to represent the physical properties that they track is that these properties will be fairly gerrymandered and complex, despite the fact that there is a *prima facie* obvious simpler candidate for the tracked property, something like mean molecular energy. This is because hot- and cold-experiences do not track absolute temperatures. First, the same stimulus applied to different areas of the body can differ in the temperature-experiences it evokes. Cold water on one's hand is felt as less cold than water of the same temperature applied to one's stomach. Rather than tracking absolute temperatures, hot- and cold-representations track temperatures relative to the bodily location of the applied stimulus. This pattern of response makes sense: Some parts of the body are more sensitive to heat and cold damage than others, so it is useful and adaptive that those areas show a greater response to temperature changes.

Second, the firing rate of any particular thermoreceptor (the transducer of thermal stimuli) depends not only on the absolute temperature of a stimulus, but also on the temperature of the thermoreceptor prior to stimulation as well as the rate of temperature change. The result is that an abrupt application of a hot or cold stimulus will make the stimulus feel hotter or colder, respectively, than a less abrupt application. This is also a sensible pattern of response: Abrupt changes in bodily temperature can be more dangerous than slower changes of the same magnitude. So it is also useful and adaptive that abrupt changes trigger experiences of greater hotness or coldness. So far, we have seen that hot- and cold-experiences can be said to track a complex function of initial temperature, temperature change, absolute temperature, and bodily location. It is important to emphasize that this is how the temperature perception system functions in normal, useful, and adaptive circumstances. Therefore, it will not do to deem some of these responses inappropriate or abnormal in some way, and thus not relevant to determining what hot- and cold-experiences track. The result is that if we want to say that these

[3]If you spend enough time touching those objects, you will perceive yourself as hot or cold too. For my purposes, I can remain neutral on the contents of such experiences.

[4]Perhaps it is possible to represent coldness without representing a particular object as cold. This possibility does not affect my argument.

experiences track a physical property, it will have to be a gerrymandered, complex, and relational property, such as (*a*) below[5]:

(*a*) A relation *R* that holds between absolute mean kinetic energy, initial absolute mean kinetic energy of bodily location on which stimulus is applied, change in absolute mean kinetic energy, and bodily location

I claim that any such proposal does not capture what hot- and cold-experiences represent.[6] When I touch the tomato, the tomato itself feels cold. Coldness is represented as a property *of the tomato*. But the property specified in (*a*) is a property of me, the tomato, the body part I use to touch the tomato, and the state of that body part prior to touching it. This already sounds problematic, but we can precisify at least part of the problem as follows: As discussed earlier, we can tell what object a represented property is bound to. If an occurrent experience is representing a relational property, then the relevant property should be experienced as bound to more than one thing. But we experience the property of coldness as bound to only one thing: the tomato. So (*a*) cannot be what hot- and cold-experiences represent.

Further, the experience I get when I pick up the tomato never causes me to believe that the tomato has the property specified in (*a*), and so this proposal falters according to our second method of finding out the contents of mental states. I need to know a lot of science in order to infer that properties like those specified in (*a*) are instantiated. And so, such beliefs would not be the result of perceptual experiences alone.

Even if we set aside for a moment the highly relational nature of the property specified in (*a*), we are still left with some problems. It is not clear that a less relational physical property, such as mean kinetic energy or energy transfer from one object to another, fares any better as a candidate for what we're representing.

(*b*) Mean kinetic energy

(*c*) Energy transfer from one object to another

Again, although involving fewer relata, (*c*) is still relational, holding between the tomato and my hand or the tomato and its general surround,

[5]See Kathleen Akins (1996) for a thorough discussion of what hot- and cold-representations track. Akins concludes from her discussion that we should stop worrying about what the temperature representation system is representing and instead focus on what it is *doing*. Importantly, Akins thinks of representation as a kind of tracking, and so her conclusion is congenial to my own. It is not clear whether she would allow representation an interesting role in hot- and cold-related behavior if representation is understood in a different way.

[6]It is important to emphasize that the question I am addressing has to do with the content of hot- and *cold-experiences* not our the content of our *concepts* HOT and COLD or *the nature of hotness and coldness*, where "hotness" and "coldness" are taken to refer to the nearest natural kinds in the area, or the like. Thus, it is not an objection to my arguments that hotness and coldness could turn out to be any of the discussed options. This sort of confusion arises more often in the case of color; see fn. 13.

while coldness seems to be a property of the tomato alone. Another problem with both (*b*) and (*c*) is that they are too sophisticated. They involve contents that arguably we cannot entertain prior to a certain kind of education. Further, (*b*) and (*c*) both fail to make sense of the beliefs we are led to have on the basis of our experiences of hotness and coldness, just as (*a*) did: After touching the tomato, I am not tempted to form the belief that it has a particular mean kinetic energy or anything to do with energy transfer. If I know enough physics, I might *then* form such beliefs. But those beliefs would be the result of inference based partly on substantive theoretical assumptions, not a result of my perceptual experiences alone.

One might object that we *do* form beliefs about the properties in (*a*), (*b*), or (*c*) on the basis of perceptual experience alone. This is because the content of our temperature-related beliefs inherit their content partly from the content of our temperature-related experiences. While this type of response is open to the tracking theorist, this claim about the content of hot- and cold-beliefs might strike some as just as implausible as the corresponding claim about the content of hot- and cold-experiences. It would be really great if we could gain all this theoretical knowledge just by touching the food in our fridges, but it just doesn't work that way.[7]

There is some indirect support from the history of science for the claim that hot- and cold-beliefs do not have anything like mean molecular energy, energy transfer, or the like as their contents. Aristotle thought that hotness and coldness were distinct and primitive properties. He did not believe that hotness and coldness were a matter of mean molecular energy, or some such thing. If the content of his experiences was something like that expressed by (*a*), (*b*) or (*c*), then why didn't he take his experiences at face value and include mean kinetic energy and the like in his physics? As late as the 19th Century, physicists believed that heat and cold were their own distinct substances: caloric and frigoric.[8] Caloric and frigoric were thought to flow from one object to another, thereby making it hot or cold, respectively. One possible (partial) explanation for the emergence of these theories is that hot- and cold-experiences represent things as being hot or cold *simpliciter*, not as having complicated physical properties. Absent better theories, these scientists were taking their experiences in this area at more or less face value. The existence of these theories very strongly suggests what was already

[7] One might suggest that we do form such beliefs, but under a particular mode of presentation that does not allow us to recognize the content of the beliefs we form. In §5.1.6, I address this type of objection. Briefly, if we invoke modes of presentation to account for what we find upon introspection, then we have to offer an account of modes of presentation. Modes of presentation are themselves representational, so the tracking theorist must account for their content in terms of tracking. But then we are back at the original problem of finding a plausibly tracked property that accounts for these represented contents. Further, on such a view, the modes of presentations seem to be doing all the work accounting for these observed phenomena, and the contents themselves end up being isolated and introspectively inaccessible spinning wheels.

[8] For an overview, see Mendoza (1961).

obvious: We do not form beliefs about mean molecular energy on the basis of hot- and cold-experiences. And that strongly suggests that hot- and cold-experiences do not represent the physical properties specified in (a), (b), and (c).

Perhaps one might suggest that experiences and beliefs are not as strongly related as I claim. Perhaps our beliefs about refrigerated tomatoes represent one property, say cold_{belief}, while our experiences about refrigerated tomatoes represent a completely different property, $\text{cold}_{experiences}$. The main worry with such a strategy is that it threatens to make the content of cold-experiences explanatorily inert. The property represented by cold-experiences on this view ($\text{cold}_{experience}$) is not discoverable through introspection, it does not explain our felt experience of coldness, and it does not hook up in any particularly interesting way with other mental states, such as beliefs about cold_{belief} objects. We could vary the content of cold-experiences without affecting anything else in the cognitive economy. In that sense, the content of cold-experiences is isolated: it does not interact with anything else. Although there might be isolated contentful states like this, I see no reason to think we're dealing with one here.

So far, I have argued that (a), (b), and (c) aren't the representational contents of hot- and cold-experiences. Here are some more things these experiences track that they might be said to represent:

(d) Hot or cold qualia or sensations

(e) The disposition to cause hot/cold qualia or sensations in me (or my community, an ideal observer, or whatever), or the manifestation of such a disposition

(f) Being of a good or bad temperature for me (or my community, an ideal observer, or whatever)

These also fail as candidates for what hot- and cold-experiences represent. The notions invoked aren't quite as sophisticated as those in the first three options, and presumably, we do believe that hot things cause our hot sensations, but these options also fail to match the experiences' representational contents. For one, (d) is phenomenologically inaccurate. When I touch the tea-cup, *the cup itself* feels hot. I represent hotness as a property of the cup, not as a property of my qualia or sensations, nor as a type of qualia or sensation. There is something it is like to represent things as hot, and I might call that a "quale" or "sensation," but that does not mean that I am not representing the cup itself as being hot.

(e) and (f) also appear to be phenomenologically inaccurate. They are just too sophisticated. Perhaps they are part of the contents of our temperature-related concepts, especially if we've studied science or philosophy, but it is implausible that these are the contents of our hot- and cold-*experiences*.

Perhaps more clearly, (e) and (f) are too relational. They involve me, or my community, or something similar, and (e) additionally involves a relation to qualia. But, again, the coldness of the tomato seems to be a property of the tomato alone. It is not bound to anything or anyone else. As far as my temperature-related experiences are concerned, the tomato could be the only thing in the world and it could still be cold.[9]

(f) deserves special attention. Taking this type of property to be the type of property represented by hot- and cold-experiences can be motivated by considering (a) again. Our thermoreceptors respond in what seems to be an advantageous way. This is because different parts of our bodies are more or less susceptible to heat or cold damage, so it is a good thing that our receptors respond differently to hot or cold stimuli applied to different areas. Abrupt temperature change is bad for us, so it is also good that faster rates of temperature change evoke a greater response. Thus, we might be tempted to say that our temperature-related experiences track something having to do with how good or bad a stimulus of a different absolute temperature than various of our surfaces is for us to interact with.

First, it is far from clear that this is what these experiences are tracking; whether a given temperature stimulus is good or bad might depend on other factors that hot- and cold-experiences do not potentially track. But let's set this aside. There are two types of properties mentioned in (f) that raise problems: *goodness/badness* and *temperature*. Temperature is something having to do with mean kinetic energy or energy transfer. For the reasons mentioned earlier, invoking temperature poses problems independent of any other problems proper to the properties in which it is embedded. Goodness and badness each pose similar problems. Our hot- and cold-experiences do not track some sort of absolute or objective goodness. Rather, if they track anything at all in the area, they track contribution to current survival, or contribution to evolutionary fitness, or maybe even something having to do with goal-accomplishment. So then goodness and badness in (f) are going to have to be further cashed out in terms of some other properties, perhaps evolutionary fitness, current survival or goal-accomplishment. The end result will look something like this:

> (g) Having mean molecular energy that is (or would be, or has been, etc.) survival-supporting (or evolutionarily adaptive, or helpful for goal-accomplishment) for me (or my community, my species, etc.)

Apart from being too relational and inheriting the problems of any account invoking mean kinetic energy, the end result is too complex and sophisticated to be the content of temperature-related experiences. When I'm at the

[9] I do not mean to imply that our perceptual experience tells us that the tomato could be the only thing in the world and still be cold. Rather, I am claiming that it is *compatible* with my experience that the tomato is the only thing in the world and it is still cold. My experience does not rule out that possibility.

grocery story touching boxes of ice cream to make sure the ice cream I'm about to buy has not already melted, I'm not representing any of that.

None of the natural candidates for what temperature-related experiences track are suitable candidates for what those experiences represent. So, what do hot- and cold-experiences represent? I suggest we take these experiences at face value and say that they represent *sui generis* properties of hotness and coldness, and intermediate properties. In any case, we can conclude that perceptual hot- and cold-representations are mismatch cases: they represent one thing and track another.

Color

In the previous subsection, I argued that hot- and cold-experiences track something other than what they represent. A similar line of argument can be run in the case of color.[10,11]

What might color-experiences be said to track? Again, we have two general options: some (probably physical) property of objects, or something having to do in some way or other with us. Here's an option of the first sort:

(h) Surface reflectance properties

A surface reflectance property might end up being one of the following:

(i) The disposition to reflect, transmit, or emit such-and-such proportions of such-and-such wavelengths of incident light[12]

(j) The categorical basis underwriting the disposition to reflect, transmit, or emit such-and-such proportions of such-and-such wavelengths of incident light

(i) is dispositional and perhaps even relational, including objects in the surround. If (i) ends up being relational, we face the same problem we faced with the relational candidates for the content of hot- and cold-experiences: Relational properties are properties belonging to, involving, or bound to more than one object, but color properties appear to bind to only the putatively

[10]The contents of color-experiences can vary along three dimensions: *hue*, *saturation*, and *brightness*. This might lead one to ask whether colors are complex properties composed of hue, saturation, and brightness. For my purposes, I can bypass this issue. Regardless of how you come down on it, there will be no plausible property that color-experiences track with which to identify colors and so we have another mismatch case.

[11]I don't take the arguments and observations in this section to be particularly novel. There is a huge literature on color, and versions of many of the arguments I offer have been much discussed. However, my focus is slightly different from the focus generally taken in this literature: rather than worrying about *what colors are*, I am worrying about *what color-experiences represent*. Given the extra assumption that what colors are is determined by what color-experiences represent, the two questions are equivalent. However, this assumption is sometimes part of what's at issue in the debate I'm concerned with.

[12]Tye (2000) has a view like this.

colored object (even though some of the alleged extra relata are on the visual scene!). If color-experiences represented relational properties involving those relata, then colors should appear to be bound to all those relata. But they do not. Blue-experiences seem to bind only to the apparently blue object.

The dispositional aspect of the suggested content is also potentially problematic. I'm not sure what it would even be like to see a disposition. The closest thing I am able to come up with is the experience evoked by Figure 5.1. The ball kind of looks disposed to fall off the pointy surface. (Still, I am more inclined to describe my experience as representing the ball as being *about* to fall.) In any case, seeing colors is nothing like this. When objects look colored, they don't look like they are disposed to do anything.[13]

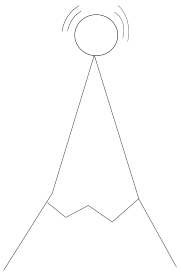

Figure 5.1: Something about to fall.

(j) avoids including dispositions in the content of color-experiences, but it does so at the cost of making color properties highly disjunctive. The relevant dispositions can be realized in many very different ways on different surfaces (compare the blue of the sea, the blue of a couch, and the blue of household cleaning fluids). What unifies many of these realizations of blue is not some microphysical similarity, but rather the disposition to reflect, transmit, or emit light in a certain way. That's why (i) does not end up being hideously disjunctive, but (j) does.

A problem common to both (i) and (j) is that they are too complex and sophisticated. A representation of a wavelength is sophisticated, as is a representation of proportions of various wavelengths. If our visual experiences represent such contents, we are oblivious to it.[14] The proposals also fail our other test. We are never tempted to form a belief about an object's surface reflectance properties based solely on color-experiences. I don't even know which surface reflectance properties correspond to which colors, and I am

[13]Some people tell me it's not so crazy to think that colors might end up being dispositions. What they have in mind is a scenario in which the referents of color-concepts end up being dispositions. We pick colors out in some perhaps indirect way, and it turns out that the properties we pick out happen to be dispositions. They are, therefore, missing the point by taking the debate to be one about the nature of colors, not about what color-experiences represent (see fn. 11).

[14]Various moves can be made involving modes of presentation and the like. I will get to those in §5.1.6.

willing to bet that you don't either. It takes serious empirical work to find out. We certainly can't discover it just by scrutinizing our color-experiences.

The second general type of approach is to say that color-experiences represent something that they track having to do with us. Here are some options:

(k) Color qualia or color sensations

(l) The disposition to cause color qualia or color sensations in me (or my community, or an ideal observer, or whatever)

Note that (k) and (l) involve positing non-representational color-related features of experiences. Colors, then, are taken to be either the qualia themselves or objects' dispositions to cause us to have those qualia.[15] First, it is not clear that there are such non-representational qualia in the case of color. When we view something yellow, there does not seem to be *both* a representation of yellowness *and* a yellowish "raw feel." The yellowness of the experience seems to be exhausted by the represented yellowness. Put otherwise, all the grayness in Figure 7.1 on p. 82 seems to be on the elephant. There isn't a represented property of the elephant as gray *and* a gray quale that doesn't represent anything.[16]

(k) is also problematic for the same reason the analogous proposal in the case of hot- and cold-experiences was problematic: it makes the wrong predictions about which objects are represented as having colors, that is, which objects color features are bound to. In the experience induced by the elephant in Figure 7.1, your gray-experience represents grayness as a property of the elephant, not as a property of you or your mind. So even if color qualia exist, these aren't the contents of color-experiences.

(l) also has all the same problems as the analogous proposal in the case of heat: it is too sophisticated. And, more clearly, (l) is relational. It involves me, my community, or some other observer. But colors seem to be properties of the allegedly colored objects, not of those objects *and* me (or some other subjects). As far as my color-experiences are concerned, the gray elephant could be the only thing in the world. The grayness is bound to the elephant, not to me and the elephant.[17]

[15] We cannot replace "color qualia" with "color representation" because then the content of color-experiences would be circular (Johnston, 1992).

[16] This is the basic intuition behind the what is sometimes called the "argument from transparency" for intentionalism (see Harman (1990), Tye (1995, 2000), and Jackson (2004, 2005)). For discussion, see §7.2.1.

[17] Similar arguments are made in Boghossian and Velleman (1989, 1991). But Boghossian and Velleman are interested in the question of what colors are, while I am interested in the question of what color-experiences represent. If we assume that color-experiences represent colors, the two discussions can be conducted in parallel. But we would be ruling out one possibility that makes certain proposals more plausible as candidates for what colors are than for what color-experiences represent: The possibility we'd be ruling out is that we represent colors without representing all their properties. On this possibility, colors could have features that we don't represent. See also fn. 11 and fn. 13.

So, regardless of whether you think the contents of color-experiences can be reduced to complexes of hue, saturation, and brightness, they cannot be further reduced to anything that those experiences can be plausibly said to track. Again, perhaps the best thing we can say about what color-experiences represent is that they represent *sui generis* color properties. In any case, we can conclude that color-experience provides us with another mismatch case.

Heaviness

We have a certain kind of representational experience when we lift heavy objects, such as large dictionaries. We can ask what heaviness-experiences track that they might also plausibly be said to represent. Here are some options:

(m) Weight

(n) Mass

But these are not plausible candidates for the representational contents of heaviness-experiences. (m) and (n) are too complex and sophisticated. They involve contents such as *mass* and *gravitational field*, which require some scientific education in order to represent. That objects have these properties is a fact not inferable from our heaviness-experiences alone. A further problem with (m) is that weight is a relational property holding between a putatively heavy object and other objects. But the heaviness we represent is bound only to the putatively heavy object. So heaviness isn't weight. And finally, (m) and (n) fail the belief test. I am not disposed to form beliefs about weight and mass on the basis of heaviness-experiences alone.

Perhaps heaviness-experiences represent something to do with the experiencing subject.

(o) Heaviness qualia

(p) The disposition to cause heaviness qualia in me (or my community, or an ideal observer, or whatever)

(q) Being hard to lift for me (or my community, or some ideal lifter, or whatever)

Let's start with (o). As in the case of color, it is not clear that there are the requisite qualia. And if there are heaviness qualia, they do not seem to be what we're representing when we experience things as heavy: Heaviness is represented as a property of the putatively heavy object, and not as a property of us, our minds, or our mental states.

(p) has all the same problems as other forms of dispositionalism. First, there don't seem to be the required heaviness qualia. Second, it does not

seem that heaviness-experiences represent dispositions, or even what that would be like. And third, (p) is a property of the putatively heavy object, the heaviness qualia, and me (or some other subjects), but heaviness-experiences represent heaviness as a property of the putatively heavy object alone. Put otherwise, heaviness is bound only to the putatively heavy object, but the property in (p) is bound to additional objects as well.

What about the property of being hard to lift? It's true that heavy things are often judged to be hard to lift, and things that are hard to lift are often judged to be heavy. But we also think things are hard to lift *because* they're heavy, and this tells against the content of heaviness-experiences just being the property of being hard to lift. Also, being hard to lift is a relational property holding between an object and a lifter (or an ideal lifter, or whatever), but our heaviness-experiences represent heaviness as a property of the putatively heavy objects alone.

Thus, I claim that heaviness is another mismatch case. Again, perhaps we need to invoke *sui generis* represented properties to account for the content of heaviness-experiences. In any case, heaviness-experiences represent something other than what they track.

5.1.3 Other perceptual mismatch cases

I've offered three examples of mismatch cases. All I really need for my argument for the conclusion that there are mismatch cases is for one of them to be convincing. By now, the form of argument should be familiar and it should be clear how to come up with additional mismatch cases. For instance, arguably, pain is also a mismatch case. Pain-experiences seem to represent *pains* in parts of one's body. But they track bodily damage, sharp objects applied to the skin, stimuli disposed to cause pain-experiences, or some such. If certain kinds of referred pain[18] are adaptive, useful, and occur in normal conditions (e.g. pain in the left arm during a heart attack), then we might have an additional mismatch between what pain-experiences represent and what they track: Referred pains might represent one bodily location but track another (see §7.3.1 for more on pain).

Similarly, one might argue that hardness-experiences represent hardness but track resistance to deflection, a particular physical structure, or some such thing. Sweet-experiences represent sweetness but track a particular chemical structure. Goodness-experiences represent goodness but track goodness-relative-to-a-goal, societal norms, or adaptive altruistic action types. Beauty-experiences represent beauty but track youth, fertility, beauty qualia, or something response-dependent. And the list goes on.

[18]Referred pain is pain that is felt as being located in a part of the body than the damage that is causing the pain-experience.

5.1.4 Non-perceptual mismatch cases

So far, I've focused on perceptual mismatch cases. This is because it is relatively easier to discover their contents than it is to discover the contents of thoughts and other such states. In this section, I describe some admittedly more controversial non-perceptual mismatch cases. Remember, I only need one such case to make my point that tracking is not the same thing as representation.

Conceptual mismatch cases related to perceptual mismatch cases

For every perceptual mismatch case, we might also find a conceptual mismatch case lurking in the vicinity. The content of the concept of blueness is undoubtedly related to the content of the perceptual representation of blueness. When we judge things to be blue, perhaps we judge them to have the same property that we visually perceive them to have when we perceive them as blue. Of course, perceptual experiences represent more specific or fine-grained colors. For example, a visual experience might represent a car as being blue_{421}, but normally a judgment will only represent it as being light blue, where blue_{421}, blue_{422}, and many other shades of blue count as light blue. How our concepts get to represent ranges of colors in this way is an open question. But for now, all I need is that there are such concepts that represent ranges of color in one way or another. Then these concepts track something other than what they represent: they represent ranges of colors but they track ranges of surface reflectance properties, ranges of dispositional properties, or some such.

We can say something similar about some of the other cases. When we judge something to be hot, perhaps we judge it to have the same type of property that we would experience upon touching it. Or perhaps we judge it as having some property in a range of such properties. Either way, the concepts track something other than what they represent.

Some additional conceptual mismatch cases

There might be more conceptual mismatch cases that we can approach independently of their perceptual analogues, if they have any. Moral concepts are good candidates. If J. L. Mackie is right, our moral concepts represent properties that are prescriptive, in that they offer subjects reasons to act in certain ways, and objective, in that they exist out there, in the world, independently of subjects and their preferences (Mackie, 1977). But our moral concepts arguably don't track any objective and prescriptive properties. Rather, they track harm, disrespect, unfairness, societal norms, personal aversions, and the like.[19] If that's right, then moral concepts are another mismatch case.

[19]Disrespect and fairness might themselves be mismatch cases. If so, we can replace their occurrence on the list with whatever it is that DISRESPECT and FAIRNESS track.

A Williamsonian concept of knowledge might also classify as a mismatch case. Timothy Williamson takes knowledge to be a basic or fundamental epistemic state out of which other epistemic states, such as belief and justified belief, are defined (Williamson, 2000). Perhaps our knowledge-judgments track true, justified beliefs that meet some gruesomely complicated Gettier-case fix requirement, but our knowledge-judgments do not represent all that. If that's right, then the concept KNOWLEDGE tracks something other than what it represents, and we have another mismatch case.[20]

Or take our concept of a self. Debates about personal identity concern questions like that of whether or not you are the same person who began reading this chapter some time ago, and if so, what makes this the case. One tempting view is that your self consists in a simple enduring substance (for a recent defense, see Lowe (1996)). Perhaps, then, we represent our friends and acquaintances as simple self substances. But, assuming there are no simple self substances, all we track are their faces, gaits, Parfittian relations of psychological continuity, or some such thing. In such a case, our concept of the self tracks something other than what it represent and we have a mismatch case.

Here's a useful heuristic in searching for mismatch cases: Look for concepts that resist reduction. The reason they resist reduction might just be that they represent something that doesn't match the nearby (probably physical) properties that they track. Other possible cases include responsibility, justice, agency, and numbers.

Revolutionary reconstruction

These examples are controversial. There are many reductionists about knowledge, moral properties, and the self. In this section, I want to offer an admittedly somewhat speculative explanation for why they are so controversial in a way that perceptual experiences are not (or at least should not be). I suggest that these examples are controversial because we have a tendency to reconstruct our concepts. When we notice that what we represent does not match what we track, we reconstruct our old concept such that it represents something closer to what we track. (We need not do this consciously. This might be something that just happens automatically.) For example, if we come to think that there is no simple enduring self that fixes the facts about personal identity over time, but that what we manage to track are Parfittian relations of psychological connectedness and continuity (see Parfit (1984)), we might reconstruct our concept of a self or a person such that things bearing these relations count as enduring persons or selves.[21]

[20] See Chapter 13 for more on Williamson's concept of knowledge.

[21] We can read Mark Johnston's (2009) *Surviving Death* in a similar way. After ruling out various accounts of what determines personal identity over time, Johnston offers a response-dependent account. I do not think he is claiming that this was our concept of the self all along. Rather, he seems to be saying that it is the nearest thing in the area

Why don't we just abandon the old concept and create a new one when this happens?[22] Concepts are connected in various ways to other concepts and states, and preserving these connections is useful. Our concept of a self or a person figures in many beliefs and norms we represent: persons have bodies; if you borrow something from a person, you should return it to the same person; if you say something to a person, he might tell someone else; and so on. If we lose the old concept and replace it with a new concept, we'll have to forge all those conceptual connections again. It's more efficient, therefore, to simply reconstruct the old concept, while retaining its old connections.

Similarly, when we learn that what causes our thermometers to react the way that they do is mean kinetic energy, our concept of hotness tends to be reconstructed so that it has a content such as *high mean kinetic energy*. It may or may not in addition continue to also inherit some of its content from the perceptual experience of hotness. That depends on how we reconstruct. However, this new concept figures in most of the same attitudes that the old concept figured in. We still believe that it's a bad idea to touch a hot stove, that cold showers increase alertness, that Canada is colder than Mexico, and so on. Likewise, our reconstructed color concepts, our reconstructed heaviness concepts, and our reconstructed moral concepts figure in much the same attitudes as the old concepts they replace.

In contrast to the case of concepts, it is much harder to change the content of your perceptual experiences. Color-experiences will continue representing *sui generis* color-like features of objects even if you've reconstructed your color concepts.

There is some debate about whether the contents of perception include more "conceptual" contents like *cat* and *TV*.[23] I do not want to enter this debate now. I just want to note that if you think that vision can represent the likes of TVs and cats, then you might also think that vision can come to represent surface reflectance properties by means of revolutionary reconstruction in much the same way as our concept of self can come to represent Parfittian relations. But even if this is so, the visual representation of color does not go away.

Again, I recognize that this story about concept revolutionary reconstruction is speculative and underdeveloped, but I will have more to say about it in Part III, where I develop a theory of concepts that explains how such revolutionary reconstruction is possible and why it's so easy. For now, the point is that if revolutionary reconstruction is the source of our disagreement

---

and thus that we should reconstruct our concepts so as to bring them in line with his response-dependent account.

[22]Here, I am thinking of concepts as individuated by their vehicles, that is, by the representations themselves. If we individuate concepts by their contents, then changing a concept's content automatically results in a new concept. It is possible to rephrase this part so as to avoid commitment to vehicles of representation, but this would require machinery introduced only Chapter 10.

[23]See, e.g. Siegel (2005), who argues that we can perceptually represent the likes of TVs and cats.

over the content of concepts such as MORAL GOODNESS, then this is bad news for the tracking theorists. Some concepts might have different contents for some of us, but all that I need is that *some* people have concepts with the contents I've described. It is enough that J. L. Mackie's concept of moral goodness was one of objective prescriptivity, that Williamson's concept of knowledge is of a fundamental epistemic state, or that some child's concept of hotness is one with the same content as her hot-experiences.

5.1.5 Another argument against tracking theories

If the claims in Chapter 4 are right, then the existence of actual cases of reliable misrepresentation is a huge problem for relation views. My aim in this chapter is to argue that there are such cases. However, it is possible to argue against tracking theories of mental representation, a type of relation view, from weaker premises. The very existence of mismatch cases is already a problem for tracking views. If tracking is representation, then tracking and representation can't come apart. But if there are mismatch cases, then they do come apart. So tracking isn't representation.

5.1.6 Responses

I have argued that there are mismatch cases in which a representation represents something it doesn't track. I will now discuss various possible responses. For ease of exposition, I will focus mainly on the case of color, but all the same objections and responses can be made regarding any mismatch case.

Deny introspective access

One might respond to my arguments with "I guess we don't have reliable introspective access to the contents of our mental states after all, at least not in your cases."

This move doesn't come without a cost. It commits the objector to an unpalatable kind of error theory, one on which we are systematically mistaken with respect to the contents of some of our most salient mental states. When I perceive something to be blue, I sometimes judge that I am perceiving something to be blue, where blueness is an intrinsic, categorical property belonging to the putative blue object alone. In contrast, I am never tempted to judge that I am having an experience of a certain reflectance property (I don't even know what the reflectance properties corresponding to various colors are). And so my introspective judgments about the contents of my color-experiences would be deeply and systematically mistaken. That is, I believe my experience has a certain content, but I am wrong. I am not having an experience representing an intrinsic, categorical property. Instead, I am having an experience whose content I cannot even find out from my current

position. (And prior to doing all this science and philosophy, I didn't even know that I couldn't find it out.)

Accuse me of an error theory

In response to my response above, someone might complain that I am also committed to an error theory. Suppose I am right about the content of color-experience, and if, as I will soon argue in §5.2, color properties are never instantiated in our world. Then our color-experiences would be nonveridical. It looks like I'm committed to an error theory with respect to color-experience. Further, if thoughts, beliefs, and judgments that something has a particular color have or include the same content that the corresponding color-experiences have, then it looks like I'm committed to an error theory with respect to those states as well. So now we have two competing error theories: one with respect to the color-experiences and judgments about colors, and one with respect to judgments about color-experiences. Which one's worse?

An error theory is not automatically bad. It is only bad to the extent to which we have independent reason to think that we are not in error. My error theory claims that a certain class of our judgments and experiences *about the external world* are systematically mistaken. The type of error theory that the tracking theorist ends up endorsing has it that we are systematically mistaken with respect to a certain class of judgments *about the content of our own experiences*. I claim her error theory is worse. We have better reason to think that our judgments about the contents of our experiences are not systematically mistaken than we do to think that our judgments about the external world are not systematically mistaken. An appearance-reality gap is more plausible than an appearance-appearance gap.[24]

Qualia

One might suggest that perceptual experiences involve non-representational raw feels, or qualia, and that it is the raw feels of these experiences that are driving our judgements in my examples, not their representational contents.

My first response to this objection borrows from the intentionalist intuition that it's not so obvious that these experiences have non-representational raw feels. It seems that their introspectively accessible features are exhausted by their representational features. Remove the represented yellowness and you thereby remove the yellow qualia. Another way to put the point is

[24]One might object that we are sometimes in error about our experiences, as in cases of change blindness and inattentional blindness, so an error theory with respect to some class of judgments about our experiences is not so implausible. However, the kind of error theory our tracking theorist is endorsing is much more extreme. She must maintain that we are in principle unable to *ever* correctly introspect on our color-experiences, not even while the experience is being undergone and one is paying full attention to it. We can only find out their contents through serious empirical work.

this: once you have an experience with yellow qualia, you are automatically representing yellowness. This suggests that there are no non-representational qualia associated with color-experience.

But even if there are raw feels, this response doesn't really avoid the problem. If raw feels account for the introspective observations that I describe, then what role does representational content play? On this view, content must be something "hidden," something not introspectively accessible. (Otherwise, we could introspect on it too and I could run my argument on those introspective results.) There are two problems with saying that these representational contents are "hidden." The first is that they start looking like explanatorily useless idly spinning wheels. They are introspectively inaccessible. They play no role specific to their content in further cognition; whatever roles they do play in generating further thoughts could have just as easily been played by states with different representational contents, or no contents at all.

The second problem is that even if there are such "hidden" contents, there additionally seem to be some representational contents that *are* introspectively accessible. Here's why I think that: If there were no such contents, then all we would find upon introspection would be raw feels. But that is not what we find. When we introspect, we don't just find a crazy mess of random raw feels. Instead, we find representations of objects, shapes, and other ways things might be. This is just another way of putting the intentionalist's transparency intuition, but it is worth emphasizing again in this context. The point is that even if there are "hidden" contents that fail to give rise to mismatch cases, experiences also evidently have a type of representational content that is introspectively accessible, and this content does seem to give rise to mismatch cases.

Modes of presentation

The tracking theorist might claim that representations have both contents and modes of presentation, and that it's the modes of presentation that are driving our intuitions about content in putative mismatch cases. In the case of color, we are representing (say) reflectance properties under a "color" mode of presentation.

What are modes of presentation supposed to be? In most cases in which we think modes of presentation are involved, objects have (putative) properties corresponding to their modes of presentation. Venus, for example, has the following two properties that we can use to pick it out: being the first star visible in the evening sky and being the last star visible in the morning sky. These two properties correspond to Venus' two modes of presentation as the evening star and as the morning star. But the problem is that the contents of color-experiences don't seem to involve similar extra properties: Reflectance properties don't have colors, and neither do dispositional properties or the other candidates for what color-experiences track. Further, even if we did

represent surface reflectance properties under a color mode of presentation, we could run all the same arguments on this represented content instead. We don't gain anything, and in fact, the situation becomes worse because now we're additionally committed to introspectively inaccessible, or "hidden," representational contents such as reflectance properties that don't do any extra explanatory work. The modes of presentation are doing all the work in explaining the phenomena we seek to explain. They are what we find when we introspect, and they are what inform our color-related beliefs and judgments. And if content is causally potent in generating behavior, they are most likely what does that too. The "hidden" contents have dropped completely out of the picture.

Silence on certain features of the represented properties

One might suggest that color-representations are **silent** on whether the represented properties are intrinsic, relational, primitive, or complex. They don't say anything one way or the other. This way, it could turn out that our color-representations represent something surprising, something complex and highly relational. The same goes for hot-and cold-representations, heaviness-representations, attractiveness-representations and perhaps other types of representations as well.[25,26]

This does not avoid the problem, however. On the proposals we considered, what we track are determinately complex or relational properties. In other words, what we track is not silent about things like complexity, but rather it is **vocal** about such things, and it's hard to imagine what it would take to track something that is neither determinately relational nor determinately non-relational. So even if the representational contents in question were silent on those features as the objection would require, we would still have a mismatch case: We track something vocal on various features but represent something silent on those same features.

Perhaps one might suggest that in the first instance, we track vocal properties, but *in virtue of* tracking these vocal properties we thereby also track silent properties. In order for this kind of strategy to work, the relevant tracking relation must isolate the silent property and not the vocal property as the relevant tracked property. If the vocal property is what is in the first instance causally related to the representation in question, what is useful for the organism's survival, and is in general more robustly connected to the representation in question than is the silent property, this strategy does not work. And so, the burden of this kind of strategy is to show that the favored tracking relation picks out the silent property and not the vocal one, and

[25]Thanks to Gilbert Harman for suggesting this objection. See also fn. 13 and fn. 17. Thanks also to Frank Jackson and Bill Fish for discussion of this objection.

[26]This is to deny what Mark Johnston calls "Revelation" in his "How to Speak of the Colors" (Johnston, 1992).

further, that this same tracking relation succeeds at this task in all potential mismatch cases. This is a tall order.

For what it's worth, it does seem to me that the representational contents in question really are vocal on whether the properties they represent are relational or complex. For example, colors really do seem to be bound to only the putatively colored objects, and thus to be features of only those objects. It is not just that they are neutral with respect to whether they involve other objects; rather, they explicitly *do not* involve other objects. We can contrast this case with cases where it really does seem that we are representing a relation. When you look at the picture below, you visually represent the goldfish as in the bowl.

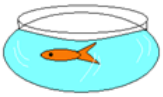

Figure 5.2: A goldfish in a bowl.

That's what it's like to represent a relational property. It is clear from the introspectively accessible content of this experience that (among other things) it represents a relation obtaining between the fish and the bowl. The representation of colors is nothing like this.[27]

5.2 From a mismatch case to anti-realism

So far, I have argued that there are mismatch cases. This already allows us to argue against tracking theories (see §5.1.5). But it's reliable misrepresentation that I need in order to argue against relation views generally, and a mismatch case is not automatically a case of reliable misrepresentation. For example, even if according to our favored tracking relation, color-experiences track something other than what they represent, it might still be the case that objects have the colors we represent them as having. We could be getting things right by accident. Reliable misrepresentation, however, requires not

[27]Similarly, we might compare color-representations with representations that clearly have complex contents, such as your visual representation of the above scene. There's an orange fish and a bowl. This total experience has an obviously complex content that allows for the possibility of decomposition in at least a few ways. We could remove the fish, or the bowl. We could also change the fish's color without changing its shape, or its shape without changing its color. As far as the content of the experience is concerned, it is possible to change the color of the fish without changing other parts of the experience. In contrast, as far as our color-experience is concerned, we cannot change part of the fish's color, so that it is represented as reflecting, say, more wavelengths of light at the 500nm. What I am suggesting here is that color-representations not only fail to seem decomposable into other parts, but that they seem to preclude the possibility of such a decomposition. Still, my confidence in this argument is lower than my confidence in the arguments in the main text, which is why this appears in a footnote.

only that a representation track something other than what it represents, but also that what it represents is *false*.

In this section, I argue that it is sometimes possible to argue from a mismatch case to anti-realism about the represented property, where **realism** about a property is the view that that property is instantiated in our world, and **anti-realism** is the negation of realism. If anti-realism about the property represented in a mismatch case is true, then the representations in question are indeed non-veridical, and we have a case of reliable misrepresentation.[28]

The best reason to think there are color instances is that we have color-experiences. Let's grant that perceptual experience at least sometimes provides us with evidence that objects are colored.[29] If we find out that our representations track something other than what they represent, then the evidence provided by our perceptions is debunked or canceled. The fact that our representations track something other than what they represent (say, surface reflectance properties rather than colors) means that what our representations represent need play no causal role in generating our mental states. That is, as far as everything causal is concerned, we would have the same thoughts and experiences involving those representations *whether or not* what they represent exists or is the case. Evidence that would exist entirely independently of what it's supposed to be evidence for ceases to provide justification. For example, suppose Professor Easy gives As to all his seniors, no matter what. John gets an A. Normally, getting an A is evidence that a student's performance was good. However, John is a senior, so he would have gotten an A *whether or not* his performance was any good. So the evidence his getting an A appeared to provide is debunked or canceled. I will argue that we are in such a situation with respect to at least some of the mismatch cases presented above. I will primarily focus on color-representation. I will argue that, given everything else we know about color-experience, we have reason to think that we would have perceptual experiences of colors *whether or not* anything was actually colored. That is, our experiences are relevantly like Professor Easy's grades: They would tell us that there are color instances whether or not there actually were color instances. This means that whatever evidence for the existence of color instances color-experiences were initially

[28] Anti-realism about a property represented in a mismatch case is stronger than necessary in order to establish a case of reliable misrepresentation. Color-representations can reliably misrepresent even if there are in fact color instances. The color instances might be in the wrong place—say in the center of black holes—or we might make systematic errors in attributing colors—perhaps everything is green, except for the things we represent as green, which are red, or perhaps everything has colors that we cannot represent given our visual systems, and that are only ever represented by snakes. These are all cases of reliable misrepresentation without anti-realism.

[29] On a Williamsonian conception of evidence, evidence must be true (Williamson, 2000). This is not my usage of the word "evidence." If this is how you use the word, read "evidence" as "apparent evidence" or "putative evidence."

thought to provide is debunked. Color-experiences offer no good reasons to think that objects are colored.

There might nonetheless be other evidence in support of the existence of color instances. I will argue, however, that there is no evidence that should motivate those not already committed to a relation view. Given a few further plausible assumptions, we can then argue that the appropriate epistemic attitude to take towards color instances is that there are none.[30]

5.2.1 From a mismatch case to a debunking argument

I claim that, given that the case of colors is a mismatch case, according to good stories about color vision, we would have color-experiences whether or not there were color instances. First, color instances needn't be invoked in a good causal-processing story of color perception. Second, color instances needn't be invoked in a good historical story of how we came to represent colors. The irrelevance of color instances to these two types of explanation directly stems from the fact that the case of colors is a mismatch case. Thus, the result generalizes: there are similar stories to be told for other mismatch cases.

The causal-processing story

A **causal-processing story** about a phenomenon is a story about what happens and in what order such that a certain effect is brought about. A causal-processing story about color perception tells us what happens in the world, how it affects us, and how one bodily or brain state leads to another, eventually resulting in the activation of color-representations. We can distinguish a causal-processing story about color perception from a metaphysical story about color perception. Perhaps, in some important and non-causal way, what God is up to is somehow constitutively relevant to our color-experiences. Even if this were the case, we wouldn't need to invoke God in a causal-processing story about color perception, although we would need to invoke God in a full story about color perception.

My claim here is that color instances, whether or not they exist, need not be invoked in our best causal-processing story of color-experience. This is in contrast with the cases of shape-, distance-, and size-representation. Shapes, distances, and sizes are **distal stimuli** of shape-, distance-, and size-experiences. As a result, a good causal-processing story of shape-, distance-, and

[30]This form of argument bears resemblance to debunking arguments used in other domains, such as ethics. The general strategy is to show that a particular type of judgment is formed in a way that is not sensitive to the truth of that judgment. This cancels or removes the justification thought to be provided by the judgment. If there is no other evidence in favor of the proposition in question, we are left with no reason to believe it. For example, in ethics, Josh Greene (2008) argues that our deontological moral intuitions are the result of unreliable processes that track morally irrelevant features like disgust. If deontological moral principles have no independent justification, then we have no reason to believe that they are true.

size-experiences will invoke the shapes, distances, and sizes of objects. The following is a sketch of such a story about size-experiences:

The size of objects causes them to reflect light in certain ways. This light stimulates a certain part of the retina. Processing occurs, and size-experiences result. Of course, many details remain to be filled in. This processing probably includes, among other things, using various cues for size, such as distance cues, in order to attribute a size to the object in question. But the point right now is that the story invokes the size of objects as a distal stimulus of size-experiences. There may be other distal stimuli of size-experiences, but that doesn't matter here. It's enough that size is one.

The distal causes of size-experiences include the *particular* sizes of objects. What would size-experiences be like if objects had different sizes? They would be different. What if there were no sizes? It's hard to make sense of this possibility, but if we can make sense of it, we should think that size-experiences would be very different. The point here is that it's not true that we would have the size-experiences we actually have if objects didn't have those sizes, either because they had other sizes or because they didn't have size at all.[31]

Importantly, these are claims about size-representation in general. I am not claiming that every particular instance of size-experiences requires the existence of sizes, or particular sizes. There can be size illusions and size hallucinations. The point is that, usually, size-experiences are causally dependent on the existence of sizes in the ways described.

Here's a similar causal-processing story about shape-experiences: The shapes of objects cause them to reflect light in certain ways, this light hits the retina, processing occurs, and finally experiences of various shapes result. Again, specifying the precise processing that occurs will be complicated. A more detailed story will invoke certain cues for shape, perhaps including distance cues for the object's various parts or cues as to what type of object the object in question is. But again, the point here is that a good perceptual story about shape-experiences invokes the shapes of objects as distal stimuli of shape-experiences. Given what we know about shape-perception, if these objects had different shapes or no shapes at all, our shape-experiences would differ. So it's not true that we would have those very same experiences whether or not there were shapes.

In contrast, our best causal-processing story of color-experience will probably not invoke colors as distal stimuli. Given that color-representation is a mismatch case, color-representations track something other than

[31] According to special relativity, the length of an object moving relative to an observer contracts the faster the object is moving, and so there is no non-relative notion of length, size, or shape. One might suggest that this makes size and shape bad examples for my purposes, since they too migh be mismatch cases. I am sympathetic to this suggestion. Our perceptual error about the world might be quite pervasive. Still, we can make sense of what *would be* a successful example (e.g. absolute shapes and sizes), and go with that. Since I only invoke these cases to draw a contrast, it is sufficient for my purposes.

what they represent; they track something other than color instances. We, therefore, can give a causal-processing story of color-experiences that only invokes what they track (or one of the many things they track), and need not invoke what they represent. We can say that objects have certain surface properties, these surface properties cause them to reflect certain proportions of different wavelengths of incident light in certain ways, this light hits the retina, some processing occurs, and color-experiences result. Again, more will have to be said about the processing that occurs, but the point here is that there is no need to invoke color instances as distal stimuli of color-experiences.

This is not yet an argument for anti-realism about colors. It's compatible with this story that objects really do have colors. The point now is just that colors are not *required* in order to explain color-experiences as far as a causal-processing story is concerned.

To summarize, a good causal-processing story about color-experience does not require the existence of color instances as distal stimuli. In the case of shape, a good processing story about shape-experiences includes shapes as the distal stimuli of those experiences. Shape-experience would be different if there were no shapes at all. If there were different shapes or no shapes, the distal stimuli of shape-experience would be different or missing, and so we wouldn't have the shape-experiences we in fact do have. The presence or absence of a particular shape in front of us makes a causal difference for our shape-experiences. Not so for colors. The presence or absence of a particular color does not make a difference for our color-experiences. Our color-experiences would proceed as usual. As far as a good causal-processing story is concerned, we would see color instances *whether or not* there were any.

Importantly, these claims are about what a *good* causal-processing story of size-representation is likely to look like. Of course, there are *possible* causal-processing stories that are compatible with all our evidence and on which size is not a distal cause of our size-experiences, perhaps stories involving evil demons and the like. Put otherwise, it can't be that any time there is *some* possible story on which we'd have some experience whether or not it was veridical we have the grounds for debunking the evidence obtained from that experience. Rather, we should say that the evidence conferred by an experience is debunked only when according to what is likely to be a *good* story surrounding the generation of that experience does not invoke what the experience is supposed to be evidence for. A *good* story should be compatible with everything else that we reasonably believe, e.g. about vision. It need not be compatible with possibilities involving evil demons (unless, of course, we have good reason to believe those possibilities are in fact the case). So the difference between the case of color and size comes to this: A *good* causal-processing story of color-representation does not invoke colors, while it is not true that a *good* causal-processing story of size-representation

does not invoke sizes (even though there are *bad* causal-processing stories about size-representation that don't invoke sizes).[32]

The historical (evolutionary) story

It is likely that a good **historical story** about how we came to have color vision likewise does not require the existence of color instances. In this case, the historical story would likely be an evolutionary story explaining how the capacity to represent colors came to be naturally selected. In other cases, the historical story may be different. For instance, it might be a story about how individuals acquired a concept or representation in question through some kind of learning process.

If a good evolutionary story about color-representation does not require the existence of color instances, then we can say that we would have *come* to represent color instances *whether or not* there were any. So, as far as the historical story is concerned, we would see color instances whether or not there were any.

Again, let's compare the case of colors to the case of other properties we represent, such as shape, size, and location properties. Here's what a good evolutionary story about spatial vision is more or less likely to look like:

Food, shelter, predators, mates, and other things important to our ancestors occupied various locations at various distances away from their bodies. It was useful to keep track of where these things were. Spatial representation conferred a survival advantage and was selected for. Of course, a full evolutionary story will ultimately have to explain why spatial representation is as developed in humans as it is, among many other things. But the point for now is that a fairly plausible evolutionary story about spatial representation will probably have it that the very thing we're representing is the same thing that was useful for us to track: spatial locations. The evolutionary pressures resulting in spatial representation were driven by the need to keep track of spatial locations themselves. Put otherwise, our spatial experiences must be more or less veridical in order to be useful. And so, a good story about the evolution of spatial representation involves spatial locations.

In contrast, a good evolutionary story about color vision will probably not involve objects actually having colors. One not entirely implausible story about the evolution of color vision has something to do with evolving to discriminate different objects of the same luminance, or perhaps evolving to detect ripe fruit against a background. Let's go with the ripe fruit story for now. Something analogous can be said if one of the other stories about the evolution of color-vision turns out to be correct. According to this story, it was useful for our ancestors to be able to discriminate ripe fruit from unripe fruit, even if they had the same shape and luminance. A person

[32]This is why a local debunking argument need not overgenerate and lead to global skepticism. Thanks to a seminar at Princeton University led by Tom Kelly and Bas van Fraassen for discussion.

could effectively do this if she was able to track the difference between ripe and unripe fruit and visually represent ripe fruit somehow differently from unripe fruit. This could be done by both tracking and representing the different surface reflectance properties of ripe and unripe fruit. But it could also be done by tracking surface reflectance properties and representing something else, like *sui generis* color properties, as long as ripe and unripe fruit were represented as having different colors. And if the arguments of the first part of this chapter are sound, this is in fact the solution that was implemented. This evolutionary story about color vision does not require objects to actually have colors. Color-experience would have been selected for *whether or not* objects have the colors we represent them as having. That's because the important thing we're keeping track of is the ripeness of fruit, particular objects encountered at various times, or some such thing, and we do that by tracking surface reflectance properties, not by tracking colors. Put otherwise, while a good historical story about location perception involves actual locations as distal causes of our coming to represent spatial locations, the analogous claim about color perception doesn't hold; a good historical story about color perception does not involve color instances.

One might even argue that representing ripe fruit as colored is more conducive to the finding of ripe fruit than representing ripe fruit as having a certain reflectance profile would have been. Simpler properties represented in experience tend to "pop out" in a way that more complex properties do not, making it easier to search for objects with those properties.[33] It's not clear, therefore, that something like a complex reflectance property could similarly pop out. So, given the choice between representing colors versus representing complex surface reflectance properties, there may even be considerable advantage to systematically misrepresenting colors over veridically representing surface reflectance properties: This would make ripe fruit easier to find. It might also be easier or cheaper for nature to design a creature that represented simpler properties than one that represented complex properties. So natural selection might favor the systematic misrepresentation of objects as having simpler properties when they in fact have more complex properties.

In summary, given what we know about color vision, it is very likely that a good evolutionary story about how we came to have color vision will not invoke color instances. Thus, as far as a historical explanation of color vision is concerned, we would see colors whether or not there were color instances.

The debunking argument for the perceptual evidence for colors

As far as the causal-processing and the evolutionary stories are concerned, color vision would proceed as usual whether or not there were colors. The

[33]See Treisman and Gelade (1980). While the theory built on these results is controversial, what's relatively uncontroversial is that color properties admit of a parallel, as opposed to serial, search strategy (color properties "pop out") while more complex properties do not.

existence or non-existence of colors has absolutely no effect on color vision as far as these stories are concerned.

Might there be another story about color-representation that does require the existence of colors? Perhaps there is a metaphysical story about color perception, a story about what color-experience *really is*, on which colors must be instantiated in order for us to perceive them. Perhaps color instances are *constitutively* involved in color-representation. I will address this suggestion later on when I address those already committed to a non-tracking relation view, since relation theorists are the ones likely to endorse this kind of role for color instances. Right now, I'm addressing everyone else. For now, I am content to argue that all the considerations independent of our metaphysical theory of mental representation point to the following conclusion: We would see colors whether or not there were color instances.

Where does this leave us with respect to our evidence for the existence of color instances? As noted earlier, the most obvious source of such evidence is our perceptual experiences of color. But perceptual experience is only good evidence if we don't have independent reason to think that we would see colors whether or not objects really had them. It turns out that, given what we know about color vision, we would have color-experiences whether or not there were color instances. This, therefore, debunks our perceptual evidence for objects having colors.

Extension of the argument to other mismatch cases

What allowed us to make this argument in the case of color is that color properties do not causally affect color-experiences in any way. This is because color is a mismatch case. If color-experiences are causally independent from color properties, then there is bound to be both a causal-processing explanation for how color vision proceeds independently of the existence of colors, as well as a historical explanation of the emergence of color vision that also does not require the existence of color instances. The actual details of the processing or historical explanations don't particularly matter. If color properties do not causally interact with color-representations, then they need not figure in either story. Further, they need not figure in a *good* story, because presumably a good story will only include actual causes of experiences.

If this is right, then analogous debunking arguments should be available for other mismatch cases. A processing story of sweetness-experiences need not invoke *sweetness* as a distal stimulus of the experiences; it need only invoke sugars. Likewise, an evolutionary story of sweetness-experiences need only invoke the need to detect sugars, high-calorie foods, or ripe fruit in order to explain the emergence of sweetness-representation. However, not all mismatch cases lead to anti-realism. The next few sections discuss what else is needed.

5.2.2 A lack of further evidence

So far, we've debunked our visual evidence for color instances. But this isn't yet an argument for color anti-realism. There could be other evidence for the existence of colored objects that does not come from visual experience. In the case of shape, we can see shapes as well as touch shapes. Our best physical science confirms that objects come in various different shapes. Even if it turned out that visual shape-experiences were not good evidence for the existence of shapes, we could fall back on our other sources of evidence for shapes. Here's a toy example to make this point: Suppose we find out that an evil demon is randomly manipulating our visual shape-experiences. The evil demon is manipulating them in such a way that is not sensitive to the shapes, if there happen to be any, that are actually in front of us. If we know about this demon, then we have a debunking argument for visual shape-experience. But we can still go out and *touch* shapes. Finding out about the evil demon does not debunk our *all* our evidence for the existence of shapes, so we still have some reason to be realists about shapes.

Unfortunately for the color realist, no other perceptual experiences offer evidence of color instances. We cannot touch, smell, hear, or taste color.[34] Neither does our best physical science tell us that there are color instances. Unlike in the case of shapes, there are no empirical reasons to think that there are color instances.

One might suggest that common sense offers some independent evidence for color realism. Whether and why the mere fact that we happen to believe something is evidence in its favor is controversial, and I don't want to enter the debate here. However, even if we grant that our beliefs can offer evidence in the way required, this will not be of much use in the present case. The reason we have the commonsense belief there are colors is that we see colors. So whatever justification our belief that there are color instances has is derived from the justification of our perceptual color-experiences. It is not obtained from some other independent source, or from nowhere at all. If we can debunk the perceptual experiences, then we've thereby also debunked any justification arising from our commonsense beliefs about color instances.

There is an important caveat, however. Perhaps our color concepts represent something other than what our color-experiences represent. Maybe we intend our color concepts to stand for the causes of our color-experiences, or the types of properties satisfying as many as possible of some list of features. This could be because we've reconstructed our concepts (see §5.1.4). That's all fine. Many of *these* properties are instantiated, but *they* aren't

[34] In some cases of synesthesia, subjects claim to, e.g., hear colors. However, whatever else we say about these cases, they do not hear colors *as colors*. Rather, what is meant by the claim that they hear colors is that when they see a particular color they also hear a particular sound. For example, visual experiences of the color red might be accompanied by auditory experiences of the sound of a trumpet. Such a subject is hearing trumpets, not hearing red. For more on synesthesia, see Cytowic and Wood (1982).

colors, as I am using the term "color." I'm interested in colors in the sense of *what is represented to us by color-experiences.*

5.2.3 From a lack of evidence to anti-realism

So far, I've argued that there is no good evidence for the existence of color instances. In this situation, I claim we should conclude that there are no color instances.[35] There is much that can be said about what the correct credence in the existence of color instances should be in this situation in which there is no evidence for or against their existence. However, I think that anyone who is on board so far and who has no antecedent commitment to a metaphysical story of mental representation that requires color realism—a type of non-tracking relation view—will be happy to accept color anti-realism, so I will not pursue the discussion further. Instead, I turn to addressing those who are thus antecedently committed.

5.2.4 For those already committed to a non-tracking relation view

So far, I have argued that if you are not already committed to a non-tracking relation view, then you have no reason to think that color properties are ever instantiated. You then have good reason to be a color anti-realist. The arguments so far are directed towards the non-committed. But those already committed to a non-tracking relation view might resist color anti-realism on the following grounds: A metaphysical theory of perception requires it. Someone arguing like this might agree that a causal-processing story about color perception does not require colors but insist that a causal-processing story leaves out the most important part of the story: the feature of the experience that *makes it* a color-experience. In what follows, I will address this kind of view.

It must be admitted that this is a possible way to argue for color realism. However, this strategy fairly quickly leads to an unattractively inflated ontology. Similar motivations will exist in favor of extending this strategy to support realism about hotness and coldness, heaviness, sweetness, and other mismatch cases that are amenable to the arguments discussed in this chapter so far. This quickly leads to what we might call **Rampant Realism**. I submit that if the theoretical options are to adopt a different view of perception or mental representation on the one hand, and Rampant Realism on the other, the former view is preferable.

It is also worth noting that positing primitive or *sui generis* facts about mental representation is less ontologically costly than Rampant Realism. If it

[35] Agnosticism or 0.5 credence about color realism would also allow me to draw my desired conclusion. The relation view requires realism to be true of all the properties that generate mismatch cases. The more mismatch cases there are, the less confident we should be that realism is true of all of them, even if we are agnostic or have a credence of 0.5 in each case in isolation.

turns out that the only view of mental representation that can avoid Rampant Realism is one on which representation is a primitive property or relation, considerations of ontological simplicity will still favor this view. Even if we had to posit sense-data in order to avoid all these primitive properties of objects, it would be worth the cost. We would only have to add one type of thing, sense-data, to our ontology, instead of many separate types of things, as required by the Rampant Realist.

A second point worth noting is that although there is still a consistent position to be had by the non-tracking relation theorist, everything that was said so far in this chapter makes his situation *worse*. In the case of shapes, we have lots of reason to think there are such things. Visual shape-experience is not a mismatch case, and there is corroborating evidence for the existence of shape instances, e.g. from touch. This makes shapes respectable things to appeal to in a theory about something else other than shapes. *Ceteris paribus*, it is not a criticism of a theory that it appeals to shapes because we already agree that there are shapes. However, in the case of colors, all the non-controversial evidence and arguments point toward anti-realism. Non-tracking relation views have to posit the existence of things we have no independent reason to think exist. And these things are supposed to be on the surfaces of all objects. They are also on the insides of objects, or at least the ones we cut open. Not only must non-tracking relation theorists posit color instances, but they must also posit entities corresponding to all other mismatch cases that we likewise have no independent reason to think exist or are instantiated. We not only have Rampant Realism, but we have Rampant Otherwise Unmotivated Realism.

In this respect, the non-tracking relation theorist's appeal to colors is almost, but not quite, as bad as certain problematic appeals to God. Suppose one posits the existence of God *only* to account for the vital force animating humans and animals. The resulting position may be coherent, but if nothing else supports the existence of God, this is an unhappy situation. It would be preferable to admit a primitive vital force. Likewise, primitivism about mental representation is preferable to Rampant Realism.

I said the situation is almost, but not quite, as bad. An important difference between the positing of colors by the relation theorist and the positing of God by the vitalist is this: There are clear alternatives to vitalism that better explain everything it aims to explain. So we have little reason to posit extra entities in an effort to cling onto vitalism. One might argue that non-tracking relation views aim to explain something that *cannot* be explained in any other way. As argued in the first half of this chapter, tracking relation views don't work, and it might be thought that relation views are the only game in town.

But they're not the only game in town. There are also production views. Here is the Worst Production View Ever: Mental representation is a *sui generis* or primitive feature of mental states. It is just there and there's no more that can be said about it. It doesn't interact with anything else we

care about or believe in. It doesn't even supervene on brain states. The Worst Production View Ever is *still* better than non-tracking relation views that are realist about all the entities mentioned above. First, it posits a lot fewer entities. It posits *sui generis* representational properties, while the non-tracking relation view posits colors, hotness, coldness, sweetness, etc., as well as what might end up being a *sui generis* representation relation to those things. Second, the new features my toy theory posits appear *in the right place*: they are new *perceptual* or mental posits, not new posits in other domains. If some phenomenon resists reduction and we need to add primitives, it's better to add them where the problem is, not somewhere else. A problem in perception should get a fix in perception, not on the surface of objects. (This is why positing a primitive vital force is preferable to positing God to explain animal movement.) Finally, Chapter 3 discusses a metaphysical problem with relation views in general. If that is at all convincing, it provides more reason to prefer my Worst Production View Ever to the non-tracking relation view.

It is worth mentioning that the non-dogmatic non-tracking relation theorist might still find something convincing in my previous arguments for anti-realism aimed at those not already committed to a non-tracking relation view. This theorist encounters mismatch cases for which we have no independent evidence of realism, such as the case of color. She encounters them one after the other. Her commitment to the non-tracking relation view constrains her credence in the existence of the entities in question in each case. But since she is non-dogmatic, her credence in the non-tracking relation theory is less than 1, and so she admits a small credence (say, 0.1) to the possibility of anti-realism about each of the relevant mismatch cases. But then if we have ten such mismatch cases, her credence that realism is true in all cases should be $0.9^{10} = 0.35$. Given her credences, then, it's epistemically more likely that there is at least one case for which anti-realism is true and the non-tracking relation view is false than that there are no mismatch cases of the relevant sort and the non-tracking relation view is true, so she should revise her beliefs. (Of course, the numbers can vary.)

This concludes my argument for anti-realism about color. If color anti-realism is true, then the case of color-experience is not only a mismatch case, but also a case of reliable misrepresentation. We represent objects as having colors, but they do not in fact have those colors, because nothing is ever colored. In Chapter 4, I argued that relation views have difficulty allowing for reliable misrepresentation. In this chapter, I argued that there are actual cases of reliable misrepresentation. Together, they form an argument against relation views generally. One way to put the problem briefly is this: Relation views account for our experiences in terms of our relations to things that exist. But there are cases, such as the case of color-experience, for which the entities that would have to exist to account for our experiences do not in fact exist. Therefore, relation views are false. We don't get to represent

colors by being related to colors or color instances. We produce contentful color-experiences.

5.3 Conclusion

I have argued that some of our representations track something other than what they represent, and that in some cases, this provides the basis for a debunking argument against our evidence for realism about those represented properties. The existence of mismatch cases is already a problem for tracking relation views. And anti-realism about colors and other relevantly similar properties is a problem for relation views generally. In short, there just aren't the right things in the world for us to be related to in order for a relation view to account for the representational contents we can and do entertain.

Chapter 6

The Significance of Tracking

In cases of reliable misrepresentation, what a representation tracks differs from what it represents. If this is right, then tracking isn't representation. But this doesn't mean that tracking isn't an important feature of mental representation. Various tracking relations obtain between representations and things in the world, and we might invoke various of them for different explanatory purposes. But since these tracking relations are not in competition with each other for some privileged status as The Representation Relation, we can be pluralists about tracking, invoking different notions of tracking for different explanatory purposes. Distinguishing tracking from representation and accepting a kind of pluralism about tracking both have a number of explanatory advantages.

6.1 Types of reliable misrepresentation

Cases of reliable misrepresentation are cases where tracking and representation come apart, often in adaptive and useful ways. In the cases examined in Chapter 5, our mental states track a *more complex* property while representing a *simpler* property. Simpler properties may be easier to represent and manipulate in thought than complex properties, and the neglected information may not be particularly useful. Knowledge of surface reflectance properties *per se* is not important for survival, but being able to reidentify objects over time, discriminate different objects of the same luminance, and pick out the ripe fruit and poisonous frogs is. If representing colors can guide our behavior towards the environment just as well as representing surface reflectance properties can, and if representing colors is cognitively cheaper or more efficient, then we should not be surprised that we systematically misrepresent objects as having colors; the extra neglected information is not useful, and it would be costly to represent.

Another type of adaptive mismatch case might arise when a relational property involving n relata is relevant to our survival, but it just so happens that in our environment, one of the relata remains constant. Then it might be more economical for us to systematically misrepresent the objects

in question as having a different property with *n-1* relata. The case of perishability is a good example here. Whether or not a particular type of food is perishable is important to us. But, strictly speaking, foods aren't perishable or nonperishable in some absolute way. Rather, they are perishable or nonperishable *relative to an environment* or *relative to a climate*. Meat is perishable-in-Africa but not perishable-in-Alaska. Still, for a group of humans that does not travel, the information provided by the extra relatum is not important. Humans living in Africa will survive equally well whether they represent meat as perishable-in-Africa (or perishable-around-here) versus as just plain perishable. In fact, they may be better off representing the meat as just plain perishable, since representing the property with one less relatum might be cheaper, more efficient, or might block certain kinds of errors. So we have a case where it is important to track a relational property (e.g. perishable-in-Africa, or even perishable-around-here), but it is sufficient to represent a slightly less relational property (e.g. perishable).[1] Heaviness might be similarly diagnosed. Although our heaviness-experiences track relational properties such as those obtaining between objects and Earth's gravitational field, Earth's gravitational field is a constant, so it need not be represented every time we represent something as heavy. For creatures that do not travel to other planets, representing a property with one less relatum than weight is just as good as representing weight, and probably cognitively cheaper.

6.2 Tracking contributes to *successful* behavior

As the above examples show, reliable misrepresentation needn't be harmful, and can often be adaptive and useful. It is at least partly *because* our mental states track biologically significant features of the world that our mental representations enable us to get by so well. It is because, say, color-representations track surface reflectance properties common to ripe fruit that our color-representations enable us to find ripe fruit. And it's because such relations obtained between our ancestors' mental representations and the properties in their environments that we have representations of those types today. We can already start to see how pluralism about tracking is useful. We can invoke various tracking relations to explain successful behavior in various complementary ways.

The picture that is emerging here is this: Tracking and representation are distinct phenomena. Representational content is at least sometimes introspectively accessible and guides behavior (if you think any content guides behavior), while what a representation tracks explains why behavior generated in part by use of that representation is successful. In §2.2.3, we considered the possibility that representation might have something to do

[1] Thanks to Sungho Choi for offering this example in support of a different claim in conversation.

with how we generate behavior that helps us get around in the world. Now we see that representations *do* have a role in generating successful behavior, but an additional, crucial part of the story has to do with tracking. It is partly because our representations track things relevant to successful behavior that the behavior they lead to is so successful. Thus, we see that a story of how we manage to get around in the world must invoke non-representational features of representations, as well as genuine mental representational features.

6.3 Notions of success other than veridicality

Sometimes perception is successful. Other times it is not. Consider, for example, color illusions. We might want a notion of success that distinguishes between normal and illusory color-experiences. We might want to say that in some way or other, illusory color-experiences fail to be successful.

Tracking theories at least implicitly equate success with veridicality. What it takes for a token of a representation to be veridical is for the representation to be tokened in the presence of things that do or would cause it to token in ideal conditions. If we equate success with veridicality, we can say that the failure in the case of color illusions is a failure in veridicality: illusory color-experiences are not veridical and that constitutes their not being successful. But if I am right about color representation, then normal color vision isn't veridical either. So this cannot be the (only) way in which illusory color-experiences fail to be successful.

There are other notions of success that we can invoke to capture the distinctions we want to make. In fact, we can invoke all the notions the tracking theorist wants to identify with veridicality, even while recognizing that they do not constitute veridicality. We can say a token of a representation is unsuccessful or defective when it tokens in the absence of what it tracks. For example, we can say that illusory color-experiences involve tokening a representation in response to objects that would not cause it to token in normal, optimal, or design conditions.

We can say everything the tracking theorist says about success, with two important differences: First, we need not equate success with veridicality. This type of success is not a *semantic* type of success. Whether or not a representation succeeds at truth or reference is a separate question. It is also not a normative notion of success, or if it is one, it involves a very shallow type of normativity. There are no deep normative facts about what representations should token in response to. Since this is not a semantic or normative notion of success, we can apply it to states that are not even representations but that nonetheless track things, such as states of one's stomach, or states of thermometers.

Second, corresponding to different tracking relations, there are different (non-semantic) notions of success. We need not privilege one such notion of success. Pluralism about tracking allows for **pluralism about non-**

semantic success. The very same tokening of a representation might be successful with respect to one tracking relation but not with respect to another.

Consider, for example, a case that has generated much discussion among tracking theorists: the case of magnetotactic bacteria. Magnetotactic bacteria have internal magnets, or magnetosomes, which align parallel to the surrounding magnetic field, much like compass needles. The bacteria propel themselves in the direction in which the magnetosomes are pointing. Magnetotactic bacteria living in the northern hemisphere have their magnetosomes oriented in a way such that they propel themselves them towards geomagnetic north. Since geomagnetic north is in the downward direction in the northern hemisphere, in effect, they swim towards deeper water, thereby avoiding the toxic (to them) oxygen-rich surface water. Bacteria living in the southern hemisphere have magnetosomes oriented in the opposite direction, and hence swim towards geomagnetic south, which also leads to deeper water and away from oxygen-rich environments. If we (a) move a northern magnetotactic bacterium to the southern hemisphere, it will propel itself to the surface and die. A great deal of debate has ensued over whether and what such a displaced bacterium is misrepresenting.[2] Is the bacterium veridically representing geomagnetic north? Is it misrepresenting magnetic north? Is it misrepresenting the location of oxygen-free water?

We can compare this case of displacement with two other cases, the first of which has also generated some discussion in the literature: (b) A northern magnetotactic bacterium is placed near a bar magnet that is oriented in the direction opposite to the geomagnetic field. (c) A northern magentotactic bacterium has a genetic mutation such that its magnetosome points to geomagnetic south; it is intrinsically identical to a normal southern magnetotactic bacterium. These two bacteria also propel themselves to the surface and die.

Candidates for what magnetosomes track include the following: (i) Magnetic north, (ii) geomagnetic north, and (iii) the presence of oxygen-free water. Recall that on the non-semantic notions of success currently under discussion, a token of a representation is unsuccessful or defective when it tokens in the absence of what it tracks. If we say the northern bacteria are tracking magnetic north, then case (c) is defective, but cases (a) and (b) are not. If we say they track geomagnetic north, then cases (b) and (c) are defective, but (a) is not. If the bacteria track oxygen-free water (or the direction of oxygen-free water), then all three cases are defective. This paragraph is summarized in Table 6.1.

[2]It is important to note that most theorists do not think that magnetotactic bacteria represent at all because they fail to satisfy other conditions required for having a mind. Still, they consider this to be a good case for comparing rival theories of mental representation. The idea must be that other, more complex cases, would behave relevantly similarly. My claims about magnetotactic bacteria do indeed literally apply to the bacteria, but they should apply to other cases of tracking as well.

		Case		
		Displacement (a)	Bar magnet (b)	Mutation (c)
What magnetosomes track	Magnetic north (i)	Success	Success	Failure
	Geomagnetic north (ii)	Success	Failure	Failure
	Oxygen-free water (iii)	Failure	Failure	Failure

Table 6.1: The success of magnetotactic bacteria in various situations according to various tracking relations.

The point right now is that once we distinguish tracking from representation, we can say that all the attributions of success and failure are compatible. Relative to different tracking relations, the bacteria track magnetic north, geomagnetic north, and oxygen-free water. Each case can be assessed for success or failure relative to each of these tracking relations. The above table is in a certain relevant way complete. It tells us all there is to know about the (non-semantic) success of the magenetotactic bacteria in the three cases. It's also worth mentioning here that recognizing the more or less equal status of various tracking relations and thus various non-semantic notions of success allows us to make finer distinctions between importantly different types of failure. The tracking theorist focuses too closely on her favored tracking relation and thus loses sight of this point.

In sum, we can invoke all the same standards of success and failure that are invoked by tracking theorists to describe what goes wrong in certain situations, except that, for us, these are not semantic standards of success. And, in fact, the situation is better for us than it is for tracking theorists, since we do not need to privilege one standard as somehow more important or more relevant. For certain purposes, one might be suitable, while for others, another might be suitable.[3]

6.4 Conclusion

In Chapter 2, we observed a phenomenon: we enjoy states that represent. Whatever else relation views are supposed to do, they are supposed to explain this phenomenon. I have argued that in fact they do not. There are paradigm cases of mental representations that are triggered by things that they do not represent, even in ideal conditions. The best way to understand these cases is in terms of reliable misrepresentation, but this is very difficult on a relation view. If I am right, tracking and representation are two distinct phenomena, each relevant to mental life and behavior in very different ways.

[3]It is also worth noting that to the extent to which pluralism about tracking enables us to clear up pre-existing debates in philosophy of mind, that is some support for the view.

Part II

The Phenomenal-Intentional Identity Theory

Chapter 7

The Phenomenal-Intentional Identity Theory

7.1 Introduction

It is sometimes thought that there are two distinct mental phenomena, or two distinct mental features: mental representation and phenomenal consciousness. **Mental representation** is the phenomenon introduced in Chapter 2, the "aboutness" of mental states. It's something that we at least in some cases can notice in ourselves. We can also notice that there's something that it's like (Nagel, 1974) to be in some states. This is (phenomenal) **consciousness**.[1] There is something that it is like to see blue, and it is different from what it is like to see yellow or to feel pain. What it's like to be in a state is the state's **phenomenal character**.[2]

[1] I set aside the issue of access consciousness. See Block (1995) for the distinction between phenomenal consciousness and access consciousness.

[2] One way to get a grasp of the notion of phenomenal character is to pinch yourself. Your attention will immediately go to your experience's phenomenal character. When I say that some mental states have phenomenal character, I mean that they have something similar in kind to this. Of course, pain-experiences have a particularly vivid phenomenal character. That's what makes them a good example. But this does not mean that all phenomenal character need be so vivid to qualify as such. Color-experiences have less vivid, but still quite vivid, phenomenal character. Non-pain tactile experiences tend to have a still less vivid phenomenal character. Proprioceptive experiences representing the locations of various parts of one's body have a still less vivid phenomenal character.

Some experiences with phenomenal character are fleeting, arguably such as the visual experience of causation. That makes the phenomenal character more difficult to discern. Some experiences are often accompanied by experiences of a different type, making it difficult to discern the contribution of the two types of experiences to the total phenomenal character. I mention these cases because I want to emphasize that I am using the term "phenomenal character" in a broad sense of the term. If it's like something to be in some state, then that state has phenomenal character. Part of what I will eventually argue is that thoughts also have phenomenal character, although it is much less vivid than that of perceptual states, and much more difficult to disentangle from the phenomenal character of other states. I don't want my notion of phenomenal character to rule out that possibility from the get-go.

There have been various attempts to unify these two phenomena. **Intentionalism** is a family of views that aims to understand phenomenal character in terms of representational content.[3] Identity versions of intentionalism hold that phenomenal character *is type or token identical to* a species of representational content. Advocates of identity versions of intentionalism include Gilbert Harman (1990), Fred Dretske (1995), Michael Tye (1995, 2000, 2009), Alex Byrne (2001), and Frank Jackson (2004, 2005). Other versions of intentionalism are supervenience and determination versions, on which phenomenal character *supervenes on* or in some way *is determined by* representational content, respectively. William Lycan (1996), Georges Rey (1998), and David Chalmers (2004b) hold versions of these views. Although I endorse an identity version of the view, everything I say is meant to apply to other versions as well. So, for my purposes, I will try to remain neutral between the various alternatives. I will speak of intentionalism as the thesis that mental representation *determines* phenomenal character, where "determination" is meant to be understood vaguely enough to cover all these cases.[4] Intentionalist views can also be categorized based on purity. **Pure intentionalism** is the view that phenomenal character is determined by representational content alone. **Impure intentionalism** is the view that phenomenal character is determined by representational content *together with some other features*.

Another approach to unifying mental representation with phenomenal consciousness is the **phenomenal intentionality theory**, which aims to understand representational content in terms of phenomenal character. Advocates of such an approach include Galen Strawson (1994), Charles Siewert (1998), Terence Horgan and John Tienson (2002), Brian Loar (2003), David Pitt (2004), Uriah Kriegel (2003) and Katalin Farkas (2008). (See also Kriegel and Horgan's manifesto (forthcoming).) There is also a possibility of pure versus impure versions of this view, depending on whether phenomenal character alone is (or determines) representational content, or whether other conditions must be met or other factors must be present. There are also other versions of the view, on which phenomenal character determines representational content in a more complex way (such as in Kriegel (2003) and Farkas (2008)).[5]

The **phenomenal-intentional identity theory (PIIT)**, the view I will defend, qualifies both as a version of intentionalism and as a version of phenomenal intentionality theory. According to PIIT, representational content is type and token identical to phenomenal character. For every

[3]The view also goes by the names "representationalism" and "representationism."

[4]Sometimes intentionalists endorse the further claim that the intentional is explanatorily prior to the phenomenal. I am not using the term "intentionalism" to include this further claim.

[5]Phenomenal intentionality theorists usually endorse the further claim that the phenomenal is explanatorily prior to the intentional. As with the term "intentionalism," I am not using the term "phenomenal intentionality theory" to entail such further claims about explanatory priority.

type of representational content, there is a type of phenomenal character that it is identical to. For every type of phenomenal character, there is a type of representational content that it is identical to. Every token of every type of representational content (or phenomenal character) is thus identical to some token of phenomenal character (or representational content). Since representational content *alone* suffices for phenomenal character, and likewise phenomenal character *alone* suffices for representational content, PIIT qualifies as a version of *pure* intentionalism and *pure* phenomenal intentionality theory.

PIIT claims that there is one thing that answers to both "consciousness" and "representation." The version of PIIT that I want to defend is one on which consciousness/representation has some of the features that we may have previously associated with consciousness as well as some of the features we may have previously associated with mental representation. A feature may have previously associated with consciousness but not with mental representation is being a product of the mind or brain. On my version of PIIT, a production view is true of consciousness/representation. One feature we typically associated with mental representation but perhaps not with consciousness is essentially displaying aboutness, or "saying something" about the world. On my version of PIIT, consciousness/representation has this sort of "aboutness"; there are no mental features that do not "say something" about the world.

My argument for PIIT will proceed as follows: In §7.2, I will argue that there are good reasons to favor *some* kind of identity theory, that is, there are reasons to think that either an identity version of intentionalism, an identity version of phenomenal intentionality theory, or PIIT is true. In §§7.3 and 7.4, I will attempt to discern which of these theories is true by considering various alleged counterexamples to the theories. I will argue that there are no successful counterexamples to intentionalism or the phenomenal intentionality theory, and thus that there are no counterexamples to PIIT, and we should accept it.

7.2 Motivations for some identity claim

7.2.1 Introspection reveals *one* mental feature

When we look at a gray object, like the elephant depicted in Figure 7.1, it seems introspectively inaccurate to say that there are two grayness-related mental features related to our experience, a phenomenal "what it's like" of grayness, and a represented content *grayness*. It is more accurate to say that there is only one grayness, and it may be correctly described as both a represented property of the elephant, and a phenomenal character.

Put otherwise, our primary evidence for the existence of mental features is introspective. We think there is mental representation because we observe it in ourselves, and, likewise, we think there is phenomenal character because

Figure 7.1: A gray elephant.

we observe it in ourselves. But when we focus on an experience, say the experience invoked by the elephant, we can only discern one grayness-related mental feature, not two. Thus, we only have introspective evidence for one such feature.[6]

We likewise see that there is only one roundness-related mental feature corresponding to the bump on the elephant's head, and so we have evidence for only one such feature as well. The same goes for other features of our experience.

Such considerations are often taken to support identity versions of intentionalism.[7] However, they equally support identity versions of the phenomenal intentionality theory, since what they support in the first instance is the claim that there is only one mental feature.[8]

7.2.2 An impressive correlation

A second reason to think that there is an intimate relationship between phenomenal character and representational content is that, to the extent to which representational content and phenomenal character may seem distinct, the two mental features are impressively correlated. If two experiences differ

[6] I do not mean to deny that there are multiple mental features relating to grayness in the following sense: there is the grayness of one part of the elephant, the grayness of another part, as well as the grayness of "it looks gray" as opposed to the grayness of "it is gray." Still, for each of these graynesses, there does not seem to be two things, a "what it's like" to experience the grayness as well as a represented grayness, nor do these different graynesses neatly divide into grayness phenomenal characters and represented graynesses.

[7] This can be thought of as a version of the transparency intuition. One common way of putting the transparency intuition is this: When we try to pay attention to our experience, all we notice are represented objects and their properties, not intrinsic features of experience (Harman (1990) and Tye (2000, pp. 46-51)). While I agree that introspection of experience does not provide us with raw feels in addition to represented contents, it is sometimes thought that these arguments suggest that introspection further reveals that contents are external world objects and properties. However, it is not at all clear that introspection can reveal that.

[8] These observations do not directly support the claim that there are no introspectively inaccessible non-phenomenal representational states or even introspectively inaccessible non-representational phenomenal states. Rather, the suggestive feature of the observation is this: in the cases that *are* introspectively accessible, the phenomenal and the intentional do not seem wholly distinct. Thanks to Uriah Kriegel for suggesting this qualification.

in phenomenal character, then they also seem to differ in representational content, and vice versa.

This is obvious upon consideration of certain mundane cases. Compare a visual experience of a red pen to an otherwise similar visual experience of a green pen. The two experiences differ in phenomenal character, and they also differ in representational content. In most normal cases, if we change an experience's content, we end up changing its phenomenal character, and if we change it's phenomenal character, we thereby also change its content.[9]

These observations extend to representational states involving "conceptual" contents. Compare the experience of seeing the duck-rabbit as a duck to the experience of seeing the duck-rabbit as a rabbit (Figure 7.2). The two experiences differ in phenomenal character (what it's like to see the duck-rabbit as a duck differs from what it's like to see it as a rabbit), and they also differ in representational content.[10]

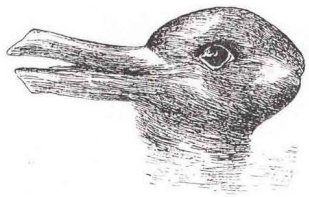

Figure 7.2: The duck-rabbit. It's like something different to see it as a duck than to see it as a rabbit.

Although mundane, these points are significant. If phenomenal character and representational content were not related in some interesting way, we would expect them to come apart in various situations. We certainly would not expect them to covary in this way.

Similar observations hold for other sensory modalities. The represented loudness or pitch of a sound correlates with the experience's phenomenal character. The same holds for tactile, olfactory, gustatory and proprioceptive experiences.[11,12] Finally, and more controversially, I think similar points hold

[9] Arguments arising from such considerations are also sometimes considered versions of the argument from transparency. See Jackson (2004), who argues that if two experiences differ phenomenally, then they also differ representationally, and so phenomenal character supervenes on representational content and some version of intentionalism is true.

[10] It is a matter of controversy whether these representational differences are differences in perceptual representational states, or differences in thoughts, beliefs, or other non-perceptual representational states. For now, I can remain neutral on this question. My claim is merely that the total representational states corresponding to the two viewings of the duck-rabbit differ in both phenomenal character and representational content.

[11] For an interesting discussion of olfactory representation, see Lycan (1996).

[12] There is also a phenomenology accompanying intending to make certain motions or perform certain actions. Whether or not this is considered a case of perception, it is also a case where differences in representational content are accompanied by differences in phenomenal character and vice versa. For a philosophical discussion of this phenomenology, see Maier (2008).

in the case of thought. However, since I want to heavily qualify this claim, I will save it for later.

For now, we can conclude that the preliminary evidence favors the existence of an intimate, perhaps constitutive, relationship between representational content and phenomenal character. Although intentionalists, such as Jackson, are primarily interested in using such observations to support the supervenience of the phenomenal on the representational, these observations also support the supervenience of the representational on the phenomenal and so support the phenomenal intentionality theory as well as intentionalism.

7.2.3 A unified theory of mind

It would be theoretically very nice if we could reduce the two problems of the mind to just one. A unified theory is simpler, and perhaps it would be a little bit surprising if there were *two* distinct and very difficult to account for mental features. The other two motivations provide reason to think that mental representation and phenomenal character are importantly related, while this motivation works in combination with them to provide an extra reason to think that this important relation is one of identity. We don't notice two distinct mental features, and to the extent to which we are tempted to ascribe two mental features to our experiences, the two features are surprisingly correlated. Given these preliminary observations, a simpler theory with only one source of mental features is quite attractive.

7.3 Putative counterexamples to intentionalism

So far, we have seen that there are various motivations for a view on which mental representation and phenomenal consciousness are intimately related. These considerations motivate both intentionalism and the phenomenal intentionality theory. In order to decide between the various possible theories, I will discuss various putative counterexamples to intentionalism and the phenomenal intentionality theory. A counterexample to (at least pure versions of) intentionalism is a pair of mental states alike in representational content but differing in phenomenal character. A counterexample to (at least pure versions of) the phenomenal intentionality theory is a pair of mental states alike in phenomenal character but differing in representational content. I will focus on alleged counterexamples that allow us to distinguish between the views in question or draw other interesting conclusions.

There are plenty of alleged counterexamples in the literature, and plenty of moves the intentionalist can make. In this section, I will only discuss two types of problems and the counterexamples they give rise to: the problem of apparently unavailable contents and the problem of non-perceptual states. One theme that will emerge from this discussion is that production view versions of intentionalism have more resources to deal with these alleged coun-

terexamples than relation view versions of intentionalism, and in particular, tracking versions.

7.3.1 Challenges arising from unavailable contents: Pain

Pains are paradigm examples of states with phenomenal character. If pains have phenomenal character but no representational content, then we might have an immediate counterexample to intentionalism: We could have two pains that differ in phenomenal character without differing in representational content (since they would both have no representational content). If that's the case, then representational content does not suffice to fix phenomenal character, and (at least pure) intentionalism is false.

Michael Tye suggests that pain-experiences represent *bodily damage of a certain type at a certain bodily location* (Tye, 2000, p. 50). Invoking the representation of *bodily damage* allows him to distinguish pains from itches and other bodily sensations that might be represented as being at the same bodily location. And invoking the representation of particular *types* of bodily damage allows him to distinguish between stabbing, throbbing, and other types of pains at the same bodily location.

It does seem that when I have a pain in my foot, I am representing my foot, and it also seems that pain-experiences are often accompanied by the thought or belief that there is bodily damage, perhaps of a certain type. But it is not so clear that pain-experiences *themselves* represent all these contents, and if they do, it is not clear that these represented contents account for the felt qualities of pain-experiences. Pain-experiences are often accompanied by the thought or belief that there is bodily damage at the pain location, but do the experiences themselves represent this damage? It's no coincidence here that Tye is a tracking theorist (see Tye (2000, pp. 64-5)). This might be why he is motivated to account for the content of pain-experiences in terms of a property in the world that things actually have and that we might be tracking. Given these constraints, bodily damage seems a likely candidate. But this option seems implausible for some of the same reasons certain tracking theorists' accounts of the content of mismatch cases seems implausible (see Chapter 5): Bodily damage does not seem to capture what our pain-experiences are representing, and in any case, it certainly does not seem to capture the phenomenal character of pain-experiences. It seems we have left the phenomenal character of pain-experiences out and in exchange let in lots of content that doesn't seem to be there or do any work in accounting for the phenomena.

Part of Tye's story is that the content of pain-experiences and other phenomenal states is *nonconceptual*. Perhaps an appeal to nonconceptual content can do some work in alleviating the above worry. For Tye, a state has **nonconceptual content** if its subject can entertain its contents without possessing the concepts involved in specifying that content (Tye, 2000, p. 62-3). One **has a concept** of P when, perhaps among other things, one is able

to identify instances of P on multiple occasions. For example, I do not have a concept of a particular shade of blue that I now perceive, say, blue_{421}, since when I see blue_{421} again, I will not recognize it as the same shade. So, when I perceive something as blue_{421}, my state's content is nonconceptual, and thus the state is a candidate for having phenomenal character (Tye, 2000, p. 61).

The first thing to note about this appeal to nonconceptual content is that it constitutes a move to impure intentionalism. On Tye's view, what accounts for the phenomenal character of pain is not only its representational content but also non-representational features of the state or the subject, in this case, perhaps something to do with the subject's ability to reidentify instances of the same type on different occasions.[13] The drawback of this sort of impure intentionalism is that considerations in favor of intentionalism discussed previously only properly motivate the relevance of representational content to phenomenal character. They do not additionally motivate the relevance of additional elements, such as whether a content is conceptual or nonconceptual. In order to properly motivate this form of impure intentionalism, something will have to be said about why a content's being nonconceptual is relevant to phenomenal character. But it is not at all clear how the inability to recognize the same type of bodily damage on different occasions should result in or have anything to do with the felt painfulness of pain, and so it is not clear that such a move to impure intentionalism is successful or motivated.

Tim Crane offers the alternative suggestion that an experience's **modality** (e.g. sight, audition) contributes to its phenomenal character. Pain-experiences represent bodily damage of a certain type in a certain location *in a certain modality* (Crane, 2008). Again, such a response involves a move to impure intentionalism, since an experience's modality will presumably be a matter of something other than what it represents. And again, it is not clear that this extra ingredient can do the work required of it. Suppose the relevant modality is the tactile modality. Then it is unclear how being represented in the tactile modality turns representations of bodily damage into experiences with the felt character of pain (rather than, say, the felt character of itches). If the modality is something more specific, perhaps a dedicated pain modality, then the modality is doing most if not all of the work in determining phenomenal character and the represented contents drop out of the picture. This would no longer qualify as a version of intentionalism.

David Bain suggests that pains represent damage in a certain bodily location under a certain **mode of presentation** (Bain, 2003). A problem for any intentionalist strategy appealing to modes of presentation is that modes of presentation are usually thought to be themselves representational.

[13] Tye's amounts to a construal of the conceptual/nonconceptual distinction as a difference in types of *states*, as opposed to a difference in types of *contents*. The distinction is usually attributed to Heck (2000). See also Bermúdez (2009) for discussion. The view I will eventually recommend in Chapter 10 allows us to construe the conceptual/nonconceptual distinction as a distinction between different kinds of contents.

Representing Venus under the mode of presentation of the morning star involves representing *the morning star*. Similarly, representing bodily damage under a "pain-y" mode of presentation would require representing some additional content. But then we need a story of what this content is such that it plausibly determines the phenomenal character of pain-experiences. And so we are basically back where we started, that is, in search of the contents represented by pain-experiences that can account for their phenomenal character.[14]

Instead, I claim we should just say that pain-experiences represent *sui generis* **pain properties** in certain bodily locations, where pain properties are those properties we are all familiar with but which cannot be further elucidated in terms of the physical properties we find when we examine our bodies by other means.[15] There are different types of pain properties, including stinging-pain properties, stabbing-pain properties, and throbbing-pain properties.[16]

Pain properties are probably never instantiated. Certain parts of our bodies sometimes sustain damage of certain types, but our pain-experiences reliably misrepresent those body parts as having pain properties instead. A state **reliably misrepresents** when all or most of its activations (or tokens) are nonveridical and occur in similar circumstances. Cases of reliable misrepresentation can arise when a mental state tracks something other than what it represents. In the case of pain, pain-representations track bodily damage of certain types, but they represent pain properties.

This amounts to an error theory about pain, which some might find unappealing. Resistance to this sort of error theory is often motivated by the idea that any representations that reliably help us produce successful behavior must get things more or less right. But this thought is mistaken. When it comes to survival and getting around in the world, *reliable* misrepresentation can be just as good as veridical representation. As long as damaged and undamaged body parts are represented differently, we have the means required for treating them differently as well, and that's what matters for successful representation-driven behavior. Put otherwise, representing a broken leg differently from an intact allows us to care for it. Representing the broken

[14]If, instead, modes of presentation are to be understood non-representationally, Bain's view constitutes a move to impure intentionalism, since something other than content determines phenomenal character, and it is not clear that this extra ingredient won't be doing all the work, as in the case of the amended proposal involving modality.

[15]On this view, something like Mark Johnston's (1992) revelation about colors is also true of pain.

[16]My proposal can remain neutral on whether pain properties are represented as objective features of one's body parts that could be discovered through third-person observation, or whether they are represented as subjective features that are only accessible first-personally. Intuitions vary, although I suspect that what's driving some intuitions in favor of pain being represented as subjective are particular theories (perhaps folk theories) of pain, rather than features of the experience itself. Thanks to Derek Baker for discussion.

leg as *damaged* as opposed to merely *painful* doesn't automatically confer any additional advantage.[17]

Such an account of pain-experiences is not available to a theorist who identifies mental representation with tracking. On a **tracking theory of mental representation**, mental representation is a species of tracking relation obtaining between a representation and its content (or instances of its content). A tracking theory cannot allow that pain-experiences represent uninstantiated pain properties, since we do not track such properties. This limitation of the tracking theory is an instance of a more general inability to allow for reliable and adaptive misrepresentation, and we should expect similar problems in other cases where the contents we need to attribute to mental states to account for their phenomenal characters involve reliable misrepresentation.

Less obviously, the account of pain I have suggested is difficult to offer on any relation view, views on which mental representation is a matter of being appropriately related to things (usually things in the external world). A relation view will have to say that we are related to either *pain instances* or *pain properties*. Let us consider each possibility in turn. On a relation view on which we are related to pain instances, there will have to be pain instances for us to be related to (otherwise, the view is not a genuine relation view, but a production view in disguise). Unfortunately, realism about pain instances just isn't plausible. But even if you're comfortable with realism about pain instances, you will also have to admit into your ontology instances of all sorts of other properties corresponding to other cases that are similarly troublesome for intentionalism. Such cases might include itches, emotions, gustatory experiences of sweetness, and colors. The more such cases there are, the more costly this strategy becomes.

Let us turn to the view, then, that representation is a matter of a relation to general or abstract properties, as opposed to particular instances of those properties. This view avoids realism about pain instances. However, such a view now owes us a story of how we get to be related to one uninstantiated property rather than another. How does one type of pain-representation (say, a representation of a throbbing-pain) get to be related to the throbbing-pain property rather than a different pain property (say, the stinging-pain property)? One way we can get to be related to properties is in virtue of being related to their instances, but this route is not available here, since the properties in question are uninstantiated.

The production view version of intentionalism avoids these problems. On the production view, unlike on the relation view, what contents can be entertained is not constrained by the objects and properties we can get ourselves related to. So, it is quite possible for mental representations to

[17]In Chapters 4 and 5, I argue that color-experiences, heaviness-experiences and many other experiences are best thought of as case of reliable misrepresentation. In Chapter 6, I argue that in many cases reliable misrepresentation is *more* useful than veridical perception (e.g. in cases where we represent a simpler content but track a more complex property).

represent uninstantiated properties, such as pain properties. And so, the production view allows us to be pure intentionalists by allowing us to attribute uninstantiated properties as the contents of certain experiences.

It may be instructive to consider a case where the challenge from unavailable contents plausibly doesn't arise: the case of triangle-experiences. In the case of experiences of triangles, there is a readily available candidate represented property that can plausibly account for the phenomenal character of those experiences, namely triangularity. Sometimes objects really are triangular, and this triangularity seems to be the right sort of thing to account for the phenomenal character of triangle-experiences. Thus, a view of mental representation on which our triangle-representations represent this property of triangularity (or instances of the property, or particular triangles, etc.) can plausibly account for the phenomenal character of triangle-experiences in terms of their representational content. Compare the case of triangularity to the case of various other experienced qualities: pain, happiness, sweetness, color, hotness, and beauty. It is not so clear whether there is a property that is actually instantiated somewhere in the subject's environment that can account for the felt quality of these experiences. In this section, I focused on the problem as it arises in the case of pain, but similar points may apply to other cases as well. Adopting a production view of mental representation allows us to avoid these problems.

7.3.2 Non-perceptual states

It is not immediately clear what the intentionalist should say about non-perceptual states, such as thoughts, occurrent non-conscious states (e.g. states in early visual processing), and standing non-conscious states (e.g. nonoccurrent beliefs). If we allow that the same representational contents can be had by various of these types of states, then we can fairly easily generate counterexamples to intentionalism. For instance, take the following cases:

(i) A perceptual state representing unique (pure) red

(ii) An occurrent non-conscious state representing unique red

(iii) An occurrent thought representing unique red

(iv) A standing belief about unique red

It seems that such states can share the same representational content but differ in phenomenal character. As things stand, we have one or more counterexample to intentionalism. Such considerations usually motivate a move to impure intentionalism. The usual strategy favored by intentionalists such as Tye is to impose certain further conditions a state must meet in order to qualify as having phenomenal character. These conditions are meant to rule in perceptual states and rule out the other types of states, thereby

	Representational	**Non-Representational**
Phenomenal	Perceptual States	
Non-Phenomenal	Thoughts	
	Standing states	
	Non-conscious states	

Table 7.1: A typical impure intentionalist target outcome.

avoiding this source of counterexamples. The target result of most such efforts is summarized in Table 7.1.

I will first argue that the impure intentionalist strategy is less than fully satisfying. Then, I will argue that that we can deal with these cases without moving to impure intentionalism. I will argue that occurrent thoughts have impoverished representational contents that match and plausibly determine their impoverished phenomenal characters, and that non-conscious states have neither phenomenal character nor representational content. Table 7.2 summarizes my target results.

	Representational	**Non-Representational**
Phenomenal	Perceptual States	
	Thoughts	
Non-Phenomenal		Standing states
		Non-conscious states

Table 7.2: My pure intentionalist target outcome.

The impure intentionalist strategy for dealing with the alleged counterexamples arising from non-perceptual states is to impose certain further criteria that a state must meet in order to qualify as a potential bearer of phenomenal character. Several prominent intentionalists employ such a strategy, including Fred Dretske (1995), Michael Tye (1995, 2000), and William Lycan (1996). The versions of this strategy that I will consider constitute a move to impure intentionalism, since the additional criteria that a state must meet in order to have phenomenal character is not a matter of represented content, but instead other features of the subject or state in question, such as the state's functional role. In this section, I will focus on Tye's account and argue that the extra criteria it invokes (1) do not result in the right distinctions, and (2) are insufficiently motivated. I will briefly suggest that similar points apply to Dretske's account.[18]

[18]The general strategy of imposing further conditions that only perceptual states meet does not automatically lead to impure intentionalism, since these extra conditions might have to do with a state's contents. Suppose all perceptual states represented colors and all thoughts represented shapes. Then we could restrict phenomenal experience to states

On Tye's view, in order for a state R to have phenomenal character it must have representational content and satisfy the following two additional conditions:

(T1) R has nonconceptual content.

(T2) R is poised to affect states with conceptual content. (Tye, 2000, pp. 62-3)

The first condition serves to rule out thoughts and other occurrent conceptual states. The second condition is meant to rule out non-conscious states, such as those in early perceptual processing.

However, both conditions rule out too much. (T1) implies that it is never like anything to think. It is undeniable that the phenomenal character corresponding to a perceptual experience of redness is much richer, more vivid, and more intense than the phenomenal character corresponding to a thought about redness. And it may also be the case that the phenomenal character of thought is not proprietary, where the phenomenal character of thought is **proprietary** just in case there isn't a phenomenal character type that accompanies all instances of what we would intuitively call the same type of thought (e.g. the thought that there is a red chair in the room). However, it is not so obvious that it is not like anything at all to think.[19]

A related problem is that conceptual contents are sometimes involved in what appear to be perceptual states, and we sometimes want to say that they are responsible for certain phenomenal features of those states. For instance, there is a phenomenal difference between seeing the duck-rabbit as a duck versus seeing it as a rabbit. Intuitively, we want to say that the phenomenal difference between the two cases corresponds to the difference in the represented contents *duck* versus *rabbit*. However, such a treatment of these cases is not open to Tye, since these differences amount to differences in conceptual contents.[20,21]

---

representing colors. Since this restriction would be on the basis of the *contents* of those states, we retain the defining feature of pure intentionalism: A state's representational content (alone) determines its phenomenal character. However, I know of no implementation of the general strategy that takes this form. Most theorists want to retain the intuitively plausible claim that thoughts, perceptual experiences, and non-conscious states can represent the same contents. (See Kriegel (2002), who argues in more detail that Tye's strategy must be understood as a form of impure intentionalism.)

[19]Supporters of the phenomenology of thought include Strawson (1994), Siewert (1998), Horgan and Tienson (2002), Pitt (2004), Dumitru (forthcoming), and Husserl (1900a,b).

[20]This point holds even if we think *duck* and *rabbit* aren't contents of perceptual states, but are rather represented by simultaneously occurring thoughts or beliefs. The point I want to make here is that contents that are conceptual according to Tye's own criteria are somehow involved in whatever states are involved in seeing a duck-rabbit as a duck, and that these conceptual contents contribute to the state's phenomenal character. See also fn. 10 of this chapter.

[21]Tye suggests that the phenomenal difference between the two viewings of the duck-rabbit can be attributed to the representation of different shapes (Tye, 2000, p. 61). Although "top-down" influences are known to influence "lower-level" perception in this way,

The present complaint with (T1) is not that it fails to meet its own target of ruling out conceptual states from potentially having phenomenal character, but rather that the choice of target has unfortunate consequences.

Tye's second condition overgenerates in a different way: it fails to meet its own target. (T2) states that in order for a state to have phenomenal character, it must be "poised" to affect conceptual states, such as thoughts, beliefs, and desires. The thought is that since non-conscious states in early perceptual processing are not in a position to affect conceptual states, such as thoughts and occurrent desires, they do not have phenomenal character.

Tye seems to have in mind a view of mental processing that goes something like this:

Figure 7.3: The view of mental processing Tye implicitly assumes.

Surely, he will agree that this picture is too simple. But the important point is that the model he has in mind posits no connection between non-conscious early perceptual states and conceptual representational states. If there were such a connection, those non-conscious early perceptual states would be "poised" to affect conceptual states, and so should be candidate phenomenal states. But this aspect of the model is simply wrong. Non-conscious states affect conceptual states in all sorts of ways.

For example, in a set of experiments involving non-conscious priming, Winkielman et al. (2005) non-consciously primed subjects with pictures of either happy or angry faces. Subjects tried a beverage and were then asked how much they were willing to pay for it and whether they desired more of it. Thirsty subjects who were non-consciously primed with happy faces were on average willing to pay more for the beverage and reported desiring to drink more of it than thirsty subjects non-consciously primed with angry faces. Importantly for us, these effects occurred in the absence of conscious perceptual or emotional effects, and so, we have a case of non-conscious states directly influencing conceptual states. This means that at least these non-conscious states are "poised." Since Tye will allow that these states are also representational and meeting condition (T1), his theory predicts that they should have an accompanying phenomenal character. But they don't.[22]

Perhaps there are ways to finesse Tye's two requirements. Perhaps we can replace (T1) and (T2) with something like the following:

(T1′) (i) in order to have vivid and proprietary phenomenal character, R must have nonconceptual content; (ii) in order to have weak and non-proprietary phenomenal character, R must have conceptual content.

it is implausible that such differences fully account for the striking phenomenal difference in the two cases.

[22]Or do they? See Pitt (Ms) and fn. 30 of this chapter.

(T2′) R must be "poised" to affect conceptual states in a particular way *W*.

While a set of some such conditions might succeed in making the right distinctions, there is a general worry that becomes increasingly troubling the more complicated our extra conditions become: These extra conditions appear to be increasingly unmotivated. In our discussion of pain, I argued that invoking nonconceptual content as a determiner of phenomenal character was unmotivated. Now, (T1′) seems even more unmotivated, since it is not at all clear why the features invoked in (T1′) should play a role in determining a state's phenomenal character. We don't want any old condition that gets the right extension. Rather, we want the condition that plausibly *explains* why some states have vivid phenomenal characters, while others have less vivid phenomenal characters, and still others don't have any phenomenal character at all. Mere extensional correctness is not enough.

Consider (T2′). It is quite plausible that conscious perceptual states affect conceptual states in a way that non-conscious states in early perceptual processing do not; call this way *W*. The idea, then is that affecting conceptual states in a yet unspecified way *W* is required in order for a state to have phenomenal character. With respect to drawing a distinction between conscious and non-conscious states, this is a winning strategy. And that's exactly what's wrong with it. The result amounts to a restatement of one of the differences between conscious and non-conscious states: conscious states affect conceptual states in way *W*, while non-conscious states do not. It does not end up explaining *why* this is the relevant difference, the difference that explains why one group of states is conscious and the other isn't. Again, we don't want any old difference that makes the right division between phenomenal and non-phenomenal states; we want the difference that makes a difference, the difference that explains why some states and not others have phenomenal character.

The upshot is that Tye's impure intentionalist strategy for dealing with non-perceptual states runs up against a particularly acute version of the general problem with impurity: While standard arguments for intentionalism offer reason to think that there is an intimate connection between phenomenal consciousness and mental representation, these arguments do not provide us reason to think extra ingredients are involved. And so, if we need to invoke extra ingredients in order to make the right distinctions between states that have or fail to have phenomenal character, these extra factors have to be motivated in some way. It is not enough that they succeed in making the right distinctions.

Other versions of the general impure intentionalist strategy succumb to similar problems. The rest of this section may be skipped by those who are already convinced.

On Dretske's version, in order for a representational state R to have phenomenal character, it must satisfy the following two conditions:

(D1) R is a systemic representation.

(D2) R is available for calibration by concept-forming mechanisms.

(D1) is meant to rule out conceptual states, while (D2) is meant to rule out states in early perceptual processing. **Systemic** representations are representations that are phylogenetically fixed, or innate. They emerge through the course of normal development without any special learning. Examples include most perceptual representations. **Acquired** representations are representations that are learned, or otherwise acquired, through a process of calibration that operates over systemic representations. Most concepts are examples of acquired representations.

As an illustrative example, consider the case of color-representations. Visual color-representations plausibly emerge through the course of normal development without any special learning, so they are systemic representations. Color concepts are arguably learned, so they are acquired representations. Visual color-representations meet Dretske's criteria, while color concepts do not. Thus visual color-representations meet both conditions, while color concepts meet (D2) but not (D1), and that is why visual color-representations but not color concepts have associated phenomenal characters. Color-representations in early visual processing are also systemic representations and hence meet (D1), but they fail to meet (D2) since they are not available for concept formation. This is why they fail to have phenomenal character.

As with Tye's theory, we might question whether the criteria are successful at drawing their target distinction, a distinction between perceptual states and all other states. There is some evidence that we can improve discrimination of different categories after training on non-consciously presented stimuli (Jessup and O'Doherty, 2009). I'm not sure if this would satisfy Dretske's notion of calibration, but if it or similar cases do, then it seems that non-conscious states are available to the process of calibration. If these non-conscious states also count as representations (and it seems they should on Dretske's view of mental representation, which is a type of tracking theory), then the view would deem these non-conscious states phenomenally conscious. But they're not.[23]

Also as in the case of Tye's view, we might additionally question whether the target distinctions are the right ones to draw. The same worries about phenomenology of thought and other uses of concepts arise here.

And again, supposing that these or some nearby distinctions make the right divisions, we might nonetheless worry that they are unmotivated. Why should whether a representation is phylogenetically fixed play a role in whether it is phenomenally conscious? What is it about availability for calibration that makes a state phenomenally conscious? That these features are relevant

[23]Although I will later argue that such non-conscious states do not have content in the same sense in which perceptual states can be said to have content, the states do have content on Dretske's theory of mental representation, which is a type of tracking theory, so they do pose a problem for his view.

to phenomenal consciousness, rather than merely contingently associated with it, has to be motivated and it is unclear how this can be done.[24]

In the remainder of this section, I address each relevant type of non-perceptual state in turn, and suggest alternative ways of dealing with them open to the pure intentionalist. My discussion will be short and programmatic. Much more remains to be said about all the states I address. My aim is to show that there is a possible pure intentionalist program that avoids the shortcomings of impure intentionalism.

Thoughts and other occurrent conceptual states

Consider a pair of experiences. The first is one of seeing a unique red, uniformly textured wall at arm's length in front of you. The second is one of occurrently thinking (or believing) that there is a unique red, uniformly textured wall at arm's length in front of you. Arguably, these two experiences have the same representational content, but different phenomenal characters.[25]

The impure intentionalist deals with this kind of case by imposing conditions apart from having representational content that a state has to satisfy in order to have phenomenal character. We have already seen that in the case of the duck-rabbit, there is some motivation to attribute phenomenal character to states involving conceptual contents. I claim that we should extend our treatment to the use of concepts in thought. Just as it's like something different to see a duck than it is to see a rabbit, it's likewise like something different to think about a duck than it is to think about a rabbit.[26]

The pure intentionalist, instead, has to account for the phenomenal difference between the experience of the red wall and the thought about the red wall in terms of differences in their content. In Chapter 10, I will argue for a theory of concepts that can be independently motivated and on which the concept of unique red and the perceptual representation of unique red have different contents, as required by pure intentionalism. I will return to pure intentionalism at that point.

Standing states

Non-occurrent conceptual states, also known as **standing states**, allow for a certain sort of counterexample to intentionalism. Compare the occurrent be-

[24]Chalmers (2004b) raises similar complaints to the impure intentionalist project construed as a reductive account of consciousness.

[25]There is room for some disagreement here. For instance, perhaps vision and thought represent distances in a different way or with different specificity. However, I believe the general point that I want to draw attention to is clear: The concept of redness that is used in thought has a very different associated phenomenal character than the visual representation of redness used in visual experience.

[26]That thoughts have phenomenal character has been defended by Strawson (1994), Siewert (1998), Horgan and Tienson (2002), Pitt (2004), and Husserl (1900a,b), and it has been rejected by most everyone else.

lief that the Parthenon is in Athens to the standing belief that the Parthenon is in Athens. The occurrent state has phenomenal character (assuming the claims of the previous subsection are correct), but the standing state does not. And so, we have a counterexample to intentionalism: We have two states that have the same representational content but different phenomenal characters.[27]

The solution I want to recommend is **dispositionalism about standing states**, the view that standing states consist of nothing more than dispositions to have certain occurrent states in certain circumstances, and after more or less time, effort, and processing (see Searle (1991) and, more recently, Kriegel (forthcoming a)). To be clear, everyone agrees that we have the relevant dispositions, and everyone also agrees that these dispositions have a categorical basis. The question is whether we should additionally posit non-conscious representational states representing some, but perhaps not all, of the contents we are disposed to produce. The dispositionalist answers "no."

There are familiar reasons to be a dispositionalist. As Searle (1991, 9. 62–63) argues, if we are not dispositionalists, we immediately face the problem of having to decide *which* standing states to attribute to subjects. The thoughts we are capable of entertaining form a range of cases differing in whether we have ever entertained the thought before, whether and after how much processing the thought is obtainable from a thought we have entertained before, how likely we are to arrive at the thought in various circumstances, and how confidently, strongly, or vehemently we would be of our thought in such circumstances. Any way of dividing the plethora of cases into those for which a subject has a corresponding standing state and those for which she does not seems arbitrary. A non-dispositionalist view of standing states would have to decide whether, half an hour ago, you believed that the Parthenon is more than a meter tall, that there is a human being within a mile of the Parthenon, that a fifth of half the square root of 100 is 1, or that $((A \vee B) \wedge (A \rightarrow C)) \vee (\neg(A \vee B) \vee \neg(A \rightarrow C))$ is a tautology. The advantage of dispositionalism is that it can acknowledge all the above-mentioned variation without having to make any principled distinctions.

A second way to argue for dispositionalism is by appeal to the context-sensitivity of memory retrieval. The consensus is that recalled memory episodes are reconstructed, partly based on cues from the immediate environment, rather than read off of a stored representation. In a classic study, Loftus and Palmer (1974) showed subjects a film of a car accident. Subjects were asked either "About how fast were the cars going when they smashed into each other?" or similar questions using more neutral language, such as

[27]The intentionalist could instead claim that in order to have phenomenal character, a state must be in some sense *activated*. This would be a move to impure intentionalism, but it does not seem as objectionable as the version of impure intentionalism discussed previously.

"About how fast were the cars going when they hit each other?" Subjects who were presented the question using the word "smashed" estimated higher speeds, with a mean estimate of 40.8 miles per hour, than those in the "hit" and other neutral conditions, with the mean estimate of 34.0 miles per hour for "hit." Those in the "smash" condition were also more likely to falsely remember seeing broken glass. This type of effect of new input on memory retrieval is known as the **misinformation effect**.

Hundreds of experiments by Loftus and others have produced similar results.[28] Not only is there a tendency for subjects presented with false information to report what turn out to be false memories, but they also often add their own embellishments.

These results strongly suggest that memory recollection is at least partly a constructive process. Events are not simply recorded in one's head and then replayed on demand. Rather, we use all the resources available to construct plausible reconstructions of what might have happened. We take hints from our current environment, such as the phrasing of certain questions. That is why we are susceptible to the misinformation effect.

The first point related to this data is that it exacerbates the problem of deciding which standing states to attribute to people prior to questioning. Should we attribute to the subjects in Loftus' original car crash study the belief that the car was going at 34 miles per hour or the belief that it was going at 41 miles per hour? And, what is more difficult, what should we say of the broken glass? Prior to questioning, did these subjects believe there was broken glass or not? The dispositionalist can say that all subjects, prior to questioning, have the same disposition to think that the car was traveling at around, say, 34 miles per hour in circumstances C and the disposition to think that the car was traveling at around 41 miles per hour in circumstances C'. There is no further question about what the subjects *really* believed prior to questioning. The same goes for the broken glass.

The second, and related, point concerning this data is that this is not how we would expect memory to behave if there was a hard and fast fact of the matter as to whether or not a subject had a particular standing state. These results help to further erode the picture of the mind as containing a giant receptacle of statements believed, and instead suggests one on which

[28] Here's another one: Loftus and Pickrell (1995) had a subject's older relative recount stories of the subject's childhood. Among several true stories, the relative recounted a false story about the subject being lost in a mall when she was 3 years old. 25% of subjects reported recollection of the event (compare with 68% of subject reporting recollection of events that actually occurred in their childhood). What is more striking is that many of the subjects who falsely recollected being lost in the mall embellished the story with their own details. One subject claimed to remember being lost in the Kay-Bee toy store and being rescued by a man in a blue flannel shirt wearing glasses.

In another experiment, subjects read a Disneyland advertisement asking them to recollect their childhood experiences at Disneyland (Braun et al., 2002). Among other things, the ad describes shaking hands with Bugs Bunny. After reading the ad, 16% of subjects falsely remembered that they had in fact shaken hands with Bugs Bunny at Disneyland. (We know these are false memories, since Bugs Bunny is not a Disney character.)

the mind is a producer of thoughts, estimates, and experiences. While these studies focus on memory, it is not implausible that something similar holds for standing desires, standing beliefs, and other such standing states.

This is all very suggestive, but I think the real reason to be a dispositionalist about standing states is this: Even if there really is a part of your brain whose activation would correspond as required to the occurrent belief that, say, snow is white. This by itself gives us no reason to think that you are representing that snow is white when the state is not activated. The internal state is best likened to non-occurrent perceptual representations, such as a red-representation. We do not think that the state whose content is *red* when activated also has that content when it is not being activated, and I claim we should say the same about non-occurrent conceptual states. Put otherwise, when you're not seeing anything red or imagining anything red, you are not in a state that represents redness. Your red-representation is available to be used to contribute content to occurrent experiences of, say, ripe tomatoes, but if it is not being used at the moment, then it does not currently represent redness. And so it's merely *potentially representational*, just as it's merely *potentially phenomenal*. We should say the same thing about the case of the states out of which occurrent memories and other thoughts are reconstructed: They are merely *potentially representational*, just as they are merely *potentially phenomenal*. They have the capacity to represent various contents when they are activated in certain combinations, but prior to such activations, they do not represent anything.[29]

Non-conscious occurrent states

Non-conscious occurrent states are states that are in some sense active, or used, but that do not have an associated phenomenal character. Such states may be involved in early perceptual processing, as well as in higher-level processing. The challenge to intentionalism arises as soon as we say that occurrent non-conscious states represent, and that they can represent some of the same contents as occurrent conscious states. Consider, for instance, a state in early visual processing representing redness and a state in visual perceptual experience representing redness. These states are are representationally alike (or near enough) but phenomenally very different.

The suggestion here is to deny that such states have content in the same sense of "content" in which conscious mental states have content. Searle

[29]One might suggest that an important difference between the visual representation of redness and the state corresponding to the belief that snow is white is that the latter has propositional form, while the former does not. However, it is not clear why this should make a difference. In the case of occurrent conscious thought, we can entertain both propositional contents (such as when we occurrently think that ice cream is tasty) and non-propositional contents (as when we occurrently desire ice cream, or just think about ice cream). Whether or not a state has a propositional form does not make a difference to whether or not it is representational for occurrent states, so I see no reason to think it should make that kind of difference for standing states.

(1990, 1992) argues that nonconscious states have only "as if" content. A bit more recently, Kriegel (forthcoming a) argues that the of non-conscious states is merely derivative from an ideal interpreter's cognitive phenomenology (this requires that there is a phenomenology of thought; see §7.4.2). Such moves are open to production views. However, on tracking versions of the relation view, it's hard to deny that non-conscious states have representational content in the same relevant way in which conscious states have representational content. They track things in the world no less reliably, and oftentimes more reliably, than conscious states. So it seems that tracking theorists have to say that they represent. Other relation views, depending on what the relevant relation is, may also have to say that such states represent. But production views can deny this. If our minds or brains produce content, then maybe they're just not producing content for non-conscious states.

To motivate the view that non-conscious occurrent states do not represent, let us first consider why we might think that they do represent. One possible reason is that they track things. Retinal states and states in early visual processing track edges, line orientations, and the intensity and wavelength of incoming light. Later states in visual processing track more complex features, such as bars, bars moving in particular directions, and depth. They get to track these complex features because of the way they are wired up to earlier states. And so, perhaps even if we are not tracking theorists, we may want to say that such states represent the features they track.

Another motivation for attributing representational content to these non-conscious states is the manifest content of the states they give rise to. States in early visual processing eventually lead to conscious experiences of color. For that reason, we might be tempted to attribute representations of color or related contents to the early states.

To combine these two motivations, we might want to attribute content to non-conscious states because of certain of their causal or functional roles. These roles include tracking things in the environment and giving rise to conscious experiences.

I want to dissolve the force of these motivations by pointing out that they can yield conflicting results. Take the case of pain. Suppose we have a representation R that tracks property P (say, P is bodily damage of a particular sort). Suppose further that R gives rise to conscious experiences of pain. If we let tracking guide our attribution of content, we should say that R represents P (or a nearby related content). If we let causal role in generating conscious states guide our attribution of content, we should say that R represents *pain* (or a nearby related content). But from §7.3.1, bodily damage is not pain. So we obtain conflicting content attributions for R. Another way to put the point is this: In the chain of processing from bodily damage to pain, there would have to be a switch in representational contents from bodily-damage-related contents to pain-related contents. But at which point in processing does this switch occur? Any choice seems arbitrary. Similar points hold for many other cases of reliable misrepresentation.

What I think we should instead say is that R does not represent anything. R tracks bodily damage, and R gives rise to the conscious experience of pain. R does many other things as well, such as interact with other non-conscious states. But that's the end of the story. There is no further fact as to what R *really* represents.

It is worth emphasizing my points of agreement with theorists who attribute content to non-conscious states: We all agree that non-conscious states play certain functional and causal roles, that they more or less reliably track certain features of the environment according to any of our favorite tracking relations. It might make sense to describe non-conscious states and processes as storing and manipulating information, in the same way that we describe the states and processes of a computer as storing and manipulating information. It might even make sense to evaluate non-conscious states for success on the basis of its processing information, as when we say that a computer program outputs the wrong answer. Out of some or all of these features, we might define a notion of **computational content**. We can all agree that non-conscious states have computational content. The disagreement, instead, has to do with whether non-conscious occurrent states

have the features that endow conscious states with representational content. I claim that the pure intentionalist can say that they do not.[30,31]

One might suggest that computational content just amounts to the same thing as representational content, but this is not obviously the case. Computational content is cheaper than representational content. It can make sense to attribute computational contents to any system with states that mediate in some interesting way between inputs and outputs, such as an elevator, a vending machine, or a computer. And we can attribute computational contents to such systems without thereby committing ourselves to the claim that they *really* represent in the same sense in which our perceptual experiences and thoughts represent. And so computational content does not obviously amount to representational content.

So far, I have suggested that non-conscious states do not have representational content, but that the notion of computational content can do much of the work that the notion of representational content was supposed to do. But one might object that there are theoretical uses of content-like notions that require attributing the same kind of content to both non-conscious and

[30]We can say the same thing about non-conscious states in blindsighters. Blindsighters claim not to see anything, but are able to answer certain types of forced-choice questions about what's in front of them better than chance. One might argue that blindsighters have some mental states with representational content but no phenomenal character. Now, everyone agrees that blindsighters have internal states that to some extent track what's in front of them and that these mental states play some role in generating responses to these forced-choice questions. But all this can be accounted for by appeal to computational content. A feature of blindsight that is sometimes downplayed in philosophical discussions is that despite their better-than-chance performance in forced-choice tests, almost all of their visually-guided behavioral capacities differ remarkably from those of the normally sighted. This very strongly suggests that blindsight is a different kind of thing than normal vision, which makes it quite plausible to say that normal vision involves mental representation, while blindsight does not.

There is another possibility for blindsight, and perhaps even for some non-conscious states. There could be isolated islands of consciousness that are not unified with conscious experiences elsewhere in the brain. Pitt argues for such an account of non-conscious representational states (Pitt, Ms). One reason to resist this kind of account is a commitment to what Daniel Dennett (2004) derisively calls the "Cartesian Theater" conception of consciousness. On the Cartesian Theater conception, there is a particular part of the brain where "it all comes together." It's as if there is a functional analog of a cinema screen somewhere in your head, where conscious experiences are projected. On such a view, it is difficult to allow for isolated islands of consciousness. Either these experiences reach the Cartesian Theater, in which case they are presented alongside other conscious experiences, or they do not reach it, in which case they are not conscious at all. In any case, if we do not have a Cartesian Theater conception of the mind, we should have no qualms allowing for isolated islands of consciousness. Just as your mental states are isolated from mine, some of your mental states could be isolated from others. This might also be a reasonable thing to say about split brain patients. Whether there are isolated islands of consciousness and how we would find out are tricky questions, but for now, I am just mentioning the possibility.

[31]Dulany (1997) and Perruchet and Vinter (2002) argue for the more extreme position that there is no non-conscious representation, and further, that there is no useful place for a notion of computational content.

conscious states, and that that's a reason to think that representational content and computational content amount to the same thing.[32]

This objection neglects that I too can say that conscious states have computational contents. Conscious states also bear causal relations to features of the environment, and they also have internal functional, causal, or computational roles. So whatever my objector thinks we need in order to explain whatever he thinks needs explaining, I can attribute it to both conscious and non-conscious mental states. In order for the objection to succeed, then, it must be that we not only need to attribute content in the same sense of "content" to both conscious and non-conscious states, but also that this sense of "content" is at least sometimes the representational sense. To assess this possibility, we would have to take a closer look at the uses to which attributions of non-conscious contents are put. I will not undertake this project here, but merely flag it as what I take to be the next step in this debate.[33]

It may be worth noting that distinguishing representational content from computational content has the added benefit of liberating computational content from some of the constraints imposed on representational content. For instance, there is a fact of the matter about what a particular representation represents. A representation's content is not open to interpretation, nor is it indeterminate what a representation's content is (although a representation may represent an indeterminate content). Thus, it is desideratum of a theory of mental representation that it yield determinate results on the question of what a mental representation represents. By distinguishing representational content from computational content, we free computational content from any such constraint. There need not be a deep and determinate metaphysical fact of the matter concerning a state's computational content. The disjunction problem and the problem of indeterminacy of interpretation do not arise for computational contents, or at least they lose much of their urgency.[34]

In this subsection, my aim has been to argue that it is not implausible to deny that non-conscious states have representational content, and that a surrogate notion of computational content can do much of the work the notion of representational content is thought to do. If this is correct, then

[32]Such an objector may have in mind a certain "spotlight" model of consciousness, on which mental processing proceeds with or without consciousness, and consciousness merely serves to bring some of these mental processes or their contents to light.

[33]Perhaps the best evidence that we need to attribute representational content to certain non-conscious states comes from non-conscious perceptual priming (see Kouider and Dehaene (2007) for a recent review). However, it is not immediately clear that the data cannot be accounted for by appeal to computational content. This is particularly plausible if we allow the conscious states a non-conscious state gives rise to play a role in our attributions of computational content.

[34]That philosophers of cognitive science need not worry about indeterminacy of interpretation of mental states is a point Robert Cummins has urged. He writes: "We needn't *worry* that we can always trade misrepresentation of x for accurate representation of Something Else; we *can* do that, but it doesn't *matter*." (Cummins, 1994, p. 288, emphasis in original)

the pure intentionalist can avoid a certain source of counterexample to her view.

7.4 Putative counterexamples to the phenomenal intentionality theory

A counterexample to the phenomenal intentionality theory would be a pair of cases phenomenally alike but representationally different. The main sources of such counterexamples arise from the cases of non-conscious states and thoughts. I will argue that with some fairly minimal further claims, the resources we used to defend pure intentionalism also help the phenomenal intentionality theorist in dealing with these cases.

7.4.1 Non-conscious states

According to the phenomenal intentionality theory, in order for a state to have representational content, it must have phenomenal character. Non-conscious states do not have phenomenal character, and so they do not represent. There are two types of non-conscious states that pose a problem for the phenomenal intentionality theory: (1) non-conscious occurrent states, such as states in early visual processing, and (2) standing states, such as non-occurrent beliefs.

Non-conscious states pose a slightly different problem for the phenomenal intentionality theorist than they do for the intentionalist. For intentionalism, the problem is that the possibility of non-conscious and conscious states representing the same contents left her open to counterexamples. For the phenomenal intentionality theorist, the problem is that there might be counterexamples *within* a class of non-conscious states: If non-conscious states represent but have no phenomenal character, then two non-conscious states might represent different contents but be phenomenally alike. For instance, there could be two early visual states, one representing a horizontal line and the other a vertical line that are phenomenally alike. This would be a counterexample to the phenomenal intentionality theory. Likewise, if standing beliefs represent but have no phenomenal character, then we could have two standing beliefs with different contents that are phenomenally alike, which would be another counterexample to the phenomenal intentionality theory. These kinds of cases are not counterexamples to intentionalism, since the intentionalist can allow that some representational contents have no associated phenomenal characters (this is why the intentionalist only faces a problem when she wants to say there can be *other* states representing these same contents, e.g. *horizontal line*, that are in fact conscious).

Although the problem arising from non-conscious states for the phenomenal intentionality theorist is different from the problem arising from such states for the intentionalist, the solutions I've recommended for the intentionalist also solve the problem for the phenomenal intentionality theorist.

In §7.3.2, I argued that the intentionalist should deny that non-conscious states represent. This allows the intentionalist to remain a pure intentionalist. The phenomenal intentionality theorist can also use the same considerations invoked in §7.3.2 to argue that both standing and occurrent non-conscious states do not represent, thereby removing the source of counterexamples. The two non-conscious early visual states are phenomenally alike (they have no phenomenal character), but they are representationally alike as well, as required (they have no representational content). The same goes for the two standing beliefs: They are phenomenally alike (they have no phenomenal character), and they are representationally alike as well (they have no representational content).

7.4.2 Thoughts

The intentionalist needs thoughts and perceptual experiences to differ in a way that can plausibly account for the phenomenal differences of thoughts and percepts that appear to have the same content. I suggested that, if we accept a certain view of concepts that I will argue for in Part III, it is plausible to say that thoughts and percepts have different contents, even in cases where they appear to represent similarly, such as in the case of the experience of a red wall and a thought about a red wall. The phenomenal intentionality theorist faces a different kind of challenge from the case of thought. If the phenomenal intentionality theory is true, then (*Distinctiveness*) must be true as well.

(*Distinctiveness*) Thoughts with different contents have different phenomenal characters.

One way to examine whether thoughts with different contents have different phenomenal characters as well is to induce two mental states that are identical with respect to perceptual content but different with respect to thought content. If the two thoughts differ in phenomenal character, we have a partial confirmation of (*Distinctiveness*).

Terrence Horgan and John Tienson (2002, p. 523) offer the following example:

(*a*) Dogs dogs dog dog dogs.

The first time you read (*a*), you probably won't understand it. Note the phenomenal character of your mental state. Now, recalling that "dog" can be used as a verb, read (*a*) again. Compare your new mental state with the state generated by your first reading. The two mental states differ in phenomenal character.

David Pitt (2004, pp. 27-8) offers additional examples of sentences that can be read with and without understanding:

(*b*) The boy the man the girl saw chased fled. (Center-embedded sentence)

(c) The boat sailed down the river sank. (Garden-path sentence)

(d) The rhodomontade of ululating funambulists is never idoneous.

(A rhodomontade is a rant, to ululate is to howl, a funambulist is a tightrope walker, and to be idoneous is to be appropriate.)

Recall that we want sentences that will generate two different mental states that are alike with respect to perceptual content but different with respect to thought content. (a)-(d) are arguably not such sentences. The two readings differ in sentence structure, and such differences can result in differences in stress, rhythm, and intonation. These can result in differences in quasi-auditory experience (and this is a perceptual difference that has nothing to do with the content of the sentence). This is especially obvious in (a).

Another concern, which also applies to (d), is that the first reading of these sentences produces a feeling of confusion or surprise (and again these experiences have nothing to do with the content of the sentences). Thus, one might respond to these examples by agreeing that there is a difference in phenomenal character but maintaining that this difference corresponds to the accompaniment or non-accompaniment of feelings of confusion, surprise, or even just an experience of understanding in general.

I don't find these objections convincing. The phenomenal character induced by reading these sentences seems to outrun any feeling of surprise, mere understanding, or quasi-audition. But I think there are examples that avoid these objections:

(e) Alice went to the bank.

You probably understood the sentence as stating that Alice went to a financial institution. But there is another reading of the sentence on which she went to a river bank. Now read the sentence again with this second reading in mind.

The two readings of (e) do not differ in structure, comprehensibility, confusion, or surprisingness. These two readings create mental states identical with respect to the visual and quasi-auditory content relating to the sentence itself but different with respect to thought content. And these mental states differ in phenomenal character. So this is a confirming instance of (*Distinctiveness*) in thought. Here's another similar case due to Charles Siewert about a woman who did not get a chance to pass her legal exam (Siewert, 1998, p. 279):

(f) Before she had a chance to pass the bar, she decided to change directions, but she was not so pleasantly surprised with where she wound up.

But we can also understand (f) as describing what happened to a woman initially heading in the direction of a drinking establishment. Again, the

two readings are alike with respect to quasi-auditory and visual content, but differ with respect to thought content.

Some might object to this style of argument by claiming to have visual experiences whenever presented with such sentences, and further that these visual experiences account for the full phenomenology of their experiences. It is difficult to decide what to make of these claims, particularly because visual representation might be constitutively involved in representation in thought, in which case visual phenomenal character and a supervening visual representational content is *required* by the phenomenal intentionality theory in the case in question.

Visual illusions, however, allow for another kind of example that holds perceptual experience constant between two cases, and thus allows us focus on any distinctively cognitive content and phenomenology. Consider White's illusion (Figure 7.4). Upon first seeing White's illusion, one might have the thought that B is darker. After realizing that the effect is illusory, one might view White's illusion while thinking that A and B are the same shade of gray. What it's like to see the picture while thinking that B is darker than A is different from what it's like to see the picture while thinking that A and B are equally dark. Yet the two experiences are perceptually alike (B continues to visually appear darker than A, even though you know it is not). And so we might infer that the difference in phenomenal character is not due to a difference in perceptual content, but rather to a difference in thought content.[35]

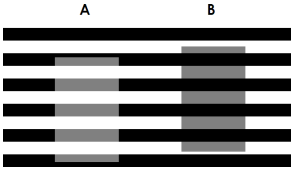

Figure 7.4: White's illusion.

There are other examples like this one. Compare the state of seeing a cup and believing it's actually there versus that of seeing a cup and believing that you're hallucinating (or alternatively, that it's a hologram). Here's another one: Compare the case of Fred who is watching a TV show (say, *The Daily Show* with Jon Steward) while believing it's the news with the case of Ted who is watching the same TV show while believing it's all fiction.

Arguments from sentences affording multiple interpretations do not control possible accompanying perceptual imagery and thus are vulnerable to the objection that such accompanying perceptual imagery is the source of differences in phenomenal character. Arguments from pictures affording

[35]You may have confounding experiences of surprise that interfere with this experiment. Wait a little bit for the feelings of surprise to fade, and try it again. (It will not be difficult to recreate the first thought that A is darker than B, even after you realize that this is an illusion.)

multiple interpretations are susceptible to an analogous kind of objection: one might object that while the above examples control for the content of visual perceptual, they do not control for any accompanying words or sentences. And so, the difference in phenomenal character within each pair of cases might be due to a difference in accompanying verbal representations. Ideally, we want a pair of cases that induce the same verbal imagery *as well as* the same perceptual imagery. If such a pair of cases differ in phenomenal character, then we have some support for (*Distinctiveness*). The following is an attempt to provide such an pair of cases.

Suppose that you do not know what a megagon is. You do know that it is a very-many-sided figure, but you do not know exactly how many sides it has. Thinking that there is a megagon present might involve the word "megagon" and the visual imagery depicted in Figure 7.5. Suppose, now, that I tell you that a megagon is a million-sided figure. If you again have a thought that there is a megagon present, your verbal and visual imagery is the same as the verbal and visual imagery in your previous experience, but the phenomenal character of your experience is different.

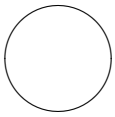

Figure 7.5: A megagon.

Of course, one can always object that additional verbal or perceptual imagery is involved in at least one of the two cases. It is unclear what additional visual imagery might occur in either case. But perhaps one might suggest that in the second experience, there is verbal imagery corresponding to the words "one million sides." However, there is little independent reason to think this might be the case. Consider an everyday example of, say, thinking about a dog. Are verbal descriptions of doghood also running through one's head? This seems implausible. By parity, we should also deny such extensive verbal involvement in the case of the second megagon-related thought.

Of course, the phenomenal intentionality theorist needs *every* difference in representational content between two thoughts to be accompanied by a difference in phenomenal character. On the basis of the arguments above, one might agree that thoughts have phenomenal character, and at least in some cases, thoughts with different contents have different phenomenal characters. However, one might protest that this does not yet show us that this is true of all cases, and hence that it does not yet show us that (*Distinctiveness*) is true. At this point, it is most fruitful to consider potential counterexamples to (*Distinctiveness*). If there is a plausible way to respond to potential counterexamples, we can strengthen the case for (*Distinctiveness*). Such

counterexamples come in two varieties, one requiring on externalism about mental representation, the other open to the internalist.

Externalist objections to (*Distinctiveness*)

Externalism about representational content is the view that the content of at least some representations is partly determined by features of the environment. The familiar Twin Earth (Putnam, 1975) case can give rise to a counterexample to (*Distinctiveness*). Earth and Twin Earth are two planets physically alike except the rivers, streams, seas and oceans on Twin Earth are not filled with H_2O, as they are on Earth, but rather with a substance with a long and complicated chemical formula that we may abbreviate as "XYZ." Prior the the discovery of the chemical composition of the watery stuff, the two planets' histories evolved in parallel.

We are to imagine an Earthling Oscar and his molecule-for-molecule duplicate on Twin Earth, who we may call "Toscar." Suppose Oscar and Toscar think the thoughts they associate with the expression "Water is wet" in their respective languages.[36] Intuitively, Oscar and Toscar are thinking different things. Oscar's thought represents H_2O, whereas Toscar's thought represents XYZ. But the phenomenal character they enjoy is the same.[37]

The objection under consideration has it that Oscar's watery-stuff concept has the content H_2O, while Toscar's watery-stuff concept has the content XYZ. But I see no reason for the defender of the phenomenal intentionality theory to accept this. We initially picked out the notion of mental representation by observing it in our own thoughts and experiences. The paradigm cases of mental representation are these ones that we have this kind of theory-independent access to. But the alleged H_2O/XYZ contents of Oscar and Toscar's mental states are not introspectively accessible,[38] and so they are not paradigm cases of mental representation. So this claim about the contents of these non-paradigm cases is a theoretical claim, and thus to a certain extent up for grabs. This means that the phenomenal intentionality theorist need not accept it. (The fact that some philosophers have strong intuitions in favor of this theoretical claim does not alter its status as a theoretical claim.)

There is much more to be said about Twin Earth cases. One thing that would make my response more convincing is an accompanying explanation of why we have Twin Earth intuitions. I will attempt to provide such an explanation in Chapters 11 and 12, where I will argue that our intuitions involve folk psychological mental state concepts, and that these concepts track something other than the psychological content we are interested in. This allows us to defuse and accommodate the Twin Earth intuition. We

[36] Whether Oscar and Toscar speak the same language is a matter of dispute.

[37] This is assuming that internalism is true of phenomenal character. For a denial of this view, see Lycan (2001).

[38] Attempts have been made to make externalism compatible with introspective knowledge of content, but they are strained. See Burge (1988), Dretske (1995) and Tye (2000) for such attempts.

defuse it by showing that it really provides evidence for something other than the claim it was supposed to provide evidence for. It provides evidence for a claim about folk psychological content, not psychological content. We accommodate it by showing that the claim about folk psychological content is more or less true. For now, suffice it to say that the claim that Oscar and Toscar entertain different psychological contents is a theoretical claim, and that the phenomenal intentionality theory is a theory that denies it.

Internalist-compatible counterexamples to (*Distinctiveness*)

Perhaps the most plausible source of possible counterexamples to (*Distinctiveness*) are cases of theory change. When one changes one's moral theory from, say, a form of moral realism to an expressivist revolutionary reconstruction, one continues to make moral judgments more or less as before. Upon witnessing teenagers torturing a cat, one might still judge that the act is wrong. Further, this judgment does not seem to differ in phenomenal character from similar judgments one made prior to changing one's ethical meta-theory. But, arguably, the contents of one's concepts RIGHT and WRONG have changed, and so the content of the judgment differs from that of similar judgments one made prior to reconstruction. If this is so, then we have a counterexample to (*Distinctiveness*). Compare the thought one would express with the words "Torturing the cat is wrong" as a moral realist to the thought one would express with the same words as an expressivist. The two thoughts are alike in phenomenal character, but differ in representational content.

Here's another similar example: Compare Dimitra and Dimitri. Dimitri's concept of a bachelor is that of an unmarried man. Dimitra has considered the cases of the Pope and men in long-term domestic partnerships and has reconstructed her concept of a bachelor to that of a man available for dating and marriage. When Dimitra and Dimitri think the thought each would express with "Ivan is a bachelor," their thoughts have the same phenomenal character. But, arguably, since they have different concepts of a bachelor, their thoughts have different representational contents.

In Chapter 10, I will argue that concepts have two relevant content-like features. They have content proper, genuine mental content, or *source content*; this is what we've been talking about all along using the term "content." But they also have something I will call *derived content*, which is a matter of actual and possibly entertainable contents related to a representation. On this view, it's quite plausible that the difference between Dimitra and Dimitri's concepts of a bachelor is only a difference in derived content, not a difference in source content. The same goes for the difference between one's moral concepts as a realist and as an expressivist—the difference might be merely in one's derived content. Further, I will argue, the type of content relevant to the phenomenal intentionality theory, and hence to the version of (*Distinctiveness*) that should concern us, is source content, not derived content. And so, these kinds of

examples are not counterexamples to (*Distinctiveness*), since they do not provide us with two thoughts differing in representational content.

7.5 PIIT

I have argued that there are no successful counterexamples to intentionalism and the phenomenal intentionality theory. Absent any other considerations, we can plausibly conclude that representational content just is phenomenal character, that is, that PIIT is true.[39]

The view that emerges is one on which mental representation, as picked out in the way defended in Chapter 2, is type and token identical to phenomenal consciousness. In defending this thesis from counterexamples, I have argued that pain and other bodily sensations represent *sui generis* properties such as pain properties, non-conscious states do not represent, or only potentially represent, and thoughts have phenomenal character corresponding to their representational content.

Finally, although we have not reduced anything to the physical, it is worth noting that we have collapsed two mysteries of the mind into just one. To some, this might seem unappealing. Some prefer a "divide and conquer" strategy to reducing the mental to the physical. However, as the discussion so far has hopefully demonstrated, such an approach is next to hopeless. We end up with tracking theories of mental representation that get the wrong answer for paradigm cases, on the one hand, and utterly mysterious nonrepresentational raw feels, on the other. Better just to stick to the one problem of mind.

[39]One might worry that representation and phenomenology might be contingently related in our world only, and that I have done nothing extra to support a metaphysical or logical supervenience thesis, let alone an identity claim. Apart from the support for an identity theory from the argument in 7.2.1, this is true. However, if we agree that representation and phenomenology exactly overlap, and there is no barrier to their identification, there seems little motivation to insist that they are nonetheless distinct. Its hard to see what better evidence we can get. Nevertheless, there are other arguments that can be made for establishing a modally stronger supervenience claim. Siewert (1998), Horgan and Tienson (2002), Searle (1990, 1992) and McGinn (1988) offer such arguments, mostly relying on thought experiments.

Chapter 8

Naturalism in the Philosophy of Mental Representation

8.1 Introduction

One of the main attractions of certain versions of relation view, in particular tracking versions, is that they are reductionistic. They offer a view on which representation reduces to a physical phenomenon, usually one having to do with causal relations. There are no ghosts, magical relations, or irreducibly mental phenomena. This type of approach is often known as **naturalism**, and is supposed to obtain much of its appeal from the success of the sciences in explaining various phenomena. From this sort of perspective, one might accuse PIIT of not being naturalistic. Unlike some other theories of mental representation, PIIT does not aim to reduce mental representation to something physical. Sure, one might concede, there are problems with these "naturalistic" theories, but at least they are naturalistic, and naturalism is really important. In order to assess this kind of objection to PIIT, we'll have to take a closer look at naturalism. I restrict my focus to naturalism as it applies to the study of mental representation. There are murky issues concerning naturalism in other areas, such as mathematics and logic, that we can avoid for our purposes. The main question of this section is, *What constraints on the study of mental representation are properly motivated by the success of science?*

There are at least two importantly different ways to understand naturalism. One is as a view or approach that endorses a particular ontological constraint on the types of entities that can be invoked in one's theorizing: only physicalistically acceptable entities may be invoked. The other way to understand naturalism is as endorsing a methodological constraint on the appropriate methods of inquiry and the admissible types of evidence that can be used in one's theorizing. For example, one such constraint might be this: when possible, the best way to find out is through the use of empirical methods. **Ontological naturalism** is a philosophical outlook that adheres

to the ontological constraint. **Methodological naturalism** is a philosophical outlook that adheres to the methodological constraint.[1] I will argue that it's methodological naturalism that is best motivated by the success of science. Ontological naturalism may also be to a certain extent supported by science, but not in a way that is of much use to us here. I will argue that my approach is methodologically naturalistic. Reductionistic theories of mental representation are ontologically naturalistic, but they are arguably not methodologically naturalistic.

8.2 Ontological naturalism

In its crudest form, ontological naturalism is a doctrine about what there is. There are physical properties and entities, there is space and time, and there are combinations of these things. There are no ghosts, angels, gods, or vital forces. There are some cases that are less clear, such as abstracta like numbers, functions, and universals. Ontologically naturalistic theories aim to understand certain "problematic" phenomena in terms of the acceptable naturalistic items.

Naturalistic Items	**Non-Naturalistic Items**	**Unclear Items**	**Problematic Items**
Microphysical properties and entities	Ghosts	Abstract entities (numbers, functions)	Consciousness
Space-time	Angels		Mental representation
Causal relations	Gods		Moral properties
	Vital forces		Aesthetic properties
	Ectoplasm		Possible worlds
			Propositions
			Colors

Table 8.1: A reconstruction of the ontological naturalist's classification of entities.

[1] Distinctions of this sort have been drawn by Shapiro (1997), Papineau (2007), Penco et al. (2007), and Horst (2008). Although there are differences in terminology and focus, the distinctions made are quite similar, namely between a more metaphysical or ontological type of thesis about what there is and a more epistemic or methodological type of thesis about how to find out about the world. Since my aim in this chapter is to evaluate the objection to PIIT that it is not a naturalistic theory of mental representation, I will draw the general distinction in the way I see most fit for my purposes. This is not meant to imply that it is somehow a more "natural" way to draw the distinction than alternative ways, nor that this way of drawing the distinction will be useful for other purposes, although both may end up being the case.

Ontologically naturalistic projects can be crudely understood as taking a Problematic Item and reducing it to some combination of Naturalistic Items, or at least showing that it supervenes on Naturalistic Items with the required strength of supervenience (see Table 8.1). As these Problematic Items are "naturalized," they are progressively moved from the "Problematic Items" list to the "Naturalistic Items" list. We are to avoid at all costs appealing to entities from the Non-Naturalistic list. The acceptability of items from the Unclear Items list is an open question.

What justifies our preference for the Naturalistic list? Why should we try to understand Problematic Items in terms of this list as opposed to the Non-Naturalistic list? The answer is supposed to have something to do with science. There are two roles science might be playing here: it might be **delivering an ontology**: it might be telling us that only Naturalistic Items exist. Or it might be **offering examples** of the appropriate types of things to invoke in explanations.

8.2.1 Science as delivering an ontology

We might take science as **delivering an ontology**. Science tells us that the Naturalistic Items exist and that the Non-Naturalistic Items do not exist. If we want to understand some new phenomenon, clearly we should prefer our theory to invoke things that exist rather than things that don't exist!

There are some straightforward worries regarding reading our ontology off of the ontology of science, and in particular fundamental physics. First, it is not clear what these ontological commitments are. Does physics tell us that there are electrons *really*? Perhaps scientists are instrumentalists about electrons and/or some of their alleged features. Or perhaps they should be. Perhaps some kind of structural realism is true: all we can know, or perhaps even all there is, is structure and phenomena. These are important and controversial issues in the philosophy of science. Given how controversial these issues are, it would be unfortunate for an ontologically naturalistic approach to mental representation to presuppose without argument a particular view in the philosophy of science.

Notice that this is not a worry for philosophers of mental representation who make ontological claims but who do not claim to be ontological naturalists of this stripe. They are at best only making claims about what we should be ontologically committed to in philosophy. They do not take their ontological claims to be somehow motivated by the ontological claims of science, so they do not need there to be ontological claims in science. In the first instance, this is only a problem for those who motivate their ontological commitments by the apparent ontological commitments of science.

Still, even assuming a philosophy of science on which science gets to have ontological commitments, it is not exactly clear what these commitments are. Fundamental physics is currently a bit of a mess, ontologically speaking. For one, we have two theories with great predictive powers, the theory of

relativity and quantum mechanics, but they are inconsistent, and we're not sure how to reconcile them. Second, we are not even sure what these theories say about ontology. Arguably, we have trouble understanding just what the theory of relativity says about space and time, and we don't even know how to interpret quantum mechanics.

And finally, even if current physics tells us something about ontology and we can figure out what it is, we don't know what future physics has in store for us. This is point is often acknowledged, but I'm not sure if reductionists about mental representation really take it to heart. So, it is unclear what exactly the constraints on the study of the mind would be, even subscribing to ontological naturalism.

Perhaps ontological naturalists in philosophy of mind are not worrying about any of these problems because they think the phenomena they want to examine are sufficiently "macro" so as to survive most revolutions in fundamental physics. One can pretty safely be an ontological naturalist about chairs while remaining fairly neutral on what to think of most of the various issues raised above. That's because chairs are sufficiently "macro" level objects that can be accounted for independent of the intricacies of fundamental physics. Perhaps mental states will turn out to be like chairs in this respect. However, one might wonder whether this optimism is misplaced. Nothing about ontological naturalism *per se* tells us at what "level of reality" the Problematic Items ought to be reduced. As far as ontological naturalism is concerned, consciousness or mental representation might be reducible to activity in relatively large brain areas, or it might be reducible to microphysical states. Suppose string theory is true. Then another ontologically naturalistic possibility is that consciousness or mental representation is reducible to string states involving other dimensions. There is no reason arising from commitment to ontological naturalism motivating a restriction to relatively "macro" level physical things.

Further, the ontological naturalist may have a misplaced sense of the authority of the Naturalistic Items list in determining our ultimate ontology. As Noam Chomsky argues, prime examples of reductions in science involve modifications of both domains. It would be more accurate, he claims, to think of these examples as examples of **unifications** of different domains, rather than straightforward reductions of one domain to another.

> Large-scale reduction is rare in the history of the sciences. Commonly the more "fundamental" science has had to undergo radical revision for unification to proceed. The case of chemistry and physics is a recent example: Pauling's account of the chemical bond unified the disciplines, but only after the quantum revolution in physics made these steps possible. The unification of much of biology with chemistry a few years later might be regarded as genuine reduction, but that is not common, and has no particular epistemological or other significance: "expansion" of physics to

> incorporate what was known about valence, the periodic table, chemical weights, and so on is no less a valid form of unification. (Chomsky, 2000, pp. 106-7)

It seems that the straightforward reduction that ontological naturalists about the mind are hoping for is uncommon, even within the sciences. Rather, it is more often the case that the "more fundamental" domain must be reformed, oftentimes fairly radically, in order for reduction to be possible; this is why Chomsky prefers the term "unification." As Chomsky also claims in the same paper in which this passage occurs, there is no reason to think that we are in any better shape when it comes to reducing the mental to the physical. A picture on which reductions are better thought of as unifications of independent and equally legitimate domains is a far cry from the optimistic hope that mental phenomena will be reduced to macro-level physics.

To summarize, that we cannot straightforwardly read our ontology off of our scientific theories. First, it's not clear that they provide an ontology in the first place. Second, even if they do provide an ontology, it's not clear what it is. While one way to avoid these worries is to stick to "macro" level physical phenomena, and invoke only those in our theories, it is not clear that this strategy is motivated, even by the ontological naturalist's own lights.

Let us set these unpleasant worries aside for now. There is clearly something right about the motivation for ontological naturalism that we're currently considering: Even if it's unclear what science tells us there is, or what to make of any such claims, it is fairly clear that there are various things it tells us probably do not exist, and that we should therefore avoid appealing to them in our theories. There are no ghosts, vital forces, and other items from the Non-Naturalistic list. (Or, at least, if we do invoke such entities, it should be on the basis of surprising and compelling evidence or arguments in favor of their existence.) This much is right about the approach.

Perhaps it is thought that we should take this same attitude towards irreducible versions of the Problematic Items. For example, one might maintain that it's bad to invoke irreducibly moral properties in your theory, and it's bad for the same reason it's bad to invoke ghosts. However, there is an important difference between the case of Non-Naturalistic Items and at least some of the Problematic Items: For the Non-Naturalistic Items, but not for at least some of the Problematic Items, eliminativism is not an option. It could, and did, turn out that there is no vital force. There is nothing wrong with eliminativism about vital forces. That's because vital forces were an **explanatory posit in a theory explaining something else**. Once a better explanation for that something else came around, we no longer needed them. Once we no longer needed them, we were left with no reason to posit their existence. The situation is importantly different for mental representation and perhaps some other Problematic Items. Mental representation is not just an explanatory posit in a theory that explains something else. Mental representation is itself a phenomenon to be explained.

We have evidence for its existence, and this evidence does not merely derive from our needing to invoke mental representation in a theory about something else. That is why eliminativism is not an option for mental representation like it was for the vital force. The fact that eliminativism is not an option for mental representation means that it is not as bad or unmotivated to postulate primitive representational properties as it is to postulate vital forces.

One might object that mental representation *is* in fact an explanatory posit in a theory that is meant to explain something else, e.g. behavior and dispositions to behavior. This theory invoking mental representation might be folk psychology, or it might be a theory in cognitive science. In Chapter 2, I argued that this is not the best way to approach the problem of mental representation. I will not reiterate my arguments here. Instead, I will just note one reason to think that mental representation isn't *merely* an explanatory posit in a theory, even if you think it is *also* an explanatory posit. Suppose something analogous to what happened to vitalism happens to these theories invoking mental representation. Suppose behavior is best explained by the non-representational properties of brain states. Any representational properties those brain states might have are explanatorily irrelevant concerning the generation of behavior. Then we'd no longer need content in order to explain behavior. But we'd *still* have the problem of mental representation. Mental representation would *still* be a phenomenon in need of an explanation. We'd still notice that we see things, hear things, and think about things, and seek to explain those phenomena. For this reason, eliminativism about mental representation is not an option.[2]

The same can be said about consciousness. Consciousness is not merely an explanatory posit in some theory; it is a phenomenon to be explained. So, eliminativism about consciousness is not an option. Colors are a different case. If our evidence for their existence is debunked then there is no phenomenon to be explained (although there is still the phenomenon of color *vision* to be explained). Perhaps we might still want to invoke colors in a theory that explains something else, such as color vision, but if a better theory comes along that does not invoke colors, we won't need them anymore. So eliminativism about colors *is* an option (see Chapter 5). Perhaps something similar can be said about other mismatch cases.

The point is that there is an important difference between items on the Non-Naturalistic list and possibly irreducible versions of at least some items in the Problematic Items list. The difference is that eliminativism is an option for the former but not for the latter. I claim that this difference is relevant to the appropriate stance to take towards theories that invoke such

[2]How we could know about our own mental contents if they do not have causal powers is an important and difficult question. If we couldn't know of causally impotent mental contents, then this is a good reason to think that they do have causal powers, since it is manifest that we do know of them. Frank Jackson uses arguments of this general form to argue that epiphenomenalist dualism is not a good response to the knowledge argument (Jackson, 2005).

entities. First, this difference means that a view on which, say, representation is irreducible is not as offensive as a view on which there are ghosts. If representation cannot be reduced to something else, then we'll just have to live with that; eliminativism is not an option.

Second, it means that in theorizing about one item on the Problematic Items list, we might be able to invoke another item on that list, as long as it is one of the items for which eliminativism is not an option. This other item will either one day make it to the Naturalistic Items list, or will have be to included as a new primitive entity. Either way, chances are it'll be sticking around. This is good for intentionalists attempting to explain consciousness in terms of mental representation. Either we'll have to reduce consciousness to some combination of Naturalistic Items, or we'll have to admit it as a new primitive. Either way, mental representation exists, and so we can appeal to it in our theories, even though we can't see how it fits in with (other) aspects of the natural world just yet. If what motivates ontological naturalism is the desire to invoke only entities that exist, then we're safe invoking mental representation. The same goes for the phenomenal intentionality theorist who wants to explain mental representation in terms of phenomenal consciousness.

Perhaps these are relatively minor quibbles with ontological naturalism. Perhaps, in the end, everything will turn out to be physical in some sense or other of the word "physical." My main complaint is not over the ontology that is presupposed by the ontological naturalist, but rather over how we are supposed to arrive at our physicalist ontology. It is one thing to eventually want a reduction, but it is another thing to want an immediate reduction. For one, it seems to limit us to relatively "macro" level properties in the reduction base. These are the ones we understand well enough to try to reduce things to. It is not clear why we should presuppose this is the way to go. But more importantly, it ignores the scientific methodologies involved in successful reductions. I will argue that science has shown us that there are no easy reductions. We will reconsider these points in more detail after considering another possible motivation for ontological naturalism.

8.2.2 Science as offering examples of successful explanations

Science offers us *examples* of successful explanations.

Case 1 Animals move. This calls out for explanation. At first, we posited a new primitive, a vital force, to explain the movement of animals. However, after some more empirical investigation, we were able to explain the movement of animals using the same laws and principles that explain the movement of inanimate objects.

Case 2 Organisms are surprisingly well-adapted to their environment. We initially explained this by positing an intelligent designer. However, after sufficient empirical investigation, we were able to explain the phenomenon within biology.

Case 3 Pumps can be used to pump liquids. The initial explanation posited a new law that nature abhors a vacuum. But it turns out that air pressure explains suction in pumps.

These are all cases where we thought we had to posit a new primitive or *sui generis* entity, property, law or principle. But the new primitive ends up being unnecessary, and we can explain the phenomena in question in terms of something else, something that happens to be on the Naturalistic Items list. It turns out that in order to explain most phenomena, we don't have to posit new primitives. So, perhaps the ontological naturalist's suggestion is that we should be weary of positing new primitives in our theories. With our 20/20 hindsight, we can see that these easy solutions invoking Non-Naturalistic Items tend to be false. So, we should avoid them. Instead, we should stick to explanations invoking Naturalistic Items.

Really, what these examples show us is that positing new entities, facts, or properties that are not independently motivated in order to explain something risks being *ad hoc*. It's easy to explain something by positing a new primitive that exactly fits the bill (e.g. a vital force explaining the movement of animals, an intelligent designer that creates well-designed organisms). A "no ghosts" restriction effectively tracks avoidance of ad-hoc-ery. But perhaps it's just avoidance of ad-hoc-ery that's motivated by the cases here.

The second thing to note is that from the epistemic positions of the theorists positing vital forces, intelligent designers, and abhorrences in nature, the correct explanations were in some important way beyond reach. Their false theories might have been the best ones available at the time. The ontological naturalist of the stripe we are now considering focuses on the *results* of non-trivial scientific work sometimes spanning centuries and ignores the processes that led to these results. This ontological naturalist seems to be hoping that, when it comes to the Problematic Items, we can *skip* the tedious journey and get right to the destination. But we neither know where exactly we're going (unless we take science to be delivering an ontology, as discussed in §8.2.1), nor how to get there.

There is a second way in which reductions are often beyond reach. As Chomsky emphasizes, successful reductions (or unifications) oftentimes involve sometimes radical modification of the more fundamental domain. If the tools for such modification are beyond our reach, perhaps because we lack the required mathematics or cannot grasp the relevant concepts, then reduction will not be able to proceed.

So the worry here is not about whether certain kinds of explanations of phenomena are preferable to others. The worry is about how these explanations are obtained. It might be that a reduction (or unification) of mental representation with physics is not possible right now. If that's the case, then it's not an objection to a theory of mental representation that it does not deliver a reduction. We'll return to these points shortly.

8.3 Methodological naturalism

Here is a statement of methodological naturalism by Penelope Maddy:

> [N]aturalism, as I understand it, is not a doctrine, but an approach; not a set of answers, but a way of addressing questions. As such, it can hardly be described in a list of theses: it can only be seen in action! (Maddy, 2001, p. 37)

Methodological naturalism emphasizes the methodology of the sciences, not particular statements or ontological claims that might arise from them. It recommends a scientific perspective, rather than a deference to the pronouncements of science. In cartoon form, it's the view that, very often, the best way to find things out is by checking.

What are the methods of science? Methodological naturalists like Maddy are weary of pronouncing necessary and sufficient conditions for what is to count as scientific methodology (Maddy, 2001, p. 48). However, there are a few quite general things that can be said. In many cases, the best way to find things out is to make observations, where this might or might not require sophisticated experimentation. Now, the suggestion that methodological naturalism should matter for philosophers is not the suggestion that we should go out and perform experiments. As far as our theories are concerned, it does not matter whether we perform the relevant experiments or whether someone else does. And it's difficult to see what kind of data might bear on certain metaphysical theses. Rather, the relevant suggestion here would be that our theories should at least be compatible with the available data. Better yet, when possible, our theories should be driven by the data. They should not be merely consistent with the data, but they should fit well and be motivated by it.

A few words are in order about the "when possible" qualification. I do not want to make grand pronouncements about other areas of philosophy, such as logic, ethics, ontology, epistemology, etc. I leave open the possibility that empirical data is not relevant, or not very relevant, to some questions. For example, perhaps applied ethics is not mainly in the business of *finding things out*, but rather it is in the business of *deciding how to live*, and that's why empirical facts are only indirectly relevant to it. And I do not claim that logical truths can be discovered empirically. I do not need the stronger claim that empirical data is relevant to all of philosophy. The question at hand is what constraints arise from the success of science for a study of mental representation. Mental representation is a subject on which empirical evidence clearly has a bearing. If the relation view is correct, then certain external world entities have to exist. We might be able to check to see if they exist. These aren't *a priori* conceptual truths we're dealing with here. These are claims about how the world happens to be. The world might or might not be that way (e.g. it might or might not contain colors). The methodological

naturalist urges that, since we can, we should go out and check, or if others have checked before us, we should examine their results.

So methodological naturalism recommends that one's theories be compatible with the evidence, and that they be driven by the evidence, when possible. What does it mean for a theory to be driven by the evidence? A theory that is driven by the evidence looks to where the evidence is pointing and goes there. It does not try to fight the evidence. It does not have to contort itself or add epicycles to deal with the evidence. A methodologically naturalistic theory flows from the evidence, not against it.

We can go back and use methodological naturalism to motivate what's right about ontological naturalism. What's right about ontological naturalism is the refusal to invoke ghosts, vital forces, and the like. Now we can see why that's a good idea. First, we have good reason to think that these entities don't exist. But second, we have no good reason to invoke them. If we are invoking them just to help a struggling theory accommodate the data, that's a bad motivation. However, this does not mean that there are no circumstances under which we should not invoke such entities. It could turn out (or could have turned out) that our best evidence supports the existence of spirits, vital forces, or the like. Right now it looks like it doesn't, but things could change, and even if they don't, they could have turned out differently. There is a possible world in which our best science accepts the existence of ghosts.

So, the success of science properly motivates methodological naturalism, and what is right about ontological naturalism is best motivated by methodological naturalism. The success of science derives from its methodology, not from its ontological commitments. It is by being open to evidence-driven change while being weary of epicycles that science has gotten where it is. Along the way, basic ontological commitments have changed drastically. We had to chuck Euclidean space, billiard-ball-like matter, and many resilient ontological intuitions. So, if we are impressed by the success of science, we should be motivated to adopt its methodology and turn to its evidence, rather than accept its ontology and then proceed to reduce all of our concepts *a priori* to the concepts of macro-level physics.

8.4 Naturalism and mental representation

Where does this leave us with mental representation? There are two points I want to make here. First, it is not clear that the so-called naturalistic theories of mental representation are naturalistic in the sense that really matters, the methodological sense. Second, it's not clear that a view that is not naturalistic in the ontological sense is incorrect for that very reason, because it might be that we are not in the appropriate position to perform a reduction right now for one of various possible reasons. This possibility makes sense from the perspective of methodological naturalism.

8.4.1 Going with the evidence

Theories billed as "naturalistic" are perhaps ontologically naturalistic, but oftentimes they are not very methodologically naturalistic. These most notably include tracking theories of mental representation. Tracking theories are arguably ontologically naturalistic.[3] But they predict the wrong results when it comes to certain paradigm mental states. That's the mismatch problem that we discussed in Chapter 5. Tracking theories have trouble explaining what we represent when we represent things as blue, warm, heavy, etc. In Chapter 5, I also explored some possible responses on behalf of the tracking theorist. We might appeal to modes of presentation, qualia, or some such. I argued that these responses are unsuccessful, but now we can say something a little bit stronger. Even if those responses were successful, they would involve contortions and otherwise unmotivated additions to the picture. This suggests that such attempts to deal with mismatch cases are fighting the data, not being driven by the data. Given that mismatch cases arise for paradigm cases of mental representation, that's extra bad news. We don't want our theory to have to struggle to deal with the available evidence, especially the evidence concerning paradigm cases. Similarly, impure intentionalist appeals to nonconceptual content or other extra conditions might be seen as a contortion to help a flailing theory, rather than being independently motivated.

8.4.2 Predicting new evidence

Another lesson we can learn from science is that predicting new evidence is an indicator of a successful theory. I will not make any claims about whether tracking theories predict anything new. In Chapter 10, we will see that the PIIT version of the production view makes some surprising predictions about the content of thoughts and concepts that are actually borne out by independent considerations. The evidence is existing evidence, but surely, when the evidence was discovered does not affect the truth or falsity of the theory. Rather, what's important is that the theory in question was developed independently of the evidence, to capture something else. Then, when it is applied to a new area, it makes predictions about phenomena that it was not originally intended to capture.

8.4.3 Reductionism: The time and the place

The project of reducing the entities in the Problematic Items list to those in the Naturalistic Items list is not clearly related to methodological naturalism. Methodological naturalism does not require one to perform complicated *a priori* or conceptual reductions of one type of thing to things of a particular ontological category. Rather, it directs one to use empirical methods to find

[3] However, see Putnam (1981, 1983).

out what things are like. While this may result in a reduction, it may not. Specifically, a reduction might require further advances or further evidence that are just not currently available. There are no short-cuts to reduction, the methodological naturalist might say. Reductions take hard work.

The problem with mental representation is that the evidence does not drive us all the way to a reduction. We might be in a situation like Aristotle's when it came to pumps. It turns out that the same pump that works at lower altitudes does not work at higher altitudes. This new evidence motivated the air pressure theory of pumps. Accounting for suction in pumps in terms of air pressure explains these differences, since there is less air pressure at higher altitudes. But it's hard to see how Aristotle could have come by that theory without the relevant new evidence.

Similarly, we have some evidence, and it points to a production view, but we might not have enough evidence to perform a reduction to the physical right now. If we try to perform a reduction now in terms of what we've got, we will probably get the wrong answer. That's what the tracking theorists tried to do.

One might object that, unlike Aristotle, at least *we know that there is an ontologically naturalistic explanation out there!* So, while it was acceptable for Aristotle to accept the abhorrence theory of vacuums, it is not acceptable for us to accept a non-ontologically-naturalistic theory of mental representation.

There are two responses to be made here. Let us assume with the objector that mental representation is, at bottom, a physical phenomenon. First, the position I am now recommending is *agnostic* on the issue of reduction. It is agnostic in that it does not say anything for or against the possibility of a reduction. It is not *anti-reductive*, but *a-reductive*. Although it is not an essential part of the theory, we can also make it **optimistic** without changing anything else. It accepts that mental representation is, at bottom, a physical phenomenon, in the broadest sense of "physical," even though it does not claim to give a reductive story.

Second, it is not clear that this additional information that there is a reductive theory somewhere out there is at all helpful in helping us find the right reductive theory. Suppose God told Aristotle "There is an explanation of pumps in terms of forces that explain other phenomena as well!" Would Aristotle have been able to hit upon the air pressure explanation? It seems unlikely. Merely knowing that a naturalistic explanation is out there is not enough to help you hit on the right one. (Perhaps God's pronouncement would have motivated Aristotle to seek new evidence, and this new evidence would have eventually led him to the right theory. But the important point here is that, even given God's pronouncement, he still had to do all the empirical work.) Similarly, it is not clear that merely knowing that there has got to be an ontologically naturalistic explanation out there can help us hit upon the right explanation of mental representation in terms of the physical. If anything, it can motivate us to use whatever means are available to try to find an answer. But we still have to do all the work.

The moral is that there are no short-cuts to reduction. Recent attempts at a reduction by tracking theorists have only led to a struggle to accommodate the data. We're being hasty. We are so eager to provide a reduction that we are giving up on mere empirical adequacy, getting the paradigm cases right. We tried to take a short-cut and it didn't work. In light of these failures, we should not be bothered by a position on which reduction is acknowledged to be currently difficult or beyond our means. And, importantly, we can hold onto all this without letting go of what's important about ontological naturalism: there are no ghosts.

The past few decades of philosophy of mind have been so motivated by the hope of reduction that one might be unsure what else there is for a philosopher of mind to do. I hope this dissertation makes clear that there is plenty to do.

Part III

The Efficient Concept View

Chapter 9

Introduction to Part III

We can entertain lots and lots of contents. These contents are varied and often sophisticated and subtle. We can think about knowledge, Fiji, Barack Obama, unicorns, and counterpossible worlds. This presents a challenge for production views: How do we manage to *produce* all these contentful states? Relation views appear to have an easier time here, at least at first glance. They say that at least some of the contents we can represent exist in the world, and we get to represent them by being appropriately related to them. The world picks up some of the slack, so to speak. I have argued that relation views face difficulties accounting for many paradigm cases. But the point right now is that it initially appears that relation views have an easier time getting us to represent these varied and often sophisticated contents. All we have to do is get appropriately related to them. Production views, in contrast, have to offer a story on which we somehow produce all these contentful states. This is a challenge for any production view, but it is particularly problematic for PIIT. How do we manage to entertain all these contents if entertaining them involves enjoying a particular phenomenal character? Does phenomenal character suffice to allow us to represent sophisticated and complex contents? Are there enough distinct phenomenal characters to account for all the varied contents we can represent?

More precisely, the problem for PIIT is this: According to PIIT, representational content just is phenomenal character. Thoughts obviously have representational content, so, on this view, they must have phenomenal character as well. Further, their phenomenal character must somehow correspond to or match their representational content. Their representational content cannot outrun their phenomenal character. Their phenomenal character must somehow suffice to fix their representational content, and their representational content must somehow suffice to fix their phenomenal character. This is because PIIT claims not only that representational content and phenomenal character are correlated, but also that they are *identical*.

Put otherwise, the case of thought leaves PIIT vulnerable to certain kinds of counterexamples. There are potential counterexamples to intentionalism and hence to PIIT that involve a thought and a perceptual state with the

same representational content but different phenomenal characters. There are also potential counterexamples to the phenomenal intentionality theory and hence to PIIT that involve two thoughts that are phenomenally alike but representationally different.

Here is another related way to put the problem for PIIT with thought: When it comes to perception, PIIT does not look terribly implausible. Having a state with the phenomenal character of blueness seems to be sufficient for representing *blue*, and vice versa: It's plausible to say that a perceptual experience with a certain phenomenal character automatically represents (Horgan and Tienson, 2002; Jackson, 2005). Once you have the blueness feel, you are automatically representing blueness. And it's also plausible that a perceptual representational state representing blueness automatically instantiates the phenomenal character of blueness (setting aside the possibility of nonconscious perception; see Chapter 7, §7.3.2).[1] However, this does not seem to be the case with thought. The phenomenal character of a thought does not seem to suffice to uniquely determine its content, and perhaps the converse is also true: The content of a thought does not seem to suffice to uniquely determine its phenomenal character.

One possible response on behalf of PIIT is to argue that thought really has much more phenomenal character than we might have initially thought. It has enough to uniquely correspond to and determine its representational content. David Pitt has a view like this (Pitt, 2004). However, for reasons mentioned earlier, I find this position implausible. It really does not seem that thought has all that much phenomenal character.

Instead, in Part III, I will argue that thought has much less representational content than we might have initially thought. Thought content is as impoverished as thought phenomenology, so thought is not a counterexample to the correspondence between phenomenal character and representational content required by PIIT. In arguing for this, I will focus on what I take to be the constituents of thoughts: **concepts**. Concepts are vehicles of representation, not contents of representation, that are often involved in thought, but that might also be involved in other states, such as states of perception or imagination.[2] Chapter 10 argues for a view of concepts that I think can solve this problem for PIIT, the **efficient concept view**. On the efficient

[1] Perhaps, like me, you think PIIT is fairly unproblematic for some perceptional contents, such as colors and shapes, but not obviously unproblematic for other contents, such as cats and TVs, that is, for "seeings as" and their analogues in other perceptual modalities. For the purposes of generating the contrast with the case of thought, we can consider only the former types of perceptual contents. The latter, more problematic, type of perceptual contents can be grouped together with thought contents for the purposes of the remaining discussion. The account I give of thought content will apply also to these latter types of perceptual contents.

[2] What I take to be most important about my position can be made compatible with a view on which there are no vehicles of representation, only contents. However, for ease of exposition, I will speak of vehicles of representation: It is controversial what exactly the contents we will be discussing represent, and speaking of vehicles allows us to pick out and individuate contents without having to specify their particular contents.

concept view, concepts and the thoughts involving them have less genuine mental representational content than we may have previously thought, but can be unpacked to yield more complex contents that more closely correspond to what we intuitively take to be their contents. For example, the concept MODAL REALIST does not represent the complex *believer of the view that possible worlds exist in the same way that the actual world exists*, but rather, some more schematic, simple, or superficial content, such as *believes some view about possible worlds*. This is true even though the former complex content determines the intuitive satisfaction conditions for MODAL REALIST and helps determine the intuitive truth conditions of thoughts involving the concept MODAL REALIST, while the latter does not. However, MODAL REALIST can be unpacked to the more complex content, where unpacking may be best understood as non-conscious, non-intentional, physiological process. Further, the results of unpacking are often experienced as further cashings out of concept's content, accounting for the illusion that we were thinking the more complex content all along.

If this is really how concepts work, then it makes sense to distinguish between two types of mental content. The first is the genuine mental representational content we've been discussing all along, which I will now call **source content**. A concept's source content is fairly impoverished. The second type might be called **derived content**. As the name suggests, derived content is content that is somehow derived from source content. For instance, it's by now commonly held that the content of linguistic utterances is derived from the source content of the mental states of speakers, hearers, or the linguistic community. The type of derived content I am interested in, however, is the derived content of mental states.[3] In the second half of Chapter 10, I argue that we can construct notions of derived content (for mental states) out of dispositions to entertain certain source contents. Thanks to relations of unpacking, these notions of derived content have some psychological reality and deserve to be classified as instances of representation. I suggest that these notions of derived content can do some of the work our original notion of mental content was supposed to do.

If all this is right, then the PIIT theorist has an easier time matching up the content of thoughts with their phenomenal character. Since PIIT is supposed to be a view about *source* content, rather than derived content, we only need phenomenal character to match up in the required way to source content. Since both are quite impoverished, as opposed to rich and complex, that such a match-up is possible is much more plausible than on other views of concepts.

Likewise, the production view has an easier time dealing with thought content, since, if the efficient concept view is right, the contents that need to

[3] I borrow the "source content"/"derived content" terminology from Kriegel and Horgan (forthcoming). In the past, I've used "original content" and "intrinsic content" content instead of "source content," borrowing from Searle and Grice, but this led to various misunderstandings.

be produced are fairly impoverished. Indeed, it is hard to see how we can account for these contents on a relation view.

Chapter 11 attempts to sketch a surrogate notion of folk psychological content, the type of content invoked in folk psychological explanations. I suggest that the folk psychological notion of content tracks something closer to derived content than to source content. This has interesting implications for the status of folk psychology and for arguments based on its predictive accuracy. Chapter 12 discusses applications of this notion of folk psychological content to the internalism/externalism debate. Chapter 13 outlines some implications for projects in conceptual analysis.

Before beginning, it's worth mentioning, first, that PIIT is highly suggestive of a view like the efficient concept view. If PIIT is right, and if thoughts have fairly impoverished phenomenal characters, then we should expect them to have fairly impoverished representational contents as well. And this is what we in fact find to be supported by evidence independent of PIIT.

Second, although the efficient concept view, PIIT, and the production view go hand in hand, not very much of Part III presupposes either PIIT or the production view. Although I believe that these views are true of source content, almost nothing I say will hinge on that. If you do not like these views, read the next few chapters as offering a theory of how to get derived content from your favorite kind of source content.

Chapter 10

The Efficient Concept View

10.1 Introduction

The efficient concept view is a view of the structure and content of concepts, the constituents of thoughts. Existing views of concepts are sometimes motivated by highly theoretical, perhaps metaphysical, considerations, such as a favored theory of truth, reference, or mental representation.[1] My approach, instead, takes what I take to be the most basic fact about concepts as a starting point: concepts play a role in mental life. They have a **psychological role**, a role in thought, memory, and phenomenology. I argue for the efficient concept view by arguing that it accounts for the data concerning the psychological role of concepts better than its competitors. This approach leads to a view that not only makes sense of this data, but also has independently plausible implications for other related issues, such as the methodology of conceptual analysis (Chapter 13), the status of folk psychology (Chapter 11), and the paradox of analysis (Chapter 13, §13.4). To the extent to which these implications are independently plausible, this further supports both the theory and the approach it stems from.

On the **efficient concept view**, the concepts used in thought are structurally fairly simple and have fairly simple contents, but they can be unpacked to yield more complex contents that we take to be further cashings out of the simpler contents. In virtue of this, we can say that concepts derivatively represent these complex contents. For example, the concept MODAL REALISM does not represent or contain parts representing all of the content that is generally associated with the concept, say, *the view that possible worlds exist in the same way that the actual world exists*. Rather, MODAL REALISM represents fairly simple contents that can be used in thought, reasoning, and the generation of behavior *in place of* this more complex associated content. Parts of this associated content can be retrieved when called for by the task

[1] For example, Jerry Fodor's theory of concepts is at least partly motivated by a certain view on what a naturalistic theory of mental content (but perhaps not mental *concepts*) must be (see Fodor (1987, Ch. 4)).

at hand, and when they are, we take them to be further cashings out of what we were thinking all along in using the concept MODAL REALISM. In virtue of this, we can say that MODAL REALISM *derivatively* represents those contents. The result is a view on which concept acquisition and possession make thinking more efficient, requiring the use of only the concepts and contents needed at a given moment, while allowing related concepts and contents to be readily available in case they are subsequently needed.

10.2 For the efficient concept view

The efficient concept view is a view about the complexity of concepts and their contents, and the role of unpacking and cashing out relations in thought. I begin by focusing on the issue of complexity. To get an intuitive grip on the notions of complexity in play, consider the relatively uncontroversial case of the thoughts that are normally induced by reading the following two sentences:

(*Beer*) There is beer in the fridge.

(*Beer and Wine*) There is beer in the fridge and wine in the cupboard.

The thought induced by (*Beer and Wine*) is more complex than the thought induced by (*Beer*) in two ways. First, it involves more, or more complex, representational states than the thought induced by (*Beer*), and second, it "says," or represents, more than the thought induced by (*Beer*). According to the efficient concept view, concepts are simpler than many theorists have thought in both of these ways.

To specify these two ways in which two thoughts and concepts can differ in complexity more precisely, let us introduce a distinction between the vehicles and the contents of representation. A **vehicle of representation** is the thing itself that does the representing, the thing that "says" something. Examples of vehicles of representation include mental states, words and signs. My discussion focuses on vehicles of *mental* representation, the things that do the mental representing. These include concepts, thoughts, and on many views, percepts. My discussion is neutral with respect to what types of things the vehicles of mental representation are. They might be brain states, functional states, symbols in a language of thought, or even states of an immaterial soul. **Concepts**, as I will use the term, are vehicles of mental representation that are constituents of, or literally parts of, thoughts (they might be constituents of certain perceptual, imaginative states, or other states such as standing mental states, as well). **Thoughts** are the states we are in or the processes we undergo when we do what we commonly call "thinking." In other words, as I am using the terms, thoughts are a type of

occurrent psychological state, that is, a state that is undergone or in some way activated or experienced.[2]

What a thought, concept, or other vehicle of representation is about, or what it **represents**, is its **content**. We can pick out the notion of content ostensively, as discussed in Chapter 2: We notice that we see things, hear things, and think things. What we see, hear, and think are the contents of our visual experiences, auditory experiences, and thoughts, respectively. As far as possible, I will remain neutral on what type of thing mental contents are. They might be worldly objects and properties, sense data, abstracta, or something else entirely.

Although I intend to stay fairly neutral with respect to the ontological types of concepts and their contents, the way I am using the terms, concepts and mental contents are **psychologically involved**, or have a **psychological role**: They are involved in reasoning, the generation of behavior, phenomenology, and other aspects of inner life. Psychological involvement is left open to include various possible *types* of roles concepts and mental contents might have: They might be causally efficacious, that is, they might *cause* mental states, behaviors, and phenomenology, as may be the case with concepts. Or they might merely reliably accompany them, as may be the case with mental contents if we think they are causally impotent. It is important to note that psychological involvement does not require being literally in the head; contents might be external world objects or abstracta that we are related to, as long as they have a psychological role.[3]

We can now specify the two previously described notions of complexity more precisely: A **representational vehicle is complex** to the extent to which it has proper parts that are also representations. For instance, if the concept WHITE CAT has a proper part that represents *cat*, then the it is complex. If this part representing *cat* in turn has a proper part representing *mammal*, then WHITE CAT is still more complex.

A **representational content is complex** to the extent to which it has proper parts that are also contents. More precisely, if a content c that is represented by a representation R has proper parts that might be represented by vehicles distinct from R, then c is a complex content.[4] For instance, *white*

[2]A note is in order about how I individuate concepts. Common ways of individuating concepts are by their contents or their contribution to the truth conditions of thoughts. Eventually, I will argue that concepts such as MODAL REALIST do not represent everything we think they do, nor do they contribute the truth conditions that we think they do, except in a derivative sense. For this reason, I do not use either criterion for individuating concepts. Instead, for now, I go with intuitive usage: When you have a thought that would be expressed by the sentence "There is a table in the room," we are using the concept TABLE. This is so even if the concept does not have the contents we intuitively take it to have.

[3]If, on your view, mental states have more than one semantic feature (e.g. a sense and a reference), take the notion of content to denote whichever one is psychologically involved.

[4]There is a question about how concepts and contents compose or combine to form more complex concepts and contents. An account of how this happens would shed further

might be a simple content, but *cat* might be complex, perhaps if it has *animal* as a part.

Returning to our initial example, we can say that the thought induced by (*Beer and Wine*) is more complex than the thought induced by (*Beer*) in the following two ways: First, the thought induced by (*Beer and Wine*) involves more vehicles of representation than the thought induced by (*Beer*): while both thoughts involve the concepts BEER and REFRIGERATOR, only (*Beer and Wine*) also involves the concepts WINE and CUPBOARD. Second, the thought induced by (*Beer and Wine*) involves more or more complex contents than the thought induced by (*Beer*): while both thoughts represent the contents *beer* and *refrigerator*, only the thought induced by (*Beer and Wine*) additionally represents the contents *wine* and *cupboard*.

I claim that thoughts and the concepts they involve are relatively simple with respect to both their contents and their vehicles. The locus of disagreement with other views of concepts can be illustrated with an example where, unlike in the case of the thoughts induced by (*Beer*) and (*Beer and Wine*), different views disagree on the relative complexity of the vehicles and contents involved. Consider the thoughts induced by the following two sentences:

(*Modal Realist 1*) John is a modal realist.

(*Modal Realist 2*) John believes that possible worlds exist in the same way that the actual world exists.

There is room for disagreement with respect to whether these two thoughts differ in complexity of contents and vehicles. **Molecularism** is the view that apparently complex concepts such as MODAL REALIST really are complex; they are composed of less complex concepts such as POSSIBLE WORLDS, EXISTENCE, etc., and that their contents are a function of the contents of their constituents. According to molecularism, then, the thoughts induced by (*Modal Realist 1*) and (*Modal Realist 2*) involve the same vehicles of representation and have more or less the same contents.[5]

On a different view, the **complex content view**, the two thoughts induced by (*Modal Realist 1*) and (*Modal Realist 2*) involve different vehicles of representation, but both thoughts nevertheless have the same content. D. O. Hebb (1949), associationist psychologist Wayne Wickelgren (1992), and more prominently in philosophy, Jerry Fodor (1998) have endorsed versions of this view, or relevantly similar views.[6] On the complex content view,

light on many of the issues discussed in this paper. However, for the purposes of this paper, I sidestep these issues.

[5]For a recent defense of molecularism, see Pitt (1999).

[6]Fodor's **conceptual atomism** is non-specific between the complex content view and a view sharing certain aspects with the efficient concept view concerning the content of concepts; this is because conceptual atomism is a view about *concepts* (the vehicles of representation), and not a view about *contents*. However, conceptual atomism together with Fodor's version of tracking theory commits one to the complex content view, since the properties and objects that our concepts are informationally related to are oftentimes very

the thought induced by (*Modal Realist 1*) does not involve all the concepts involved in the thought induced by (*Modal Realist 2*). For instance, the thought induced by (*Modal Realist 2*), but not that induced by (*Modal Realist 1*), involves the concept EXISTENCE. Still, the complex content theorist claims that the two thoughts have the same content. The complex content theorist, unlike the molecularist, is happy to allow for the complexity of contents to drastically come apart from the complexity of their vehicles. She might make a comparison with non-mental representations, such as maps, for which relatively simple vehicles of representation (e.g. a triangle) can represent more complex contents (e.g. *a camp site*). Similarly, she might say, simple vehicles of *mental* representation can represent complex contents.

The third view I will discuss, which is also the view I will defend, is the **efficient concept view**. On this new view, the thoughts induced by (*Modal Realist 1*) and (*Modal Realist 2*) differ both in the complexity of their vehicles (as the complex content view maintains), *and* in the complexity of their contents (as neither opposing view maintains). The efficient concept view agrees with molecularism that complex contents tend to be associated with complex vehicles, and it agrees with the complex content view that concepts are relatively simple vehicles of representation. The result is a view on which concepts and hence the thoughts involving them are less complex with respect to both vehicles and contents than we may have initially thought. The efficient concept view adds to this picture of the complexity of concepts and contents a claim about the importance of unpacking and experiences of cashing out, which I will describe in more detail later.[7]

We can put the locus of disagreement between these views more concretely by introducing a new notion, that of **associated content**. Whatever we say about what concepts represent, we can all agree that concepts have various contents that are closely associated with them. For instance, regardless of whether we think that the concept MODAL REALISM represents *the view that possible worlds exist in the same way that the actual world exists*, we can all agree that it is closely associated with that content; this closely associated content is what I call the concept's "associated content." One way to frame the main question of this chapter, then, is this: *What is the relationship between a concept and its associated content?* According to molecularism and the complex content view, but not the efficient concept view, concepts represent their complex associated contents. More precisely, the efficient concept view accepts, while the other two views reject, the following claim:

---

complex. Even the property our concept RED is informationally related to is a complex surface reflectance property, or perhaps a dispositional property to cause certain reactions in us. (I suppose the conceptual atomist who endorses an informational semantics can avoid commitment to the complex content view by saying that all the properties we are informationally related to match our concepts in complexity.)

[7]Wickelgren (1992) also emphasizes the role of unpacking but not that of experiences of cashing out. In conversation, Chris Viger has described to me a related and as yet unpublished view of his, the **acquired language of thought hypothesis**, on which concepts represent words and unpack into other concepts and perceptual representations.

(*Simpler Contents*) (Many) concepts don't represent their associated contents; rather, they represent contents that are distinct from and less complex than these contents.

A caveat is in order: There is room for disagreement about what exactly a given concept's associated content is. For instance, on some views, BACHELOR's associated content is *unmarried man*, while on other views, it might be *man available for marriage* (so as to exclude the Pope and men in long-term domestic partnerships from counting as bachelors). There is also room for more fundamental disagreement with respect to associated content: Implicit in the above two views of the associated content of BACHELOR is the view that associated contents are definitions specifying necessary and sufficient conditions for category membership. However, over the past few decades, work in psychology has motivated views on which the relevant associated content of concepts is a prototype depiction (e.g. of a prototypical bachelor) (Rosch, 1975; Rosch and Mervis, 1975), a set of exemplars (e.g. particular instances or types of bachelors) (Medin and Schaffer, 1978), or a part of or an entity in a theory (e.g. an entity defined in terms of its role in a folk theory of marriage) (Murphy and Medin, 1985). On these views, the associated content of BACHELOR would be a prototype, exemplar, or part of a theory, respectively. However, these disagreements are orthogonal to the questions I am asking about the complexity of concepts. For each of these allegedly importantly related contents, we can ask whether concepts really represent them, or if they are merely related to them in some other interesting way or in no interesting way at all. Put slightly differently, representing these allegedly important contents are all different ways in which concepts can be fairly complex, but it is the question of complexity that is at issue, not the particular ways in which concepts might be complex. For this reason, the notion of associated content is deliberately left vague so as to cover all these possibilities. You can think of the notion of associated content as a *placeholder* for the type of complex content that is importantly related to concepts according to your favorite of these or similar alternative views.

We can also introduce the notion of a concept's **associated vehicle**. Intuitively, a concept's associated vehicle is the vehicle of representation that explicitly represents that concept's associated content. More precisely, a concept's associated vehicle is the vehicle of representation that has a part representing every part of a concept's associated content. Molecularists think that concepts just are their associated vehicles, while complex content theorists and efficient concept theorists deny this.

The three different views can agree on what a concept's associated vehicle is, for instance that VIXEN's associated vehicle is FEMALE FOX. The disagreement is over whether concepts are identical to their associated vehicles, for instance, whether VIXEN is one and the same vehicle of representation as FEMALE FOX. The molecularist maintains that concepts are identical to

their associated vehicles, while the other two views deny this.[8] In other words, the efficient concept view and the complex content view accept, while molecularism rejects, the following claim:

(*Simpler Vehicles*) (Many) concepts are distinct from and less complex than their associated vehicles.

It may be helpful to provide an illustrative example. Suppose that we agree that the associated content of the concept VIXEN is *female fox*. The efficient concept view accepts both (*Simpler Vehicles*) and (*Simpler Contents*). So, it holds that VIXEN is a simpler vehicle than, and hence also not composed of the same vehicles as, FEMALE FOX, and that *vixen* is a simpler content than, and hence also not in some way composed of the same contents as *female fox*. Molecularism rejects both claims, so the molecularist will maintain that VIXEN and FEMALE FOX are composed of the same vehicles and have the same content. The complex content view accepts (*Simpler Vehicles*), but rejects (*Simpler Contents*); it maintains that VIXEN and FEMALE FOX are distinct concepts, with FEMALE FOX exhibiting more complexity than VIXEN, but that they share the content *female fox*. Figure 10.1 is a simplified depiction of the concept VIXEN on the three views.[9,10]

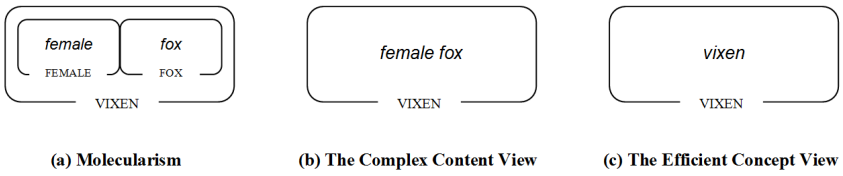

Figure 10.1: The concept VIXEN according to (a) molecularism, (b) the complex content view, and (c) the efficient concept view. Concepts are represented by rounded rectangles with their names appearing on the bottom edge in small caps and the names of their contents appearing inside the rectangles in slanted text. This diagram does not depict the important role of cashing out relations on the efficient concept view (see Figure 10.4 for that).

As I've already noted, the efficient concept view adds to this picture a claim about the importance of unpacking.

[8]The three views also disagree on whether subjects must possess all their concepts' associated vehicles; molecularism affirms this, while the other two views deny it.

[9]Molecularists who think that FEMALE or FOX are themselves complex concepts would add more embedded rectangles to the diagram. Likewise, different versions of the complex content view might be best represented using slightly different diagrams.

[10]There are other possible views. For instance, one might accept both (*Simpler Vehicles*) and (*Simpler Contents*), while denying or staying neutral on the relevance of unpacking to conceptual representation. It is also possible to be a complex content theorist while attributing a prominent role to unpacking. This may be the best way to interpret Wickelgren. I specify the efficient concept view's competitors in the way that I do in order to make explicit the various theoretical choices that one can make in order to arrive at the efficient concept view in a way consonant with the arguments of this paper.

(*Unpacking*) Concepts can be unpacked into their associated vehicles.

Unpacking has to do with the retrieval of more complex content or information from less complex vehicles. I'll say a bit more about unpacking shortly.

Before proceeding, I should note that the notions of associated content and associated vehicle play a dual role in my discussion. First, they help precisify the locus of disagreement between the various views. Second, they play an important role in the efficient concept view, on which concepts *unpack* into their associated vehicles and *derivatively represent* their associated contents. We will also soon see that on the efficient concept view, various types of associated contents and associated vehicles can peacefully co-exist, which is another reason I leave the notions vague so as to cover different possible views of the contents importantly related to concepts.

In the rest of §10.2, I argue for the efficient concept view by arguing first for (*Simpler Vehicles*) and then for (*Simpler Contents*). §10.2.1 argues for (*Simpler Vehicles*), §§10.2.1 and 10.2.2 argue for (*Unpacking*), and §10.2.3 argues for (*Simpler Contents*). My claim is that the efficient concept view allows us to best explain various observations concerning the psychological role of concepts and contents. Concepts and contents behave in certain ways when they are used in thoughts and this allows us make inferences about their structure and content. In §10.3, I expand on the role of unpacking and cashing out.

10.2.1 For (*Simpler Vehicles*)

There are various possible lines of argument for (*Simpler Vehicles*). Wickelgren argues for the claim using optimality considerations. First, having concepts that conform to (*Simpler Vehicles*) would be the most efficient way to represent thoughts that would otherwise be too complex to entertain given our limited attention span (Wickelgren, 1992, pp. 21-2). Second, having concepts that conform to (*Simpler Vehicles*) would be the most successful way to minimize associative interference between thoughts that share parts given certain other constraints (Wickelgren, 1979, p. 45). Assuming that our representational system efficiently meets the aforementioned challenges and that there are no countervailing constraints, we can conclude that (*Simpler Vehicles*) is in fact true of us. While I endorse this line of argument, it introduces additional complexities that I prefer to avoid for present purposes.[11]

Fodor mainly argues for (*Simpler Vehicles*) by arguing against various views on which concepts are complex, such as definitional views, prototype views, and theory theories (Fodor, 1995, 1998, 2008). While there is something

[11]For instance, to compare the efficiency of representing complex associated contents the way the efficient concept view or the complex content view suggest to the way molecularism suggests, we should compare the cost of having more representational vehicles, as is likely to be the case on the first two views, to the cost of activating more vehicles in any given thought, as on molecularism.

close to a consensus in the mind sciences that (at least most) concepts do not represent definitions,[12] Fodor's other arguments are a matter of some controversy.

I think that there is an easier route to (*Simpler Vehicles*) that is sufficient for our present purposes. This route draws on a line of argument Fodor mentions while arguing against views on which concepts represent definitions:

> Does anybody [...] really think that thinking BACHELOR is harder than thinking UNMARRIED? Or that thinking FATHER is harder than thinking PARENT? Whenever definition is by genus and species, definitional theories perforce predict that concepts for the former ought to be easier to think than concepts for the latter. (Fodor, 1998, p. 46)

What I want to take from Fodor's suggestion is the idea that the relative difficulty of two thoughts can tell us something about their relative complexity. In §10.2.1, I expand on certain related lines of thought. I argue that observable differences in difficulty between different thoughts readily support (*Simpler Vehicles*). These observations cut against not only views rejecting (*Simpler Vehicles*) on which concepts represent definitions, but any view rejecting (*Simpler Vehicles*). In §10.2.1, I argue that results on memory research further support this view of concepts.[13]

Differences in the difficulty of thinking thoughts

I will argue that certain observations about the relative difficulty of thinking various thoughts are best explained by views accepting (*Simpler Vehicles*).

My first claim, which should be relatively uncontroversial, is that **some thoughts are more difficult to think than others**. Difference in the difficulty of thoughts can be operationalized in various ways: Distracting tasks are more likely to prevent you from thinking a more difficult thought than from thinking an easier thought; thinking a more difficult thought is more likely to distract you and impair your performance on other tasks than thinking a less difficult thought; you are more likely to fail at producing a more difficult thought than an easier thought. For now, though, the notion of a difference in difficulty can remain a fairly intuitive notion.

My second claim, which should also be fairly uncontroversial, is that **differences in the complexity of thoughts give rise to differences in difficulty**. For instance, the thought induced by (*Beer and Wine*) is more difficult to think than that induced by (*Beer*). You are more likely to fail at thinking the second thought, especially if you are otherwise occupied or distracted, and thinking the second thought is more likely to distract you

[12] See Smith and Medin (1981) and, more recently, Murphy (2004) for reviews.

[13] My arguments depend on observing certain states in ourselves. I think this is sufficient here, but it would be nice to have controlled experiments substantiating my conclusions. An experiment that would directly address these issues would compare the way two concepts with the same associated content behaved in thought. If such concepts behaved differently, we would have reason to think they are in fact distinct concepts.

from other tasks. It is not unreasonable to at least partly attribute this difference in difficulty to the two thoughts' difference in complexity with respect to vehicles.

It is important to note that I am concerned with the differences in difficulty of thinking particular thoughts, not the differences in difficulty of the (say, linguistic) processing that leads to those thoughts, or the differences in difficulty of thinking subsequent thoughts. For example, reading and understanding the sentence (*Beer*) viewed in a mirror is likely to be more difficult than reading and understanding it as it appears in this text, but this is not a difference in the difficulty of entertaining the thought *there is beer in the refrigerator*. Likewise, if the thought induced by (*Beer and Wine*) but not the thought induced by (*Beer*) gives rise to subsequent thoughts (perhaps one remembers an embarrassing incident involving the consumption of wine), this is a difference in the difficulty of the subsequent thoughts, but not a difference in the difficulty of the initial thoughts induced by (*Beer*) and (*Beer and Wine*).

One might worry that it can be difficult to tell these various types of difficulty apart. In the case of the thoughts generated by (*Beer*) and (*Beer and Wine*), the difference in linguistic complexity between the two displayed sentences probably also contributes to the difference in difficulty between the two total experiences. However, we can argue that it is very plausibly not the sole contributor. Imagine a scenario in which the two thoughts in question are generated as relatively spontaneous judgments, rather than by means of reading a displayed sentence. Suppose you are in your house wondering whether there are any alcoholic beverages around. Then, in one case, you think to yourself *there is beer in the refrigerator*. In the second case, you think to yourself *there is beer in the refrigerator and wine in the cupboard*. The thought you had in the second case is more difficult to think than the thought you had in the first case. You are more likely to fail at thinking it if you are distracted or in some way incapacitated, and it is more likely to distract you from other tasks.[14]

We are now ready to move on to the argument for (*Simpler Vehicles*). I claim that pairs of sentences such as (*Modal Realist 1*) and (*Modal Realist 2*) induce thoughts that differ in difficulty, and that this difference in difficulty is best explained by a difference in the complexity of the two thoughts. Consider, as a more drastic and hopefully more obvious case, the thoughts induced by the following two sentences:

(*Philosophers*$_{hard}$) Annie believes that possible worlds exist in the same way that the actual world exists, Benny believes that the right act is the one that maximizes expected utility, and Cathy believes that the content of a mental state is at least partly determined by facts about the subject's environment.

[14]Thanks to David Pitt and Jack Spencer for discussion on the factors that can influence difficulty.

(*Philosophers*$_{easy}$) Annie is a modal realist, Benny is an act utilitarian, and Cathy is an externalist about mental content.

Assuming you know what modal realism, act utilitarianism, and externalism about content are, the two thoughts have the same associated content.[15] But the first thought was more difficult to think than the second thought.

We can get at the difference in difficulty in at least two different ways. First, we can compare the two thoughts that you had just now. Second, we can compare your case to the case of **Eugene the undergraduate freshman**, who lacks the concepts MODAL REALISM, ACT UTILITARIANISM, and EXTERNALISM ABOUT CONTENT, but has the concepts POSSIBLE WORLDS, UTILITY, and MENTAL CONTENT. Eugene is only able to think the first thought, and it is more difficult for him to do so than it is for you to think the second thought.

If this is right, then we have a case of two thoughts with the same associated content that nonetheless differ in difficulty. One explanation of this difference in difficulty is that the easier thought is less complex with respect to vehicles than the more difficult thought. When it comes to vehicles of representation, you are literally thinking less when you think the easier thought. This sort of explanation is open to the efficient concept view and the complex content view, but not to molecularism, since it rejects (*Simpler Vehicles*).

Before considering some alternative possible explanations of this difference in difficulty that are open to the molecularist, let us consider a related phenomenon: Thoughts whose associated vehicles differ in complexity need not differ in difficulty. The following sentences generate such a pair of thoughts:

(*Vegetarian*) John is a vegetarian.

(*Utilitarian*) John is a utilitarian.

The observation here is this: It is not more difficult for us to think the thought invoked by (*Utilitarian*) than it is to think the thought invoked by (*Vegetarian*), even though the first thought has a more complex associated vehicle than the second. But if molecularism is true, then we should expect it to be more difficult to think the thought invoked by (*Utilitarian*), since, regardless of what other factors contribute to difficulty, the sheer amount of representational vehicles involved should result in some difference. In other words, molecularism falsely predicts that the two thoughts differ in difficulty. Figure 10.2 illustrates the differences in the complexity of vehicles of representation that are activated when thinking the two thoughts on the three different views.

Let us now consider some responses on behalf of the molecularist. I said that factors other than complexity can affect the total difficulty of an

[15] If you disagree with this, change the example as you see fit.

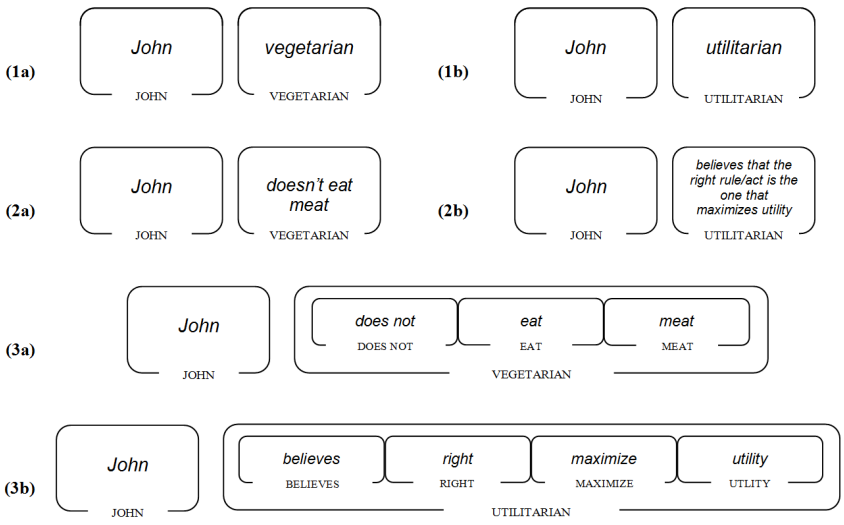

Figure 10.2: The concepts involved in the thought induced by (*Vegetarian*) on the efficient concept view (1a), the complex content view (2a), and molecularism (3a). The concepts involved in the thought induced by (*Utilitarian*) on the efficient concept view (1b), the complex content view (2b), and molecularism (3b). On molecularism, but not on the efficient concept view or the complex content view, additional complexity in associated content results in additional complexity in the vehicles of representation that must be activated in order to use a concept.

experience. One such factor is the amount of linguistic processing required in order to generate the thought in question from a displayed sentence. The molecularist might suggest, then, that the observed difference in difficulty between the thoughts produced by ($Philosophers_{easy}$) and ($Philosophers_{hard}$) is explained by the different amount of linguistic processing involved. We can agree that such differences do indeed result in differences in difficulty of the total experience. However, I claim that they do not account for the difference in difficulty of the thoughts themselves. As in the case of the thoughts induced by (*Beer*) and (*Beer and Wine*), we can consider a situation in which the thoughts in question are generated absent the linguistic processing involved in reading or hearing sentences. Suppose Jane explains her views to Eugene, the undergraduate freshman, in terms that Eugene can understand. Eugene suddenly comes to the shocking realization, *Jane believes that possible worlds exist in the same way that the actual world exists!* Compare Eugene's case to **Phillipa the Philosopher's** case. After listening to Jane explaining her views to Eugene, Phillipa suddenly comes to the shocking realization, *Jane's a modal realist!* I claim that Eugene's realization was more difficult to think than Phillipa's, even though there was no difference in linguistic processing involved. Eugene was more likely to fail to have the thought if distracted or otherwise impaired, and he was more likely to fail at other tasks competing for his attention. If this is right, then not only is there a difference in the difficulty of the linguistic processing required to generate the

two thoughts, but there is also a difference in difficulty between the thoughts themselves. Similar arguments apply to the example of the thoughts induced by (*Philosophers*$_{hard}$) and (*Philosophers*$_{easy}$).[16]

An alternative explanation of the differences in difficulty available to the molecularist is this: When concepts are acquired, their constituent representations become more strongly connected such that they can be activated together more easily. This may suffice for the undergraduate case, since there the two thoughts differ with respect to whether their subjects have acquired certain concepts. However, it does not by itself explain the difference in difficulty of thinking the two thoughts now to yourself; you have already acquired the relevant concepts, and so the molecularist taking this line should predict that thinking both would be equally easy.[17] And it does not speak to the samenesses in difficulty of the thoughts produced by (*Utilitarian*) and (*Vegetarian*).

At this point, a general note about the examples used is in order: For convenience, I have chosen concepts that we can readily associate with definitions (e.g. MODAL REALIST). However, my argument does not depend on associated contents actually being definitions. For instance, if you think concepts represent prototypes, then think of the allegedly more difficult thoughts as those involving the representations of the elements in the prototype. In the case of MODAL REALIST, this prototype might include adherence to a typical version of modal realism (say, David Lewis'), an attraction to possible world semantics, a willingness to bite bullets, and perhaps an unkempt hairstyle. In all such cases in which the associated vehicle of MODAL REALIST is more complex than MODAL REALIST, it will be more difficult to think a thought involving MODAL REALIST than a thought involving its associated vehicle; I claim that this is because they're distinct and differ in complexity.

In sum, the cases we've examined show that the difficulty of thinking a thought comes apart from the complexity of the thought's associated vehicle. The efficient concept view and the complex content view can explain this pattern of similarities and differences in difficulty better than their common competitor, molecularism.

---

[16] If you think that thought is just inner speech, then you might not like this. However, if that's your view, then you should already agree that (*Philosophers*$_{hard}$) and (*Philosophers*$_{easy}$) induce different thoughts in us, since they induce different word-representations, and further that (*Philosophers*$_{hard}$) induces more word representations than (*Philosophers*$_{easy}$). This is enough to secure (*Simpler Vehicles*).

[17] One might combine this response with a denial of the possibility of thinking the hard thought once one is capable of thinking the easy thought, but this seems introspectively implausible. It does seem that we are capable of thinking two distinct thoughts by reading each sentence. Thanks to David Pitt and Helen Yetter for pressing this and related objections.

The effects of concept acquisition on memory

Not only does concept acquisition allow one to think easier thoughts with a particular associated content, it also confers short-term memory benefits when it comes to remembering that associated content. In the psychology literature, this phenomenon is sometimes known as **chunking**, or the recoding of a larger amount of information into a smaller amount (Miller, 1956).[18]

Wickelgren (1992) and others take chunking to provide evidence for (what I'm calling) (*Simpler Vehicles*). A loose reconstruction of the general line of argument goes like this: Concepts (the chunks) and their associated vehicles (what the chunks "stand for") behave differently when it comes to memory: Concepts are easier to store and later recall than their associated vehicles. So, the two must be distinct. Further, the best explanation of the difference in difficulty in memory storage and retrieval is this: Concepts are simpler than their associated vehicles, and so, (*Simpler Vehicles*) is true.

There are other possible ways to make sense of chunking that do not require (*Simpler Vehicles*) and that are thus open to the molecularist. Perhaps during the process of concept acquisition, a concept's associated vehicles become more strongly linked together. This makes it easier for them to be activated together in the future, making memory storage and retrieval easier than it previously was. While this is a possibility, the explanation in terms of (*Simpler Vehicles*) is preferable for at least the following reason: It allows us to provide a unified explanation of differences in difficulty in occurrent thoughts and the memory benefits of chunking. This is desirable because, first, the same concepts that are used in occurrent thought are also used

[18]The idea was anticipated by D. O. Hebb (1949), who maintained that new ideas are represented by new vehicles of representation (for him, cell assemblies) that are linked to the vehicles representing their unchunked contents.

Chunking has been extensively studied in chess players. De Groot (1966) and Chase and Simon (1973) studied differences between chess players of different skill-levels. One task was a short-term memory task: For a few seconds, subjects viewed a board configuration that either corresponded to a stage in a possible competitive game or that was randomly generated. Then the subjects were asked to reconstruct the configuration on a new board with the original board out of sight. Experts greatly outperformed novices in the possible competitive game condition but not in the random condition. These results are best explained by chunking: Experts have chunk representations in long-term memory that "stand for" common multi-piece chess configurations. When presented with a board configuration, they can store the chunks in short-term memory instead of the positions of each individual piece. To reconstruct the configuration on a new board, they unpack the chunk concepts into their target contents, which specify information about the individual pieces. Novices have smaller or no chunks, and so even if they can store the same number of chunks in short-term memory, less can be retrieved from those chunk concepts. But when the pieces are placed randomly on the board, the advantage of experts over novices disappears, since the experts have no chunk concepts corresponding to parts of those meaningless configurations.

Similar studies have found similar results in bridge (Engle and Bukstel, 1978; Charness, 1979), electronic circuit diagrams (Egan and Schwartz, 1979), music (Sloboda, 1976; Beal, 1985), basketball (Allard et al., 1980), computer programming (McKeithen et al., 1981), map use (Gilhooly et al., 1988), and dance (Allard and Starkes, 1990).

in short-term memory (it would be a strange view that denied this), and second, the capacity to think easier thoughts seems to arise at the same time as the short-term memory benefits: they arise when one "masters" a notion, when one "gets it," or in other words, when one acquires a concept. And so, we might expect it to be the same feature of concept possession that confers both advantages, which, if I am right, is the ability to think simpler thoughts.[19]

Subsection summary

Phenomena related to differences in difficulty and the short-term memory benefits of chunking support views of concepts that accept (*Simpler Vehicles*), such as the complex content view and the efficient concept view.

10.2.2 Unpacking

In order for chunking to be a useful way to remember complex associated contents, it must be possible to move from the chunked thought or concept to the information that it "stands for," and this is in fact what happens. The process is sometimes called "**unpacking**." Unpacking need not be an intentional or personal-level process. You usually don't *decide* or *intend* to unpack a concept. Rather, at least in most cases, unpacking is a non-conscious, sub-personal process. The process may admit of a computational further explanation, or it might not. Perhaps the best further explanation is neurobiological. For now, we can just take unpacking to be a non-conscious "stab in the dark" type of process: we're wired up in such a way that when we need to retrieve further information relating to a concept we're using, unpacking occurs automatically.

Unpacking links concepts to other representations.[20] Theorists such as Wickelgren (1992) have suggested that the unpacking relation might define a more or less hierarchical conceptual structure, linking concepts standing for more general categories to more specific categories.[21]

[19]This explanation of the memory benefits of chunking obtains further support from observations concerning the phenomenology of memory retrieval. When one tries to remember what happened at a baseball game, one may first retrieve (say) that there was a home run. Then, if necessary (say, if one were trying to describe the game to someone lacking the concept of a home run), one may subsequently retrieve that a batter hit the ball, then passed first base, then passed second base, then passed third base, and finally arrived at home, scoring a run. If molecularism were true, it would be hard to understand how retrieval could proceed in two stages like this, since the results of the first retrieval presumably would involve all the same concepts and contents as the results of the second retrieval.

[20]I use the word "unpacking" to stand for a relation between concepts and their associated vehicles, rather than a relation between contents.

[21]Barsalou (1999) proposes a relationship that looks like unpacking between concepts (which he calls "simulators") and perceptual representations (simulations triggered or produced by the simulators). Chris Eliasmith is developing a neuro-computational model on which concepts are "semantic pointers," pointing to perceptual representations. His view

The view so far is one on which concept possession consists in having representations that can be used in lieu of more complex representations in thought and memory, but that unpack into those more complex representations when necessary.

10.2.3 For (*Simpler Contents*)

So far, we have that concepts are fairly simple. In this subsection, I argue that their *contents* are fairly simple as well.

I claim that views that reject (*Simpler Contents*) but accept (*Simpler Vehicles*) end up attributing content that is not psychologically involved. This is evidenced by the fact that retrieval of further information that is represented by a concept's associated vehicle is sometimes needed in order to perform certain tasks. For instance, consider the thought generated by (*Tables*):

(*Tables*) The facts about tables supervene on the physical facts.

Suppose I now ask you, "What's supervenience?" It takes at least a little bit of time for you to answer. I claim that what is happening is this: You are unpacking your concept of supervenience, and in doing so, you're retrieving parts of its associated content that weren't there in the content of the occurrent thought you had upon reading the sentence.

Unpacking is required not only for explicit retrieval of associated content, but also for certain other uses of concepts, for instance in reasoning or in generating behavior. For example, if I ask you whether a necessary God supervenes on everything, it might take you some time to answer. That is at least partly because you have to unpack your concept SUPERVENIENCE to retrieve information about what supervenience is.[22,23]

That concepts need to be unpacked in order for us to make use of their associated contents suggests that those associated contents were not represented by the concepts in the first place. Unpacking a concept involves activating a concept's associated vehicles and *only then* entertaining its associated contents. The alternative proposal that associated contents are represented

---

can be understood at least partly as providing a theory of the unpacking relation. I think it's also fair to say that he endorses (*Simpler Contents*), although the notion of content at play may be different than mine, and he takes the contents of concepts to determine how they unpack.

[22]Unpacking tends to occur only when needed. If you and I are rattling off names of all the modal realists we know, we may not need to unpack our concept of modal realism very much. In contrast, if we are arguing over whether or not modal realism is true, we might need to do quite a bit of unpacking.

[23]A related phenomenon is that we occasionally **fail to unpack**. Sometimes, perhaps when we are tired, intoxicated, or otherwise impaired, we are able to think the thought induced by (*Table*) but we are unable to retrieve further information about what supervenience is. It's unclear how to make sense of these cases on views on which we were entertaining SUPERVENIENCE's associated contents in having the original thought.

by both a concept *and* its associated vehicles attributes psychologically uninvolved contents to concepts: These attributed contents are not involved in reasoning, the generation of behavior, or phenomenology. Of course, there are *some* contents that play these psychological roles, but it turns out that these contents are merely the results of unpacking.

This line of argument can be precisified and generalized to form an argument for the following claim about conceptual complexity:

(*COM*) The complexity of a concept C's mental content does not exceed the complexity of C.

Recall that a vehicle of representation is complex to the extent to which it has parts that are also representations and a content is complex to the extent to which it has parts that are also contents. (*COM*) claims that a concept's content is not more complex than the concept itself. (*COM*) is entailed by a perhaps more intuitive thesis, which we might call the **explicit representation thesis (*ERT*)**. (*ERT*) states that all of a concept's content is *explicitly* represented by the concept, meaning that it has representational parts representing every part of its content. If you like (*ERT*), then you should like (*COM*) as well.[24]

From (*COM*) and (*Simpler Vehicles*), we can argue for (*Simpler Contents*) as in the following argument sketch: Concepts are fairly simple (from (*Simpler Vehicles*)). The complexity of a concept's content does not exceed the complexity of the concept itself (from (*COM*)). Therefore, the contents of concepts are also fairly simple.

Before arguing for (*COM*), I want to emphasize that we are dealing here with psychological phenomenona. Again, concepts and contents are psychologically involved: They are involved in reasoning, the generation of behavior, and phenomenology. From a perspective that emphasizes psychological involvement, positing concepts and contents that are not psychologically involved is unmotivated and should be avoided. The crux of my argument is that positing contents that violate (*COM*) is unmotivated in just this way.

Suppose, for an empirical analogue of *reductio ad absurdum*, that we have a simple concept C that represents the complex conjunctive content *a and b*. Since C is simple, it does not have proper parts that are themselves representations. Since *a and b* is complex, it has parts that are contents, such as, let us say, *a*. C is depicted in Figure 10.3.

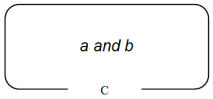

Figure 10.3: The concept C.

[24]My claim is that (*COM*) in fact holds for our concepts, not that it holds for other mental or non-mental representations in us or in other actual or possible creatures.

One aspect of psychological involvement is involvement in reasoning, behavior, or further thought. For a content to be psychologically involved in this way, it must be **usable**, either in conscious thought or experience, or subconscious or nonconscious processing. Whatever it takes for a concept to be used, there must at least be a vehicle of representation that is simultaneously activated. If vehicles of representation are brain states, then the claim is that when a content is used, a brain state that represents it is activated. So, then, we can say that in order for a content to be *usable*, there must be a representational vehicle that represents the content in question and that *can* be activated.

The problem with attributing complex contents to C, then, is this: C does not have any representational proper parts, and so it does not have any parts whose content is *a*. So, the content *a* that is allegedly had by C is not usable, and hence not psychologically involved in the way described. So, considerations from reasoning, behavior, and the generation of further mental states do not support attributing the complex content *a and b* to C.

Although this argument uses as an example the case of a simple concept representing a less simple content, it is easy to see how the argument generalizes to deem unmotivated any attribution of content in violation of (*COM*). Complexity in content that outruns the complexity of the concepts representing it results in more proper parts of contents than there are vehicles that can support them. And so some parts of the content are going to end up "irretrievably trapped in the concept" and psychologically uninvolved, making it unmotivated to posit them.

Of course, there might be another representation, say A, that represents *a*, and that is triggered by C. Then *a* is usable and so it can have a psychological role of the sort we're discussing. But then it is in virtue of being represented by A, not C, that the content *a* has this psychological role. And so we still have no motivation for positing *a* as part of the content of C. This is precisely what happens when the concept in question can be unpacked. For instance, SUPERVENIENCE can be unpacked into a complex content involving the content *possible worlds*. Unpacking involves having further thoughts involving other concepts, such as POSSIBLE WORLDS, and it is being represented by POSSIBLE WORLDS that allows *possible worlds* to have any psychological role, not its being represented by SUPERVENIENCE. And so, that SUPERVENIENCE can be unpacked to yield the content *possible worlds* does not motivate attributing *possible worlds* as part of the content of SUPERVENIENCE.

So far, I have argued that considerations arising from behavior, reasoning, and further thought and experience do not motivate attributing contents in violation of (*COM*). However, there is another aspect of psychological role: **phenomenology**. We can understand phenomenology broadly as including both "what it's like" to be in a certain state (Nagel, 1974), as well as an experienced awareness or "grasping" of contents. In order to motivate content attributions in violation of (*COM*) on the basis of phenomenology, then, we would have to maintain that concept use sometimes gives rise

to phenomenology more complex than the concept. I claim that this is implausible. It would be strange if a state's associated phenomenology were more complex than the vehicles that subserved it. Phenomenology is quite plausibly tied to and limited by features of the vehicles of representation in accordance with (*COM*), just like other aspects of psychological involvement.

Readers who are not satisfied with the immediately preceding argument might be satisfied with an argument for the weaker conclusion that phenomenological considerations do not motivate content attributions in violation of (*COM*) in the sorts of cases we've been considering. When we think the thought invoked by (*Table*), we do not seem to enjoy a complex phenomenology corresponding to the use of the concept SUPERVENIENCE. We do not "grasp" the complex associated content all at once and in full detail. This is evidenced by the fact that when we unpack our concept and experience a representation of what supervenience is, we enjoy a different phenomenology, one that is richer and more complex. This is more obvious when we consider concepts with more complex associated contents, such as ARGUMENT FROM ILLUSION. Consider the thought induced by the following sentence:

(*Illusion*) The argument from illusion has puzzled many philosophers.

When we enjoy the thought induced by (*Illusion*), we do not immediately feel that we "grasp" all the premises of the argument in our thought, nor do we have other phenomenological experiences corresponding to all the parts of the associated content. Again, this is evidenced by the fact that when we unpack the concept ARGUMENT FROM ILLUSION, we enjoy an experience with a *different*, more complex, phenomenology.[25] And so, even if you don't agree with my previous arguments against phenomenological considerations supporting content attributions in violation of (*COM*) generally, you should accept that phenomenological considerations do not justify content attributions in violation of (*COM*) in the kinds of cases we've been examining.

This concludes my arguments for (*COM*). We can now argue for (*Simpler Contents*) as follows: Suppose a concept C is complex to degree n. C is less complex than its associated vehicle, which has complexity $m > n$ (from (*Simpler Vehicles*)). C's associated contents are as complex as C's associated vehicles, that is, they have complexity of at least m. But C's content cannot be more complex than n (from (*COM*) and the first premise). Therefore C does not represent its associated contents, that is, (*Simpler Contents*) is true.

One might quarrel with the terms of the debate. One might suggest that mental representation is not a psychological phenomenon, or that psychological involvement comes to something other than what I suppose. My short reply to these kinds of worries is this: It's fine with me if you use the term

[25]It may not be possible to unpack the concept ARGUMENT FROM ILLUSION all at once. However, even a partial unpacking will yield a thought with a more complex phenomenology.

"mental representation" to denote something else. Still, you should agree with me that there is *some* important phenomenon that is psychologically involved in more or less the way stated. I am talking about *that*.[26]

10.2.4 The content of concepts

According to the efficient concept view, concepts do not represent their associated contents, but rather something simpler. There are various options with respect to what kind of contents concepts represent. These contents might be vague or specific parts of a concept's associated contents (e.g. the content of MODAL REALISM might be *possible worlds view*). They might be or involve words (e.g. *"modal realism"* or *view called "modal realism"*). They might be new simple contents (e.g. *modal realism (its own thing)*). Different concepts or even what we would intuitively call the same concept used on different occasions might have different kinds of contents. Everything I say is compatible with any of these options, as well as with various combinations of them (as long as the resulting contents attributed to concepts are compatible with the observations made so far).[27]

10.2.5 Section summary

I have argued for most aspects of the efficient concept view: Concepts are distinct from their associated vehicles and don't represent their associated contents. Instead, a concept's associated contents are usually only available through unpacking. What remains is to describe the role of experiences of cashing out. I turn to this in the next section, where I argue that experiences of cashing out, together with unpacking, allow us to define new notions of content.

10.3 Derived Content

So far, I've argued that the thoughts we think involve fairly simple concepts with fairly impoverished contents. We might be a bit unhappy with this

[26]This is only my short reply. I do not deny that much depends on the initial terms of the debate and our initial understanding of mental representation. Chapter 2 is devoted to defending my starting point over other starting points that are popular in the literature and Chapter 8 is a general discussion and motivation of my methodology. I do not have the space to repeat the arguments in those chapters here. Instead, I will offer the following consideration, which I think can be supported in the context of this chapter alone: My approach to mental representation leads to increased understanding of the purpose and functioning of concepts, and, as we will soon see, sheds new light on problems in other domains. To the extent to which the approach I take is fruitful and increases our understanding, it thereby gains some support.

[27]In Chapter 7, I argue for the phenomenal-intentional identity theory (PIIT), on which representational content is identical to phenomenal character. This view offers a ready account of what the contents of concepts are: they are "what it's like" to be in a state in which they are activated. The efficient concept view goes hand in hand with this view of mental representation, though it does not require it.

picture. Some projects that we might care about presuppose or require a less impoverished notion of content, e.g. projects in action theory, where complex contents are attributed to beliefs and desires.

Further, we certainly seem to think that some contents are in some sense equivalent to others. When one concept unpacks into another, we sometimes take the results of unpacking to be a further **cashing out** of what we were thinking before. Such contents *present themselves as*, or *come as* further cashings out of previously entertained contents. For example, we might think, *Modal realism, the view that possible worlds exist in the same way that the actual world exists, that's what I was thinking all along*. Put otherwise, experience does not present itself as one mental snapshot after another. Rather, some thoughts are experienced as related to previous thoughts in various ways, including as being further elaborations of those thoughts.

Cashing out is not the same thing as unpacking. While unpacking is (at least usually) a nonconscious process, cashing out is *experienced*. Further, while unpacking is usually accompanied by experiences of cashing out, the two can come apart. One concept might unpack into another closely associated concept without the subject undergoing an experience of cashing out. For example, after a traumatizing experience, I might come to strongly associate cats with mutilation, and my concept CAT might unpack into MUTILATION. Still, I do not experience *mutilation* as a further cashing out of *cat*.

Nevertheless, the picture that emerges is one on which unpacking and cashing out for the most part conspire to generate the illusion of richly contentful thought. We entertain a thought, such as JOHN IS A MODAL REALIST, and when we ask ourselves what we were just thinking, the process of unpacking is triggered, and we experience *John believes that possible worlds exist in the same way as the actual world exists* as a further cashing out of our initial thought. Given this situation, it is a bit too pessimistic to conclude that the full story about concepts is that they have a fairly simple structure and represent fairly simple contents. Instead, I want to suggest that, given the roles in mental life of unpacking and experiences of cashing out, we can sensibly define new notions of derived content that capture the types of content we intuitively take ourselves to be entertaining, as well as other useful notions of content.

To do this, let us first note a distinction between source representation and derived representation. **Source representation** is representation in the first instance, or representation of the "metaphysically" most basic type. A theory of source representation tells us where representation first enters the picture, how it gets injected into the world. So far, we have been talking about the source representational content of concepts. **Derived representation**, as the name suggests, is representation obtained or derived in some way from source representation. For example, it is commonly thought that the content of sentences is derived from the content of mental states. Corresponding to the distinction between source and derived representation, we can introduce a distinction between source and derived content: A representation's **source**

content is what it source-represents, while its **derived content** is what it derivatively represents.[28]

In this section, I will suggest that although concepts do not source-represent their associated contents, they derivatively represent at least some of them. Before beginning, however, I want to emphasize that source representation and derived representation might be fundamentally different types of phenomena. In other words, the occurrence of the word "representation" in both "source representation" and "derived representation" is somewhat incidental. Still, later on I will argue that the notion of derived representation satisfies our intuitive notion of representation, despite its dissimilarities with source representation.

10.3.1 Defining new notions of content

As previously discussed, in certain circumstances we experience some contents as further cashings out of some other contents. For instance, in circumstances D, subject S might experience *unmarried man* as a further cashing out of *bachelor*. We can specify the **cashing out relation (C)** as follows:

(C) aCb for some subject S in some circumstances D iff S experiences b as a cashing out of a in circumstances D.

C holds between contents. The contents that are experienced as cashings out of some content might themselves cash out into other contents. In other words, a might cash out into b, and b in turn might cash out into c. Call each successive instance of cashing out a **C-step**. We can say that a is related to c by two C-steps. a might be related to more contents by more C-steps. We can then define a new relation, C^*, which is the ancestral of C:[29]

(C^*) aC^*b (for S in D) iff a is related to b by any number of C-steps (for S in (any of) D).

We can now use C^* to define a derived content schema:

(*Derived Content Schema*) A mental representation A with source content a has derived content b (for a subject S) iff aC^*b (for S in (any of) D).

According to this notion of derived content, concepts derivatively represent all of their contents' potential cashings out, as well as all the potential cashings out of their potential cashings out, and so on. For example, if a subject is disposed to experience *unmarried man* as a further cashing out of *bachelor*

[28] I borrow my terminology from Terrence Horgan and Uriah Kriegel. This distinction more or less corresponds to the Searlean and Gricean distinctions between **intrinsic** (or **original**) and **derived** content.

[29] C is a psychological relation, in that there are experiences of cashings out. C^* is an abstraction from a series of psychological relations, but it need not itself be thought of as a psychological relation.

in the appropriate circumstances, then her concept BACHELOR derivatively represents *unmarried man*. If she is disposed to experience *human male* as a further cashing out of *man* in the appropriate circumstances, then *man* is part of the derived content of her concept BACHELOR as well (see Figure 10.4).

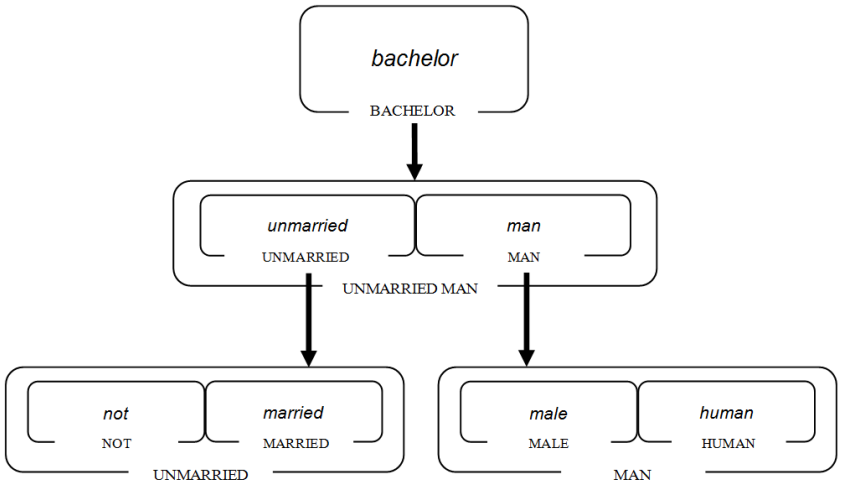

Figure 10.4: The concept BACHELOR and its cashing out relations. The content of the concept on the tail of a solid arrow cashes out into the content of the concept on the head. BACHELOR derivatively represents everything connected to it by a series of such arrows. While Figure 10.1's depiction of the efficient concept view focused on illustrating its claims about complexity, the efficient concept view is ultimately best depicted with a diagram like this one.

This is only a schema. We obtain different notions of derived content by restricting the relevant circumstances. To see why this is useful, suppose that **Susan** is disposed to experience *selfish pigs* as a further cashing out of *bachelor* in circumstances involving a few incredibly bad dates and a few incredibly strong beers. She might think to herself: *Men become nice only once they're married. Any man who's single is a selfish pig. That's what bachelors are, selfish pigs!* If we do not restrict the circumstances in which experiences of cashings out are relevant to a concept's derived content, Susan's concept BACHELOR counts as at least partly derivatively representing *selfish pigs*. Note that it is irrelevant whether Susan is ever in such circumstances; the mere fact that in such circumstances she *would* experience *selfish pigs* as a further cashing out of *bachelor* is sufficient to make her concept BACHELOR count as derivatively representing *selfish pigs now*. This problem afflicts us all: any one of us *could* find ourselves in *some* situation where they would experience surprising contents as further cashings out of *bachelor*. Indeed,

with sufficient intoxication, almost anything could present itself as a further cashing out of almost anything else.[30]

For this reason, it is useful to have notions of derived content that impose restrictions on what are to count as the appropriate circumstances. We might consider the appropriate circumstances to be normal circumstances (on some notion of normalcy), sufficiently nearby circumstances (on some notion of nearness), circumstances not involving learning or theory change, or circumstances that the subject in question considers to be in some way important or authoritative. There is no in principle right or objectively privileged way to restrict the circumstances. Different restrictions yield different notions of derived content that might be useful for different purposes.

For instance, consider Tyler Burge's character **Bert**, who believes that he has arthritis in his thigh (Burge, 1986). If we impose a fairly strict restriction on circumstances, one that excludes circumstances involving intoxication, theory change, learning, additional experiences of various sorts, and the like, then in the relevant circumstances, Bert might experience *inflammation of the joints or thighs* as a further cashing out of *arthritis*. This fairly strict restriction might be useful for certain purposes, such as predicting Bert's immediate behavior. For example, we can predict that Bert is likely to describe his malady with, "I have arthritis in my thigh." However, for other purposes, such as predicting Bert's long-term behavior, other notions of derived content might be useful, and so we might instead impose a different restriction on circumstances, one that includes nearby circumstances involving some learning (e.g. the circumstances in which Bert has a dictionary handy, asks the relevant authority, or asks his mother). In these circumstances, Bert will experience *inflammation of the joints* as a further cashing out of *arthritis*. So then Bert's concept ARTHRITIS would count as derivatively representing *inflammation of the joints*.[31] This notion of derived content might be useful for longer-term predictions of Bert's behavior. We might predict that Bert is

[30]It is not only intoxication or other forms of impairment that we might want to block from determining a concept's derived content. The possibility of acquiring new true or false beliefs through conventional learning methods has a similar effect. Suppose that Bob currently believes that electrons are wave-particles as described by some current theory. Suppose that in the (merely possible) circumstances in which he learns that scientists claim to have discovered that electrons are actually little angels sent from God to provide us electricity, Bob experiences *little angels sent from God to provide us electricity* as a further cashing out of *electron*. Whether or not Bob is ever put in such circumstances is irrelevant; the very fact that he *would* have such experiences affects the derived content of his concept ELECTRON *now*.

[31]If we include both the immediate circumstances, in which Bert would experience *inflammation of the joints or thighs* as a further cashing out of *arthritis*, as well as the more distant circumstances, in which Bert would experience *inflammation of the joints* as a further cashing out of *arthritis*, then we obtain conflicting derived contents for Bert's concept ARTHRITIS. If this result is undesirable, it can be avoided by including only one of those sets of circumstances, or by providing a decision procedure to choose between such contents or circumstances in cases of conflict. Perhaps (in certain circumstances) Bert would consider certain circumstances to be more authoritative than others; we might invoke such preferences in our decision procedure.

unlikely to seek long-term arthritis treatment as a result of the pain in his thigh. This example illustrates the usefulness of different notions of derived representation for different purposes. It is also interesting to note that some restrictions result in an externalistic notion of derived representation.[32,33]

We can also fill in D in such a way that it includes all and only the cashings out obtainable while being locked in a room for an indefinite amount of time without making new observations. These circumstances yield something approximating the results of idealized rational deliberation. We would then get a notion of derived content that captures something like the **primary intension of two-dimensional semantics** (see e.g. Chalmers (2004a)). It is sometimes argued that the primary intension of 2-D semantics is the ideal type of content to use for the purposes of at least the beginnings of certain sorts of inquiry (Jackson, 1998). If this is indeed the case, it is useful to have a notion of derived content that can capture them. For one, we can agree to use this notion of derived content when constructing and evaluating metaphysical arguments, which would prevent what (if all this is right) would turn out to be needless debate over the semantics of the terms and concepts involved. Further, it would grant some psychological reality, albeit of a limited sort, to the notions of content used in such metaphysical debates.

We can also fill in D so as to allow for interaction with particular artifacts, yielding a notion of content analogous to the externalistic notion of belief at play in Clark and Chalmers' extended mind (1998). Clark and Chalmers describe a case of **Otto**, an Alzheimer's patient who writes important information in a notebook that he carries with him at all times. They argue that it is reasonable to count the results of consulting the notebook as part of Otto's set of beliefs. Clark and Chalmers' view is a certain sort of externalism about beliefs and other propositional attitudes, which they call **active externalism**. My framework can allow for an analogous type of active externalism about *contents*. Suppose Otto has notes such as "Arthritis: an inflammation of the joints" written in his notebook. When someone asks him about arthritis, he consults his notebook and experiences *an inflammation of the joints* as a further cashing out of *arthritis*. By including interactions with the notebook in the relevant circumstances, we obtain an active externalistic notion of content for Otto, which might be useful for predicting his behavior

[32]Burge uses the Bert example to motivate externalism about source content. Here, we are using it to motivate externalism about *derived* content, although only for certain purposes, such as predicting behavior in the long-term. In Chapter 12, I argue that this is the strongest conclusion such arguments can establish.

[33]When it comes to predicting behavior, the ideal circumstances are those that the subject is likely to find herself in at the time of behavior. If we are interested in finding out what Bert will do in the next five minutes, then we don't want to include circumstances that are only likely to be encountered further in the future, such as the circumstances in which Bert visits his doctor. But if we are interested in Bert's long-term behavior, such as the likelihood of his seeking long-term arthritis treatment, we should also include the circumstances in which he is likely to find himself by the time he could potentially engage in those behaviors.

(assuming he reliably checks his notebook when necessary), as well as for specifying what we intuitively take to be the truth conditions of his thoughts, and indeed, what he himself would accept as the truth conditions of his thoughts.[34]

If we allow D to include circumstances in which we undergo various perceptual experiences, then a concept's derived contents can be perceptual contents. For instance, if, upon seeing blue in the appropriate circumstances, we think *that's what blue is*, then at least part of the derived content of the concept BLUE is the visually represented content.[35] In this way, we can attribute the intuitively correct derived content to perceptual recognitional concepts.[36]

Recall, also, that we left open the question of what particular associated contents thoughts and concepts have. The discussion has so far been neutral with respect to debates between prototype theories, exemplar theories, theory theories, and definitional theories of concepts. Now we can further say that apart from asking whether the theories are true of the source contents of any concepts, we can ask whether they are true of the derived content of any concepts, on some notion of derived content. Of course, which notion of derived content we should use to evaluate these theories will depend on our interests and purposes. For instance, if we are interested in the quick categorization of objects, then a different restriction on circumstances might be relevant than if we are interested in the use of concepts in drawing inferences. This allows for the possibility that different psychological theories of concepts are true relative to different notions of derived content that are appropriate for different phenomena in which we might be interested. Further research is needed to determine whether these possibilities in fact obtain.[37]

Another interesting project is to use our schema to construct a notion of derived content that closely tracks **folk psychological content**, the type of content we intuitively attribute to our own thoughts and to those of others, often in order to explain or predict behavior. Such an account might offer an explanation of the surprising predictive accuracy of folk psychology (see, e.g. Fodor (1987, Ch. 1), and Jackson (1998)) without requiring us

[34]This example is closer to home than we may like. It is not unusual for academics to write a paper and then a few years later remember that they still endorse what they wrote, but notice that they have forgotten exactly what their arguments were or even how they used their terms. Active Externalism for Academics can solve this problem by attributing to them the beliefs and derived contents they would have if given the opportunity to re-read their papers.

[35]This and other cases of merely recognitional concepts are cases in which unpacking and cashing out come apart, since (except for those with an unusually good imagination) your concept BLUE does not unpack into a visual representation of blueness.

[36]I am skating over complications arising from the fact that BLUE is supposed to represent a *range* of perceptually representable colors, not just one color, but it is easy to see that the resources for such subtleties are available. It is now just a matter of refining our notion of derived representation to get the results we want given our purposes.

[37]This is very much in the spirit of Machery (2009), who argues that different methodologies used to study concepts actually tap different things.

to attribute folk contents as the actual source contents of mental states. It would also offer a story about the relationship between folk psychology and what is really going on in the head that is neither eliminativist about folk content (since folk content is a type of derived content and not wholly unrelated to what is going on in the head), nor wholeheartedly realist about it either (since folk content is not source content, but rather one of many possible types of derived content, each on par with the others, metaphysically speaking).[38]

I have described different notions of derived content that can be defined using the derived content schema. This is only a start. Depending on our purposes, we can elaborate upon the schema in various ways.[39] I want to emphasize that the resulting notions of derived content are not in competition with one another. Rather, we can and should accept **pluralism about derived content**. Different notions might be useful for different purposes. I have offered a preliminary survey of some of the different notions that seem to currently be in use, and how we can accommodate them on my framework. The more general point is that, although concepts have fairly impoverished source contents, they have the next best thing: rich, varied, and robust derived contents.

10.3.2 Is derived content really content?

One might worry that these derived contents do not really deserve to be called "contents." What we have are genuine mental contents, dispositions to unpack, and dispositions to accept the results of certain unpackings as further cashings out of the initially entertained contents. But, one might wonder, *How does* that *result in anything like representation?* In a certain sense, it doesn't. I'm not reporting any new metaphysical facts about the world. If you care about the metaphysics of content, then it's source content that you care about, not the derived content I'm discussing here.

Still, at least some notions of derived content behave much like other cases that we are happy to describe using the word "representation." Given that we do not and probably cannot occurrently think everything we might want to be thinking, we do the next best thing: We use certain thoughts and concepts to *stand for* more cumbersome thoughts and concepts. We use our concepts *as if* they represented their derived contents. There are two aspects to this use. First, derived contents can be retrieved when needed. There is a way to move from a concept to its derived content, a way to

[38] I expand on these points in Chapter 11. In Chapter 12, I argue that the fact that the folk notion of content partly tracks derived content has important implications for arguments relying on folk intuitions about contentful states.

[39] For instance, we might want to separate out contributions to derived content that look like what is sometimes called "modes of presentation." *Modal realism* might cash out as both *the view that possible worlds exist in the same way that the actual world exists* and *the view about possible worlds that David Lewis made famous*. As is, our current derived content schema treats both contributions to content similarly.

"decode" it. This is unpacking. Second, there is a sense in which we accept the "decoded" content as the content of the initial representation. We mentally relate the two. This is analogous to relating a meaning to a word and thereby conferring derived content to the word in question (or at least contributing to the determination of the word's meaning).

Here's an analogy that illustrates how derived representation behaves like representation: On Google Maps, various icons explicitly represent banks, grocery stores, and drinking establishments. This explicit representation is analogous to the source content of mental states. When we click on one of these icons, say an icon representing a drinking establishment, we get more information that we accept as being about the item in question, say that it is called the "Wig and Pen." This is analogous to subsequent contents that we take to be further cashings out of previous thought contents, and that can, therefore, be considered the derived content of the original representation.

There is a sense in which we can say that the icon represented the additional information that the drinking establishment is a pub called the "Wig and Pen" even before we clicked on it. After all, we can manipulate and reason with the icon, and that simulates manipulating and reasoning with the additional information. For example, if we move the icon to a new location, it is as if we are moving the icon *together with the additional information* to a new location. After all, when we click on it again, the same information appears, and we accept it as being relevant to the icon just as before. Similarly, we use our concepts to reason, solve problems, and remember things, and that simulates reasoning, problem solving, and remembering things using the associated contents of those concepts. For instance, without unpacking the concept SUPERVENIENCE, we can reason as follows:

1. All physicalists hold that the mental supervenes on the physical.
2. John is a physicalist.
3. Therefore, John believes that the mental supervenes on the physical.

If we were to unpack the concept SUPERVENIENCE at any point in our reasoning, we would get the same result, and we would accept it as being at least part of what we meant by *supervenience* all along. Just as we can manipulate the Google Maps icon while retaining the results of clicking on it, so too can we reason with the concept SUPERVENIENCE while retaining the results of unpacking it. And just as manipulating the icon in Google Maps simulates manipulating the results of clicking on it, so too does reasoning with the concept SUPERVENIENCE simulate reasoning with the results of unpacking it. Since concepts behave in this way, it makes sense to think of derived mental representation (at least of some varieties) as satisfying something like an intuitive notion of representation, even though it does not qualify as genuine mental representation.

I said that there are two features of derived mental representation that contribute to its being appropriate to call it a species of representation: unpacking relations, and experiences of cashing out. To see the relevance of each, it is instructive to consider cases where one is absent. Imagine a subject whose concept BACHELOR does not unpack into UNMARRIED MAN or any other representation, but who occasionally has out-of-the-blue experiences of cashing out relating the content *bachelor* to *unmarried man*. This person would accept *unmarried man* as a further elucidation of *bachelor*. Whether or not we want to say that his concept BACHELOR does indeed derivatively represent what he takes it to, his derived representational content is practically useless. Whenever we ask him what a bachelor is, he stares at us blankly, and he is unable to classify men he knows to be unmarried as bachelors. The difference between a normal subject and this subject is analogous to the difference between a competent speaker of a language and a creative but absentminded genius who tries to invent a new language by writing down thousands of definitions on scraps of paper that she subsequently loses and whose content she forgets. Every once in a while, she finds a scrap of paper with a definition on it, but she quickly loses it again, and she can never find them when she needs them. Perhaps, as a courtesy, we might say that she have successfully stipulated the meanings of various new terms. But as a representational system, her language is useless to her. She cannot write in it or speak in it. She cannot use it to make inferences. And so, whether or not we want to say that cashing out is sufficient for derived mental representation, it is clear that a representational system with experiences of cashing out but without unpacking relations is practically useless as a representational system. Unpacking contributes to its being appropriate to classify derived mental representation as a type of representation.

Next, imagine a subject whose concept BACHELOR unpacks into UNMARRIED MAN but who has no corresponding experiences of cashing out. She might suddenly conclude that John, who she knows to be a bachelor, is unmarried. From her perspective, this inference might look a bit like a stab in the dark. At best, she might think that being a bachelor and being unmarried co-occur, but, from her perspective, she does not acknowledge any *semantic* connection between the two. Although the unpacking relations between her concepts make those concepts useful to her getting around the world, since she does not recognize any semantic relations between them, we might be less tempted to recognize the relations between her concept BACHELOR and the content *unmarried man* as instances of representation, as opposed to mere association. In short, her concepts and other representations are connected in the right way, but the experiential component common to many cases of conventional or stipulated representation is absent. We might be tempted to say that her concept BACHELOR as representing *unmarried* man in the same way that smoke represents fire (see Chapter 1), but this falls

short of the kind of representation we would be happy to attribute to other representational systems that we can be said to use, such as languages.[40]

10.4 The production view and the efficient concept view

In Chapter 9, I raised a worry for the production view. It seems that we represent varied, subtle, and sophisticated contents, and it's not at all clear how this is possible if representing contents requires producing contentful states. We might be so impressed by the breadth and complexity of contents we can represent so as to think our minds could not possibly produce such contentful states. The relation view, on the other hand, seems to have an easy time accounting for our subtle and complex thought contents. We get to represent such contents by being appropriately related to similarly sophisticated and complex objects and properties in the world.

However, if the efficient concept view is correct, the situation drastically changes. The contents that our thoughts and concepts represent are not as sophisticated and complex as we may have initially thought. Rather, they are quite impoverished or schematic. This has the following two effects on the debate between the production view and the relation view: First, it makes the production view much more plausible. The contents we represent are simpler and more elusive, and this makes it much less surprising that they are merely a product of the operation of our brains or minds.

Second, not only does the situation improve for the production view, but it also worsens for the relation view. While it may be fairly clear how we can become related to certain complex and determinate properties in the world, such as the property of being a female fox, it is less clear how we can become related to the elusively schematic contents that our thoughts and concepts represent on the efficient concept view. It's not clear, for instance, that there is a property corresponding to the vague and schematic content *vixen*. Even if there is such a corresponding property, this property is not the same property as the one corresponding to *female fox*, and so it is less than clear how we get to be related to it, when all we really ever encounter are female foxes. In any case, it seems somewhat odd to speak of concepts source-representing full-fledge properties in the first place, and this suggests that even if the relation view can accommodate the efficient concept view, the resulting combination might appear quite unnatural.[41]

In summary, although the efficient concept view does not automatically entail the production view, it fits quite well with it. On the production view, contentful states are products of the mind or brain. These products need not correspond to everyday properties, such as the property of being a

[40]Thanks to Kati Farkas, Frank Jackson, and Matt Ishida for helpful discussion on the topics of this section.

[41]Thanks to Harold Hodes for discussion.

female fox, but might instead correspond to the vague and schematic contents represented by concepts on the efficient concept view.

10.5 PIIT and the efficient concept view

In Chapter 7, I argued for the phenomenal-intentional identity theory (PIIT), the view that representation is identical to consciousness. PIIT faces a challenge when it comes to thought: Thoughts seem to have rich, subtle and complex representational contents, but relatively impoverished phenomenal characters. The efficient concept view offers the following response for PIIT: Thoughts and the concepts they involve don't have that much representational content after all. And so, although I have not specifically argued for this, it now seems fairly plausible that the content they do have corresponds to their phenomenal character in the way required by PIIT.[42]

As promised in Chapter 7, §7.3.2, the efficient concept view allows the intentionalist and PIIT theorist to respond to the challenge posed by the red wall case. The challenge is this: A perceptual experience representing a uniformly painted, uniformly textured, unique red wall at arm's length away intuitively has have the same representational content as a thought that there is a uniformly painted, uniformly textured, unique red wall at arm's length away, but the two experiences have different phenomenal characters. This is a counterexample to intentionalism, and hence PIIT: we have two cases that are representationally alike but phenomenally different.

[42]Kriegel and Horgan (forthcoming) suggest a framework for thinking about theories of mental representation that distinguishes the source of mental representation (where mental representation first enters the picture, how it gets injected into the world) from derivative sources of representation. For the phenomenal intentionality theory, the source of representation is phenomenal character. Kriegel (forthcoming a) suggests that other types of representation can be derived from the conscious content attributions of an ideal interpreter.

David Bourget (2010) employs a similar strategy for the pure intentionalist to deal with non-phenomenal apparently representational states, particularly standing states. He claims that phenomenal character can be identified with *underived*, but not derived, intentionality. On his view, content can be derived in at least the following four ways: (1) linguistic deference, (2) composition (thus, a state source-representing *unmarried man* would have derived content, on Bourget's view), (3) dispositions or causal connections, and (4) the "matching" relationship between narrow and broad content (roughly, the concept WATER might have the derived or underived content *the clear watery stuff* as far as matching is concerned, but it has the derived content H_2O). Bourget's notion of derived representation is much broader than mine. On my view, the relevant notion of derived content is a specific case of Bourget's (3) and possibly (4). On Bourget's characterization of derived content, compositional contents are automatically derived contents, while this is not the case on my characterization. For this reason, it is difficult to compare his proposed solutions to problems for the pure intentionalist to my own. In particular, in the case of thought, Bourget claims that since thinking is a temporally extended process, its content is derived. The inference seems to rely on thought being compositional (in a temporally extended way). This does not address the question I am asking, which is about the temporal parts of an extended thought process, and whether their content is derived. In any case, the spirit of our views is very similar.

The efficient concept view allows for the following response: The two cases may involve the same derived contents, but they do not involve the same source contents, and intentionalism and PIIT are theses about source content. The visual representation of unique red represents the precise property of redness, while the concept does not specify what exactly it is to be unique red. As discussed above, perhaps the concept RED is a recognitional concept, such that the circumstances of unpacking relevant to its derived content (on some interesting notion of derived content) include circumstances where one is having various perceptual experiences.

What are we to make of our intuition that the two cases represent alike? In Chapter 11, I will argue that our intuitions about mental state contents do not track source content, but rather something closer to derived content. If this is right, and if, as I have suggested, there is a notion of derived content on which the two experiences derivatively represent the same contents, then our intuitions that the two experiences of a red wall represent alike are not far off. And so, this strategy for dealing with the red wall case allows the intentionalist and the PIIT theorist to accommodate our intuitions.

The efficient concept view also allows us to deal with the internalism-compatible alleged types of counterexamples to (*Distinctiveness*) discussed in Chapter 7, §7.4.2, which is helpful to both the phenomenal intentionality theorist and the PIIT theorist.

(*Distinctiveness*) Thoughts with different contents have different phenomenal characters.

This type of counterexample involves two subjects (or one subject at different times) that have different views about what constitutes a certain type of thing, say, a bachelor. Dimitri takes bachelors to be unmarried men, while Dimitra takes bachelors to be men available for dating or marriage. Intuitively, when Dimitri and Dimitra have a thought they would each express with "Ivan is a bachelor," their thoughts have different contents but the same phenomenal character. And so, we have a counterexample to the phenomenal intentionality theory, and hence to PIIT. If two states can have the same phenomenal character but differ in representational content, then representational content does not supervene on and is not identical to phenomenal character.

According to the efficient concept view, there is a distinction between source content and derived content, and so there are two corresponding ways of construing (*Distinctiveness*):

(*Distinctiveness*$_{source}$) Thoughts with different source contents have different phenomenal characters.

(*Distinctiveness*$_{derived}$) Thoughts with different derived contents have different phenomenal characters.

Since the phenomenal intentionality theory and PIIT are theories about source content, not derived content, (*Distinctiveness*$_{source}$) is the only version of (*Distinctiveness*) that these theories need. But, according to the efficient concept view, contents such as *unmarried man* and *man available for dating or marriage* are not source-represented by the concept BACHELOR, but rather, only derivatively represented. And so it is quite plausible that the difference between Dimitri and Dimitra's thoughts are differences in derived content only, not in source content. If this is right, then the Dimitri/Dimitra case is not a counterexample to (*Distinctiveness*$_{source}$), and that's the only version of (*Distinctiveness*) that matters for the phenomenal intentionality theory and PIIT.

What about our intuition that Dimitri and Dimitra represent differently? As I will argue in Chapter 11, our intuitions about content do not track source content, but rather something closer to derived content, and so our intuition that Dimitri and Dimitra represent differently (to the extent to which we have such an intuition) only supports the falsity of (*Distinctiveness*$_{derived}$), which quite plausibly is false. Indeed, cases like the Dimitri/Dimitra case suggest that it probably is false.

10.6 Conclusion

In slogan form, the efficient concept view allows you to think more than you can occurrently think. Concepts and their contents are fairly simple, but they unpack and cash out into more complex contents, which they can be said to derivatively represent. In this way, concept possession allows us to get around limitations on how much can be thought, remembered, and manipulated at a time, while simulating thinking, remembering, and manipulating thoughts of of virtually unlimited complexity.

The efficient concept view fits nicely with the production view and PIIT. The resulting picture is one on which mental representation is a product or process of the mind or brain that is identical to phenomenal consciousness. This identity between representational content and phenomenal character holds not only in the case of perception, but also in the case of non-perceptual states such as thoughts.

Chapter 11

Folk Psychological Content: A Case Study

In the previous chapter, I argued that thoughts and the concepts they involve have surprisingly impoverished contents. I tried to ease the pain of this situation by suggesting ways of defining notions of derived content out of actual and potential mental contents. I offered a general schema that can be filled out in various ways to yield different notions of derived content. I will suggest that with some other fairly uncontroversial ingredients, we can offer an account of what folk psychological notions of content track that explains their perhaps surprising success in predicting behavior. While I focus on folk content, similar procedures can be used to define notions of derived content useful for other areas of inquiry, such as decision theory, personal-level psychology, or epistemology.

One reason I focus on folk content, rather than other types of content, is the prominence of this notion in philosophical discussions of mental representation. Some theorists approach the problem of mental representation with the notion of folk content already in mind (see Chapter 2 §2.2.1). Such approaches are often justified by appeal to the predictive accuracy of folk psychology. The example of airplane tickets is often used here. If I buy a ticket to Montreal for a year from now, and you agree to pick me up at the Montreal airport, then, using your folk understanding of my mind, you can predict that in a year from now, I will be at the Montreal airport. It is thought to be fairly impressive that we can predict the movement of large bits of matter, sometimes years in advance, in such a way. This is taken to support some kind of realism about the posits of folk psychology.

Although I have chosen a different approach to the problem of mental representation, and I believe it has proven fruitful, one might still wonder why folk psychology is so successful given my view that the only psychologically real and even potentially causally potent content is not folk content, but rather the much more impoverished source content of mental states. One might turn this into an objection as follows: The predictive accuracy of

folk psychology supports realism about the existence, causal potency, and explanatory relevance of the content-like posits of folk psychology, namely folk contents. Folk contents must be psychologically real in a way that precludes their being abstract and possibly fairly arbitrary constructions out of other real entities. But I claim that thoughts source-represent fairly impoverish contents, and so it seems I am denying the psychological reality of folk contents. So, one might argue, I must be wrong.

This argument is not conclusive. At best, we have arguments for conflicting conclusions, one in favor of the exclusive psychological reality of the sometimes impoverished source contents I want to attribute to mental states, and one in favor of the psychological reality of folk contents. Still, it is worth defusing the argument. We can do this by offering an account of how folk psychology can be predictively accurate *despite* failing to track the only psychologically real content, source content. This is what I will do.

Here's how I will do it: I will set aside the question of what the folk psychological notion of content really is. In other words, I am not interested in providing a conceptual analysis of the folk concept of content. On my view, there may be more than one "correct" analysis, depending on which notion of derived content we are interested in, or there may be no "correct" analyses at all (see Chapter 13). Instead, I will focus on the question of what the folk concept of content tracks.[1] The aim is to explain folk psychology's predictive accuracy by showing that its concepts track something interestingly correlated with behavior, despite not tracking source contents.

11.1 What the folk notion of content tracks

My suggestion is that the folk notion of content tracks a combination of psychologically real mental content, derived content, and referents of derived contents. In the next section, I will argue that this is a good combination of things to track in order to predict behavior in situations of limited knowledge.

11.1.1 Extra ingredient #1: Derived content

According to folk psychology, subjects count as believing many of the unpacked contents of their occurrent beliefs, desiring many of the unpacked contents of their occurrent desires, and so on. For example, if John has the occurrent belief that Fred is a bachelor, we might count him as believing that Fred is an unmarried man. We use this belief content to predict John's behavior. If John, in an attempt to play matchmaker, is introducing his sister to all the unmarried men he knows, we can predict that he will introduce her to Fred. Indeed, we would have great difficulty tracking the source content of John's concept BACHELOR instead and using that in our predictions. If the version of PIIT I have been defending is right, BACHELOR's source content

[1] See Chapters 4–6 for the distinction between tracking and representation.

might be a vague phenomenal feel that varies between individuals or the same individual at different times, which would be a difficult thing to keep track of.

I said there are many different possible notions of derived content. These are obtained by filling in D in the following schema (see §10.3.1):

(*Derived Content Schema*) A mental representation A with source content a has derived content b (for a subject S) iff aC^*b (for S in (any of) D).

What's the right way to fill in D for the purposes of constructing a notion of derived content that can capture what the folk notion of content tracks? We might proceed to answer this question through the study of folk intuitions, perhaps using the methods of conceptual analysis. We can consider various cases and consult our intuitions on whether we take them to be authoritative on folk content in the right way. For example, we probably don't want D to include circumstances involving drugs or too much learning. I would expect the result to be fairly messy, and perhaps to differ for different people. It might also turn out that there is more than one folk notion of content that we use in different circumstances.

11.1.2 Extra ingredient #2: Referents

Sometimes we count subjects as having mental states that in some way constitutively involve particular objects, kinds, or other worldly entities. For example, you might count yourself as having a visual experience that involves *this* very typed word. Proponents of this sort of content sometimes call it **object-dependent content**, but we can extend the notion to cover contents that are supposed to include worldly entities other than objects, such as events or kinds. For ease of exposition, I will focus on objects in what follows.

That our folk notions of content at least sometimes track referents may be related to the much discussed intuition that it is sometimes appropriate to move from statements such as "Ahmad believes that a is F" to "There is something such that Ahmad believes of it that it is F." It may also be related to our at least occasional factive uses of mental state verbs like "see." Indeed, I would speculate that our initial attraction to relation views of mental representation may have something to do with such intuitions.

Which object-dependent contents are we willing to include in our content ascriptions? I claim that these are usually the objects subjects manage to refer to. In the previous example, you had a visual experience that succeeded in referring to the italicized token of the word "this." You count that token of "this" as part of your visual experience, not some other token of the word "this" or some other thing entirely.

I said that there were many possible notions of derived content, as well as a notion of genuine mental, or psychological, content. Which type of content is such that the folk notion of content tracks its referents? Again, the details

are a question for empirical study of folk intuitions. However, we can fairly confidently say that the relevant type of content is not just source content. The source content of thoughts does not sufficiently determinately pick out objects or other worldly things, or if it does, it picks out the wrong things. Rather, we must include much unpacked content in order to get content that stands a chance of picking out the referents we intuitively take mental states to have. So we are most likely tracking the referents of derived contents, on some notion(s) of derived content.

To what extent our folk notion of content tracks the referents of derived content is an open question. A reason to think our concept of folk content does not always and only track these referents is that we are happy to attribute to subjects mental states about nonexistent individuals, such as Santa Claus. So our Santa Claus belief attributions do not track reference to Santa Claus. At best, they track *attempted* reference. But that is just to say that they track the source and derived content of subjects' concepts of Santa Claus, since at least in cases like this, these are the determinants of (attempted) reference.

11.2 Why these extra ingredients do not hinder the predictive accuracy of folk psychology

I have suggested that folk psychological notions of content track some combination of source content, derived content, and referents of derived content. Still, source content is the only content occurrently running through subjects' heads, the only psychologically real content. All other parts of what the folk notion of content tracks are either abstractions of actual and potential source contents, or worldly entities. So if any type of content determines or is somehow intimately related to the generation of behavior, it looks like it should be source content. This makes the success of folk psychological predictions of behavior somewhat puzzling. In this section, I will argue that the extra ingredients that folk psychological notions of content track do not hinder folk psychology's predictive accuracy. In fact, given certain constraints, they enhance it.

11.2.1 Why it's not so bad to track derived content

If we had unlimited information about a subject's mental states, we could somewhat reliably predict her behavior by keeping track of her moment-to-moment occurrent contents. These contents sometimes generate or are otherwise appropriately related to behavior such that we can use the mental state contents to predict the subject's immediate behavior. However, this will not always suffice, because sometimes unpacking is required in order to generate appropriate behavior. In other words, sometimes we don't want to predict immediate behavior, but behavior occurring in the near or distant

future, after the relevant sorts of unpacking have occurred. For these sorts of predictions, we would ideally also keep track of the subject's dispositions to unpack in various circumstances, as well as the circumstances she is in. Then, to the extent to which source content is predictively related to behavior, we will be able to predict the subject's behaviors.[2]

Of course, given our limited information concerning subject's occurrent mental states, we cannot always (or perhaps ever) use this method to predict behavior. The folk notion of content does the next best thing: it tracks a combination of source contents and some types of derived contents. It indiscriminately tracks both source content and the contents available through unpacking in certain circumstances.

By conflating mental content and derived contents, some information is lost, such as information about when particular parts of the attributed contents are likely to be occurrently entertained. However, this information is not particularly useful for most ordinary folk psychological predictions. Unpacking usually occurs automatically and relatively reliably when required, so it does not matter if we count a subject as thinking the unpacked contents of her thoughts all along, since whenever these potential results of unpacking are needed, they tend to be retrieved anyways.

Here is another way to put the general point: In the last chapter, I suggested that occurrently entertaining an impoverished source content and retrieving derived contents only when needed simulates a situation in which we are occurrently thinking all the derived contents at once but are focusing only on the parts that are currently needed. Folk notions of content track something closer to what we're simulating. But since our actual psychologically real states simulate what folk notions of content partly track, folk notions of content form a fairly good basis for behavioral predictions.

Further, it is worth repeating that the alternative, which is to separately keep track of mental content and the potential results of unpacking in various situations, is unfeasible. It would require more knowledge than we have available of subjects' moment-to-moment occurrent states and how they are wired up to other states. Given that this is practically impossible and anyways inefficient, it is not a bad solution to track a lumped together combination of mental content and derived content.

Of course, predictions based on different notions of derived content will vary in their success. In general, predictions will be most successful to the extent to which the relevant circumstances involved in determining derived content include the circumstances in which the subject finds herself when she produces her behavior (or circumstances that are relevantly similar).[3] Just

[2]Notice that while the assumption that a subject is rational might help us predict her inferences and actions given her precise mental contents, it does not tell us how she will unpack her concepts, since there is no fact of the matter as to what the "correct" unpacking is.

[3]Of course, if a subject unpacks similarly in circumstances that are not included in the circumstances determining derived content, our predictions might still be accurate.

how successful folk psychology is at predicting behavior is an open question, as is which type of derived content folk notions of content track. For my argument to be successful, folk psychology's degree of predictive accuracy should more or less match the degree of predictive accuracy that can be attained through reliance on whatever type of derived content folk notions of content track. (Of course, the success of folk psychological predictions of behavior depends on more than which type of derived content folk content attributions track. It also depends on the accuracy of the principles of folk psychology that allow us to move from mental state attributions to behaviors.)

11.2.2 Why it's not so bad to track referents

We sometimes want to make predictions about **successful behaviors** (Williamson, 2000, pp. 75-80). Whether or not a behavior is successful can depend on whether or not the derived content of the mental states giving rise to it succeed in referring. To borrow an example from Timothy Williamson, if Fred is thirsty and has a visual experience and desire that manage to refer to the glass of water on the table in front of him, we can attribute to him an object-dependent desire for *that* glass of water. This allows us to predict that he'll be drinking water soon, as opposed to merely *trying* to drink water.

Further, whether or not a behavior is successful affects **subsequent behavior**. If we can predict that Fred's attempt to procure the water on the table will be successful, then we can also predict that Fred won't be searching the fridge for water (because his desire for water will have already been satisfied).

As this example illustrates, having our folk notions of contents track the referents of subjects' source or derived contents allows us to make predictions about successful behaviors and about behaviors further in the subject's future. This is precisely because these sorts of predictions require information about more than just the subject's occurrent psychological states: they also require information about the world, and in particular, about whether the world corresponds in the right way to the subject's states. This is precisely what our attributions of folk content track when they track referents. As in the case of folk content tracking derived contents, the referents are not always tracked by a separate state. We do not always have one mental state tracking a subject's derived content, say, of the perceptual visual experience of a glass of water on the table, and then a separate state tracking the fact that the first state succeeds in referring to the glass of water on the table. Rather, when we think *Fred sees the glass of water*, we have one state that tracks a combination of both types of state, both Fred's derived contents, and the

But this would be a case of epistemic luck; we got the right answer based on the wrong evidence.

states of the world that make them accurate. And this is useful because we can predict both the success of Fred's behavior and his subsequent behaviors.

11.3 Conclusion

I have argued that folk notions of content track a combination of source content, derived content, and referents of derived content. The result is a view of the status of folk psychology that is neither wholeheartedly realist, since folk psychological notions of content track a mixture of factors, only one of which is genuinely psychologically involved, nor eliminativist, since folk psychological notions of content are not wholly divorced from psychological reality, but instead track a construction obtainable from actual and potential mental contents along with certain relevant states of the world that is very useful in predicting behavior given our limited access to what's really going on in the head. This explanation of the fairly impressive success of folk psychology without full-blown realism about folk content thus defuses any arguments for realism about folk content based on such success.[4]

I repeat that the case of folk content is just an example of notions we can construct from mental content using the resources I have provided. Similar strategies can be used to construct other notions of content that are used in other areas of philosophy or other disciplines, such as action theory, metaphysics, epistemology, and sociology. The crucial factor differentiating the types of contents useful for these different purposes will be the circumstances of potential unpacking that we are interested in. To the extent to which we are interested in different circumstances for different purposes, we can make use of different notions of content.

[4]To the extent to which what's really going on in the head diverges from what folk psychology tracks, we might predict some of the failures of folk psychology. This would further aid in defusing arguments from the success of folk psychology. But this is a topic for future research.

Chapter 12

The Internalism/Externalism Debate

In the previous chapter, I argued that the folk psychological notion of content does not track source contents, but rather a combination of source contents, derived contents, and referents of derived contents. This has implications for certain arguments relying on folk intuitions about psychological states. This chapter is a discussion of how the efficient concept view, as well as the results of the previous chapter concerning folk content, transform the internalism/externalism debate. In particular, I will argue that arguments for externalism relying on intuitions at best only establish conclusions about folk content, not source content, and further, that the internalist about source content can accommodate these intuitions. This might make everyone happy, or, what is more likely, it might make everyone unhappy. I like to think of this discussion as proposing a way for internalists and externalists to meet in the middle.

12.1 Reconciling internalism with externalist intuitions

The debate between internalism and externalism is a debate over what kinds of factors play a role in determining what mental states represent. **Externalism** is the view that environmental factors sometimes play at least a partial role in determining content, whereas **internalism** is the denial of externalism.

Externalism has recently become very popular. The main arguments in favor externalism are thought to be thought experiments. A challenge for internalists is to explain away or accommodate the externalistic intuitions. In this section, I will provide a possible way for internalists to do both.

12.1.1 Externalist intuitions

Here is a variant of what is perhaps the most famous externalist intuition pump in Hilary Putnam's Twin Earth thought experiment in "The Meaning of 'Meaning'" (Putnam, 1975):

Oscar and Toscar are intrinsic duplicates on Earth and Twin Earth, respectively. Twin Earth is a planet in our universe that is just like Earth, except that the clear, watery substance flowing from the taps and rivers is not H_2O, but rather some other substance with a long and complicated chemical composition that can be abbreviated with "XYZ." The intuition is that when Oscar and Toscar think the thought associated with the phrase "Water is wet," they think thoughts with different contents. Since Oscar and Toscar are intrinsic duplicates, this difference must be a result of their different environments. Thus, environmental factors partly determine mental content.[1]

12.1.2 Worries with externalism

Despite these perhaps strong intuitions, there are good reasons to resist externalism. First, there is a *prima facie* tension between externalism and our capacity to know what we're representing introspectively. If what my mental state represents depends on features of my environment, then in order to find out what I represent, it looks like I have to investigate my environment. This conflicts with how we think we usually find out what we're representing: by introspection. Here's another way to bring out the weirdness: If externalism about water is true, it looks like Aristotle did not know what his concept of water represented, since he didn't know about water's chemical constitution. That seems absurd. Similarly, advances in science may falsify the judgments about our mental contents we are currently disposed to make (suppose it turns out that water isn't H_2O). There have been attempts to reconcile externalism with self-knowledge, but they are strained.

A second common worry with externalism is that it is not clear how contentful states can be psychologically involved. On a natural way of individuating behaviors, Oscar and Toscar behave exactly alike. So whatever representational difference the externalist attributes to them is causally impotent and generally psychologically irrelevant. As with the problem of self-knowledge, attempts have been made to reconcile externalism with the psychological involvement of mental content, but they too have been strained.[2]

[1] This argument was originally supposed to support externalism about linguistic content, but the same type of argument can and often is taken to support externalism about mental content.

[2] Sometimes the problem is put in terms of the causal potency of mental contents: If any kind of content has a causal effect on behavior, it's internalistic content. This is because the environmental factors externalists deem relevant to content do not differentially affect

Finally, to the extent to which the arguments in Chapters 3–4 are convincing, we may have additional reasons to want to resist externalist theories, since they are versions of relation view. I will not repeat these arguments here. Right now, I'm not trying to make an argument against externalism. There is much to be said on both sides of the debate. The reason I am bringing up these worries is to eventually show that the internalist's way of accommodating externalist intuitions avoids them.

12.2 Rendering externalist intuitions harmless

In Chapter 11, we saw that the folk psychological concept of content doesn't track source content, but rather a combination of source content, derived content, and referents. So results from thought experiments invoking our pretheoretical intuitions about the content of mental states can at best only track facts about folk content, and so can only support conclusions about folk content. Internalism and externalism are meant to be views about source content. So, it is a mistake to think that intuitions invoked in Twin Earth-type thought experiments bear on the internalism/externalism debate. In other words, Twin Earth intuitions are compatible with internalism, and so they are harmless to the view. At best, what we have are arguments for a different kind of externalism, **externalism about folk content**.

12.3 Accommodating externalist intuitions

Not only can the internalist render externalist intuitions harmless, but she can also accommodate the type of externalism they might be said to properly motivate, namely externalism about folk content. There are two main ways for environmental factors to influence folk content.

12.3.1 The first strategy

One way to accommodate externalism is a variant of the 2-D semantics strategy employed by Jackson and Chalmers (see, e.g. Chalmers (2004a)): We say that Oscar and Toscar's concepts corresponding to the word "water" have as their derived content something like the clear watery stuff around here, or the clear watery stuff on Earth in the actual world. The efficient concept view allows us to additionally emphasize that these descriptions need

what's going on inside the head, and it's what's going on inside the head that causes behavior. While this is true, it will fail to move someone who doesn't think that semantic properties have causal powers in the first place. This is why I prefer to make the broader claim that the only content that can be psychologically involved, where psychological involvement can involve a mere reliable correlation with behavioral effects (for more on psychological involvement, see Chapter 10, §10.2). Externalistic content is not only causally irrelevant, but psychologically uninvolved generally.

not be occurrently running through Oscar and Toscar's head, since these contents may be merely derived contents.

A subject might be disposed to experience *the clear watery stuff around here* as a cashing out of *water* and so WATER would derivatively represent *the clear watery stuff around here*.[3]

We can treat deference to experts, a linguistic community, or the Bible in a similar way. Hilary accepts that elms are trees of the kind that experts call "elms." Fred defers the content of his concept of goodness to Mackie's book. Jane defers the content of her highly theoretical concepts to the theory she has scribbled in her journal, which she has now forgotten.

To repeat, what the efficient concept view contributes to the overall picture is a view of where the relevant descriptive content that secures reference is to be found: It is to be found in folk content, a species of derived content.[4]

12.3.2 The second strategy

Another way to get externalism about folk content is to specify the appropriate circumstances such that they can differ between intrinsic duplicates. Recall the derived content schema:

(*Derived Content Schema*) A mental representation A with source content *a* has derived content *b* (for a subject *S*) iff aC^*b (for *S* in (any of) *D*).

[3]In Chapter 10, I suggested that we might be able to add more features to the derived content schema, perhaps so as to make it recognize different types of derived content, such as definitions, conditions on what it would take to be a cashing out of some content, or modes of presentation. If we do this, then we can alternatively (or perhaps additionally) say that *the clear watery stuff around here* might be disposed to present itself as a condition on what it would take to be a cashing out of *water*. Subjects might express this state with, "I don't really know what it is to be water, but whatever it is, it's to be the same kind of thing as the clear watery stuff around here." This would be an expression of knowing what it takes to be what water is without knowing what water is. Then *the clear watery stuff around here* would be part of WATER's derived content of a different type.

[4]The efficient concept view also confers an additional advantage for employing a 2-D semantics style strategy. Notice that for the strategy to succeed in securing externalism about derived content, Oscar and Toscar must manage to refer differently by *around here* or *this sample*. Oscar must be able to refer to *his* environment on Earth or *his* sample in front of him, while Toscar must be able to refer to his environment on Twin Earth or his sample in front of him. Otherwise, if Oscar and Toscar refer alike with *around here* (say, because around here unpacks into a description of their home planet, and both Earth and Twin Earth satisfy both their descriptions), then they refer alike with *the clear watery stuff around here*, which is precisely what we need to deny in order for this strategy to secure different derived contents to their "water" concepts. One thing the efficient concept view can say, then, is that *around here*, *this sample*, and other contents that appear to involve an indexical element, all unpack into something to do with the subject, like *the place where* I *am*, *the sample that is at such-and-such distance in front of* me, etc. Then, all we need is for subjects to be able to directly refer to themselves. In other words, we can reduce other instances of apparent direct reference to only one, more plausible, sort of direct reference, namely direct reference to oneself (or perhaps even just one's own mental states). This does not yet solve the problem, for we still have to provide an account of direct reference to oneself or one's experiences, but it limits the problem.

Intrinsic duplicates might differ with respect to *D*. Here's how this strategy might deal with Twin Earth: Suppose the relevant circumstances for Oscar include circumstances in which he has future scientists from *his* world handy, while for Toscar, the relevant circumstances include circumstances in which he has future scientists from *his* world handy. These are different circumstances. In one case, they involve scientists from Earth, while in the other case, they involve scientists from Twin Earth. In those two different circumstances, different contents would present themselves as a cashing out of water to Oscar than to Toscar.

The point is that we can use this strategy to get Oscar and Toscar to have different derived contents on the same notion of derived content. The notion of derived content is the *same* because we specify these circumstances using the same description, namely, the circumstances in which *the subject* has future scientists from the subject's own environment handy. But since the subject and *the subject's own environment* differ for Oscar and Toscar, the actual circumstances picked out are importantly different.

This type of strategy is most helpful for accounting for cases of **active externalism**; see §10.3.1.

12.4 Conclusion

In this chapter, I argued that externalist intuitions are not evidence for externalism about mental content. If anything, they are evidence for externalism about folk content or some types of derived content. This is enough to block arguments against internalism about mental content that rely on such intuitions. I then further suggested that it is possible to be an internalist about mental content and an externalist about certain types of derived content. I described two different ways in which this is possible.

Chapter 13

Conceptual Analysis

Conceptual analysis can be generally characterized as the project of finding the content or meaning of a concept or linguistic expression through methods such as introspection and consideration of possible cases. In what follows, I will focus on the analysis of concepts rather than linguistic expressions. Such projects can be of independent interest, as when one wonders what we really think persons are. They can also be of instrumental interest to other projects; for instance, if we are interested in finding out whether there are persons, we might start by asking what our concept PERSON represents in an effort to precisify our question.[1] Sometimes it is also hoped that conceptual analysis will deliver substantive *a priori* metaphysical truths.

In most cases, the targets of conceptual analysis happen to be some the associated contents of concepts. The target analysis of VIXEN, for instance, might be *female fox*. On the efficient concept view, these contents are often merely *derived* contents. This readily explains several features of conceptual analysis that are in some cases more difficult to explain on molecularism and the complex content view: (1) Progress takes work (§13.1); (2) the results of conceptual analysis can be apparently conflicting with no clear resolution (§13.2); and (3) satisfactory results can appear to be next to unattainable (§13.2). We can also explain and resolve the paradox of analysis (§13.4) and offer what might appear to be an appropriately deflationary account of the analytic/synthetic distinction (§13.3). Finally, we can shed some light on concept reconstruction (§13.5).

13.1 Conceptual analysis takes work

We cannot just immediately analyze a concept. Rather, conceptual analysis takes work. We have to perform various exercises: we have to consider possible cases, possible analyses, possible counterexamples to those analyses, and so on. This might seem surprising on some views of concepts such as

[1] Conceptual analysis, and particularly this approach, has been recently and forcefully defended by Frank Jackson (1998).

molecularism and the complex content view. On those views, a concept's associated content is occurrently running through our heads when we use that concept. But if these contents are occurrently running through our heads every time we use a concept, then why can't we just activate that concept, say by using it in a thought, and immediately *tell* what the correct analysis will be?

Relatedly, when presented with a putative analysis of a concept, we are not very good at judging its correctness. For example, an analysis of KNOWLEDGE as *true, justified, belief* was generally thought to be correct until Gettier's counterexamples (Gettier, 1963). This is surprising on molecularism and the complex content view. If either of those views is correct, then it is mysterious why can't we just directly compare a proposed analysis of a concept with the content occurrently running through our heads when we use that concept. In other cases, we do seem able to make fairly reliable comparisons of the sameness or difference of occurrent contents. For instance, if we are shown two color samples side by side, we are able to judge whether we represent them as the same color or as different colors, as in the picture of two gray squares in Figure 13.1.

Figure 13.1: Two gray squares.

Just by looking at the two gray squares, we are able to judge that we represent them as being different shades. We can immediately tell that we are entertaining two different color contents in each case. Why are we not able to make such direct comparisons of sameness or difference in content in the case of concepts?

I do not deny that various maneuvers can be made to make these two related observations about conceptual analysis compatible with molecularism and the complex content view. However, the efficient concept view makes good sense of these observations without any fancy additional moves. On the efficient concept view, the associated contents that conceptual analysis aims to get at are oftentimes merely derived contents (or perhaps folk contents; see Chapter 11). These derived contents are not occurrently entertained when we use the concept in question. Instead, these contents need to be somehow retrieved using various methods, such as the consideration of possible cases. That's why conceptual analysis takes work, and that's why we cannot immediately judge a proposed analysis by comparing it with the contents of a thought involving the concept in question. We first need to perform activities that trigger unpacking. This in turn explains why we are often surprised by the results of conceptual analysis. For instance, we were surprised to find out that we're inclined to judge that the Pope is not a bachelor, and thus that the

"correct" analysis of BACHELOR is not *unmarried man*. In order to obtain this sort of derived content, we had to specifically consider the case of the Pope and see what happened when we unpacked in those circumstances. Similarly, after encountering Gettier cases, many of us were similarly surprised to find out that the analysis of KNOWLEDGE is not *true, justified belief*.

13.2 Conceptual analysis faces in principle difficulties

Sometimes it seems that, despite our best efforts, an analysis is not forthcoming, such as in the case of KNOWLEDGE. These difficulties are predicted and explained by the efficient concept view.

One difficulty arises from a tension between two methods that might be used in conceptual analysis: One method, which we may call the **brainstorming method**, involves brainstorming conceptual relations between the concept in question and other concepts. For instance, we might ask ourselves what a bachelor is and come up with the condition that a bachelor must be a man.[2] The brainstorming method aims to directly find what we might call a concept's **intension**, or the rule that determines whether the concept applies. (I assume that the ultimate goal of conceptual analysis is to find the intension of concepts, which are taken to be or or in some way determine their contents.)

A second method, the **method of cases**, involves considering possible cases and asking ourselves whether they satisfy the concept in question. For instance, we might consider the Pope and ask whether we consider him to be a bachelor. We then look for features that all items that we want to consider bachelors have in common. The method of cases aims to find a concept's intension by first finding the items to which it seems to apply.

These two methods can pull in different directions. It might turn out that the first method returns something like *unmarried man* as the content of BACHELOR, while the second method returns most unmarried men and some married men who have been separated for 10 years and whose wives refuse to grant them a divorce for financial reasons, but fails to return the Pope, a recently widowed 90 year old man, and a 25 year old man in a long-term domestic partnership as items to which BACHELOR applies, and thus recommends something like *man available for dating or marriage* as the content of BACHELOR.

In the case of BACHELOR, we might go back and review our conclusions obtained with the brainstorming method. Perhaps a more careful application of the method will reveal that a bachelor is a man available for dating or marriage, or something more complicated that better fits our intuitions about particular cases. However, this is not always so easy. Consider the

[2]This method is similar to the method of brainstorming folk platitudes about various things and ramsifying over those platitudes in Lewis (1972) and Jackson (1998).

case of KNOWLEDGE. The brainstorming method yields a content like *true, justified belief*. However, as our intuitions concerning Gettier cases show, our knowledge-judgments do not just track our judgments of whether something is a true, justified belief. Reapplying the first method in this case arguably does not offer much else. The two methods appear to be in tension.

The efficient concept view explains the tension as follows: Concepts are **autonomous**: They are distinct representations from representations of their associated contents and are thus free to enter into tracking and other relations without their associated vehicles thereby entering into such relations. The autonomy of concepts gives rise to the conflict between the brainstorming method and the method of cases. Further, this conflict is not a mere curiosity, or a mere challenge in an ultimately valid methodology, but rather an in-principle difficulty arising from the very nature of concepts.

Returning to the case of KNOWLEDGE, since KNOWLEDGE is autonomous, it is distinct from TRUE, JUSTIFIED, and BELIEF, and may enter into relations of various sorts without TRUE, JUSTIFIED, or BELIEF thereby also entering into those relations. For instance, KNOWLEDGE might enter into tracking relations without TRUE, JUSTIFIED, or BELIEF, or some combination of those concepts, thereby entering into those same tracking relations. Figure 13.2 illustrates the possible case in which KNOWLEDGE partly tracks reliability (of the mechanisms generating the putative knowledge), but does not cash out into *reliability*. The solid line represents cashing out relations and the dotted line represents tracking relations. In this example, the concept RELIABILITY has a causal effect on the concept KNOWLEDGE without being related to it by the cashing out relation. Thus, KNOWLEDGE tracks, but in no sense represents, reliability.

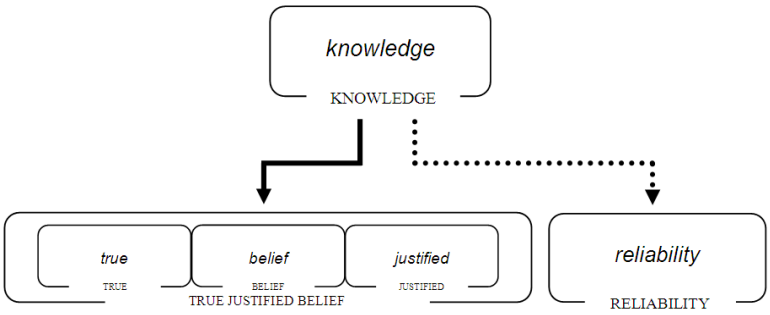

Figure 13.2: The concept KNOWLEDGE might track reliability without the composite concept TRUE, JUSTIFIED, BELIEF or any of its parts thereby tracking reliability.

Tracking relations are sometimes mediated by representations, as in the above case, but sometimes they are not. For example, the depth-experiences involved in 3-D vision track both occlusion of one object by another and binocular disparity (the disparity between the images projected on either retina). However, whereas we not only have states that track, but also states that represent, occlusion of one object by another, we do not likewise have

states that represent binocular disparity. The causal path between binocular disparity and depth-experience is not mediated by a representation representing binocular disparity. Figure 13.3 might be a partial processing diagram for depth-experiences, where black boxes stand for non-representational states. (This is just a toy example. Of course, 3-D vision is much more complicated.)

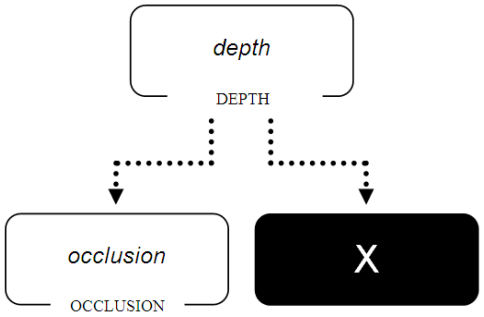

Figure 13.3: A toy 3-D vision processing diagram. The blacked out box stands for a non-representational state.

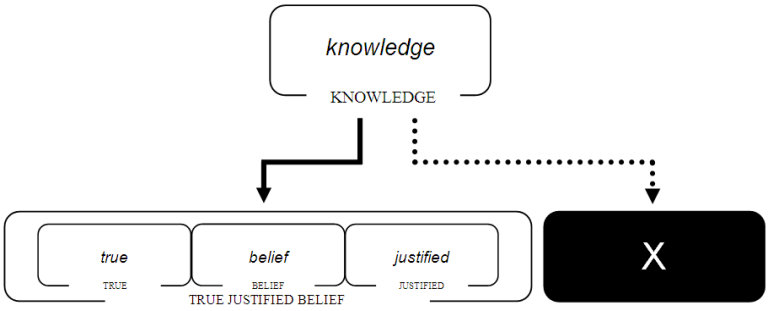

Figure 13.4: The concept KNOWLEDGE might bear a tracking relation to reliability that is not mediated by a representation of reliability.

Something similar might be going on with KNOWLEDGE (Figure 13.4). X is a non-representational state that tracks some property or properties in the world. In other words, X tracks, but X does not represent. On this model, KNOWLEDGE tracks some function of what TRUTH, JUSTIFICATION, BELIEF, and X track. While *knowledge* at least partly cashes out into *true justified belief*, it never cashes out into whatever the non-representational state X tracks. This is why the cases to which the concept KNOWLEDGE seem to apply cannot be captured by any specification of the intension of *knowledge* in terms of what *knowledge* cashes out into, and that's why the brainstorming method and the method of cases are in conflict in the case of KNOWLEDGE.

Indeed, perhaps the non-representational state X additionally tracks truth, justification, and belief. We can then have intuitions about which cases are cases of knowledge while completely bypassing representations of

truth, justification, and belief. The fact that KNOWLEDGE is autonomous allows for this possibility. Perhaps the concept WATER is something like this. When we reflect on water, we might tap into certain theoretical knowledge, and thus unpack water as H_2O. However, when we consider various cases, our judgments about whether or not something is water might not coincide with our judgments of whether it is or how much of it is H_2O. (Barbara Malt (1994) found that subjects' judgments of whether something is water does not track judged percentage of H_2O, but rather other features related more closely to our goals and interests.)

A final possibility that can generate problems when it comes to conceptual analysis is that a concept like KNOWLEDGE might be taken to be primitive or otherwise unanalyzable by its subject. For example, Timothy Williamson's concept of knowledge might be unanalyzable. If he uses the brainstorming method to analyze his concept, he will come up empty-handed. Still, his concept KNOWLEDGE tracks various features, and so, the method of cases will yield results for Williamson similar to the results it yields for other people. Figure 13.5 is a depiction of what Williamson's concept of knowledge might look like.[3]

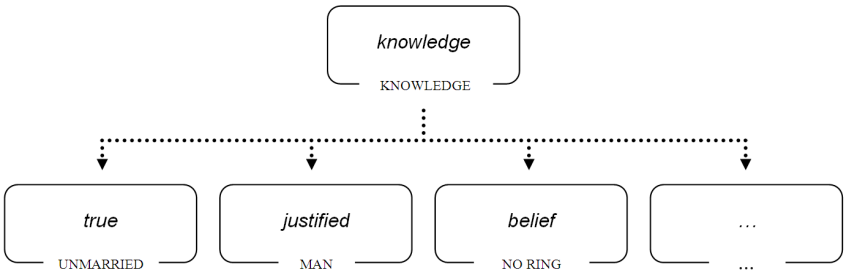

Figure 13.5: Timothy Williamson's concept of knowledge.

Might something similar be the case with BACHELOR? Perhaps we start off with the concepts UNMARRIED and MAN, and form the concept BACHELOR that unpacks into UNMARRIED MAN. But since BACHELOR is autonomous, it is free to form its own tracking and conceptual relations, independently of the tracking and conceptual relations of UNMARRIED and MAN. It might form conceptual relations with availability for marriage, for instance, or tracking relations with males that don't wear wedding rings but not with unmarried males in a long-term domestic partnership. Of course, this story is highly speculative, but the point is that there could be concepts like this.[4]

[3]Further, and relatedly, the efficient concept view allows for Williamson's concept of knowledge to figure in theoretical relationships with other states without TRUTH, JUSTIFICATION, and BELIEF thereby figuring in such relationships.

[4]Concepts obtained through explicit definition are good candidates for such cases. I am reminded of an event in a seminar Karen Bennett taught at Princeton on the metaphysics of mind. While discussing the point that a necessary God would supervene on everything, a graduate student objected, "That doesn't fit my intuitions about supervenience. That

In sum, on the efficient concept view, concepts are autonomous: They can enter into tracking, conceptual, and other relations without their associated concepts thereby entering into such relations. This allows for tracking relations and cashing out relations to come apart, which results in the brainstorming method and the method of cases yielding conflicting results. This is an in-principle limitation of conceptual analysis, given the efficient concept view.

13.3 The analytic/synthetic distinction(s)

Analytic truths are truths that are true in virtue of the content of the concepts involved alone, such as that a vixen is a female fox, while **synthetic** truths are truths that are not analytic, such as that there are foxes in New Jersey. The questions of whether there is a principled analytic/synthetic distinction, and, if so, whether there are any analytic truths are closely related to issues surrounding the legitimacy of conceptual analysis.

The efficient concept view offers something like an analytic/synthetic distinction, although it may not be quite what the proponent of such a distinction was hoping for. First, there is the "easy" case: The efficient concept view allows that the mental contents of various concepts might match the mental contents of other concepts in the way required to generate an analytic truth, perhaps such as in cases such as the thought expressed by "Female foxes are foxes." However, these are not the cases of primary concern, so I will set those cases aside for now.[5] Instead, I will focus on "hard" cases involving distinct concepts with distinct mental contents, such as the thought expressed by "Vixens are female foxes."

The efficient concept view allows us to draw something like an analytic/synthetic distinction in "hard" cases as well. On the efficient concept view, there is a fact of the matter about whether something is part of a concept's derived content relative to some notion of derived content. However, there are a few important deviations from standard conceptions of the analytic/synthetic distinction: On this view, there are multiple analytic/synthetic distinctions, one corresponding to each notion of derived representation. For instance, on one notion of derived representation, it might be analytic that bachelors are unmarried, while on another notion, it might not be (instead, it might be analytic that bachelors are available for dating or marriage). Of course, we might preferentially care about some notions of derived content, such as ones corresponding to folk notions of content, and hence we might care more about some analytic/synthetic distinctions than others. But there is no somehow objectively privileged analytic/synthetic distinction, just as

shouldn't count." His concept of supervenience was obtained by explicit definition, yet it had taken on a tracking life of its own.

[5] "Easy" cases can be straightforwardly understood as cases of Kantian "containment"; the representations and contents of the analysis are literally contained in the representation and content of the analysandum.

there is no objectively privileged notion of derived content, and further, it might turn out that we care about different notions of derived representation and hence different analytic/synthetic distinctions in different circumstances.

The upshot is that the efficient concept view allows for something like an analytic/synthetic distinction, with the crucial proviso that in most cases, at least in the "hard" cases, such a distinction is only relative to a notion of derived content. Whether such a distinction can do all the work it was hoped an analytic/synthetic distinction could do remains to be seen.

13.4 The paradox of analysis

This story about conceptual analysis also resolves the **paradox of analysis**, usually associated with G. E. Moore (1953). The thought expressed by "A vixen is a female fox" is thought to express a correct analysis of VIXEN, and it is also thought to tell us something informative, namely about the content of VIXEN. But it at least initially appears paradoxical for an analysis to have both these features: **correctness** and **informativeness**. If an analysis of a concept is correct, then it must have the same content as the concept of which it is an analysis. So, if the analysis of a vixen as a female fox is correct, then FEMALE FOX and VIXEN must have the same content. But if an analysis specifies the same content as the concept of which it is an analysis, then the analysis should not be informative (e.g. the analysis of a vixen as a female fox should be no more informative than the analysis of a vixen as a vixen, which is trivial).

The efficient concept view provides a ready solution to this "paradox": VIXEN and FEMALE FOX have different source contents, but the same derived contents (on most notions of derived content that we care about). First, the correctness of the analysis expressed by "A vixen is a female fox" might amount to the following: VIXEN and FEMALE FOX have the same derived content (relative to some notion of derived content that we care about). I will argue that this sort of correctness is compatible both with our intuitions of the informativeness of the analysis, and with the actual informativeness of the analysis.

On the efficient concept view, we can explain why we have the intuition that the analysis of a vixen as a female fox is informative, even though we think it is correct. What drives our intuition is the difference in source content of VIXEN and FEMALE FOX. The source content running through our heads when we use the concept VIXEN and the concept FEMALE FOX is different, and so it seems we learn something we were not already thinking when we find out that a vixen is a female fox. (This is related to the fact that conceptual analysis takes work; see 13.1.) In other words, since it is not immediately obvious from the content of the thought expressed by "A vixen is a female fox" that the analysis is correct, we judge that it is informative. We can contrast this case with the "easy" case of the thought expressed by "A vixen

is a vixen," which does not seem to be informative. This thought involves the same concept, VIXEN, twice, arguably contributing more or less the same source content in both cases. We can immediately tell that the thought is true and analytic, but since the same source content is running through our heads *twice*, it does not strike us as informative. This observation further confirms my suggestion that the source of our intuition that the analysis of a vixen is a female fox is information is the difference in source content between VIXEN and FEMALE FOX.

The above explains our *intuitions* about informativeness, given correctness. But there is also a sense in which our intuitions are more or less right: We may not have previously known that VIXEN and FEMALE FOX have the same derived content (relative to some notion or notions of derived content that we care about). This is a new fact that we in some sense learn. Of course, we do not believe that analyses provide information about *derived content*, as opposed to source content, but we do believe that they provide information about something content-like, and since we probably do not distinguish between mental content and derived content for these purposes, we are not too far from the truth (see also Chapter 11 on the folk notion of content).

In this way, the efficient concept view allows for a resolution of the paradox of analysis. It explains how an analysis can be both informative and correct. Notice that molecularism and the complex content view cannot say something similar, since on both views, VIXEN and FEMALE FOX have the same source content, and it is this difference in source content that is crucial for the efficient concept view's resolution of the paradox.

13.5 Reconstruction

In Chapter 5, §5.1.4, I suggested that we have a tendency to reconstruct our concepts. I offered a somewhat speculative explanation of why it is preferable to reconstruct a concept rather than to abandon it and create a new one with a different content: Concepts are connected in various ways to other concepts and states. Preserving these connections is useful. So it is preferable to reconstruct an old concept while retaining its connections rather than abandon it along with its connections, form a new concept, and then forge those connections anew. For example, we believe that touching a hot stove is harmful. If it turns out that our concept HOT fails to refer, we might be tempted to reconstruct our concept to have the content *high mean kinetic energy*. Such reconstruction is preferable to abandoning the old concept HOT and creating a new concept HIGH MEAN KINETIC ENERGY because in the first case we have to reform the belief that touching a stove with the property in question is harmful, whereas in the second case, we can retain that belief.

13.5.1 Reconstruction is easy on the efficient concept view

If we accept the efficient concept view, we can make some more sense of how such reconstruction is possible, and to the extent to which such reconstruction is common, we obtain more support for the efficient concept view.

On the efficient concept view, concepts are autonomous. In §13.2, I emphasized one feature of autonomy: concepts can track things that their associated vehicles do not track. Here, I will emphasize another feature of autonomy: concepts can engage in conceptual relations that their associated vehicles do not engage in. For example, suppose BACHELOR has the derived content *unmarried man* (according to some notion of derived content). Now, suppose BACHELOR is connected to representations of not having wedding rings and being smelly. It will be possible to change the BACHELOR-to-UNMARRIED MAN connection without altering the BACHELOR-to-NO-RINGS connection, as illustrated in Figure 13.6.

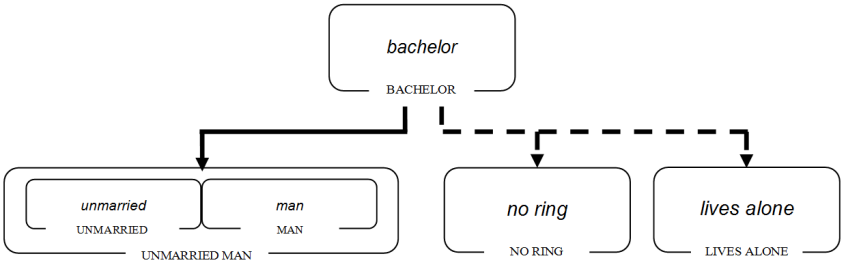

Figure 13.6: The dashed lines represent causal or conceptual relations to concepts that are not also cashing out relations. The concept BACHELOR can form causal and conceptual relations with NO RING and LIVES ALONE without requiring that UNMARRIED and MAN form such relations.

This confers two types of benefit. First, it becomes possible to alter stored information about bachelors without altering BACHELOR's derived contents. Second, and more interestingly, it makes it possible to alter the derived content of BACHELOR without altering BACHELOR's other connections. For instance, by replacing the connection to UNMARRIED MAN with a connection to AVAILABLE MAN we can alter the derived content of BACHELOR without affecting the belief that bachelors don't wear wedding rings; see Figure 13.7.

Again, this is possible because BACHELOR is a distinct representation from UNMARRIED and MAN. So BACHELOR is free to enter into relations with other concepts without requiring that UNMARRIED MAN also enter into these relations. Put otherwise, since BACHELOR, and not UNMARRIED MAN, figures in the belief that bachelors do not wear wedding rings, the belief can survive loss of the concepts UNMARRIED and MAN. We could reform our concept of bachelors to include all robots with the date-seeking switch turned to "on" (perhaps if we discover that all things we previously wanted to call "human

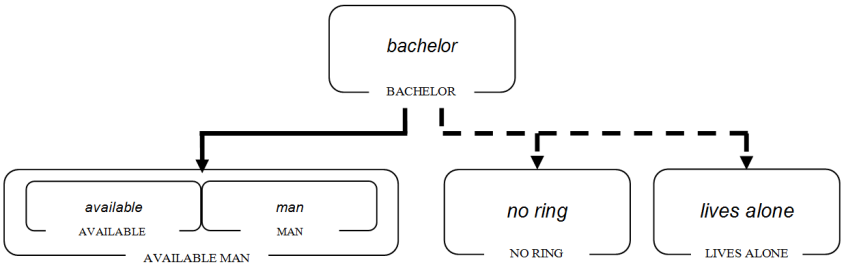

Figure 13.7: The concept BACHELOR subsequent to concept change. Changing the cashing out relations does not affect other causal or conceptual relations.

males" were robots) and this would not affect our beliefs about bachelors and wedding rings.[6]

It is worth mentioning that a story like this fits well with the phenomenology of theory change. Suppose you drastically change your moral views: from a moral absolutist, you become a moral relativist. You have reconstructed your moral concepts. Still, at least most of the time, your first-order moral judgments remain subjectively the same. When someone purposely tortures your cat, you think *That's wrong!* And your thought is phenomenologically the same as the thought you would have had had you remained a moral absolutist. Indeed, your change in moral theory might only manifest itself during specifically meta-theoretical discussions.[7]

13.6 Reconstruction versus discovering more derived content

A further consequence of the efficient concept view is that there is no principled distinction between reconstructing a concept and discovering more derived content that was there all along. In Burgess and Rosen's terminology, we might say that there is no principled distinction between a revolutionary reconstruction and a hermeneutic reconstruction, where a hermeneutic reconstruction is a description of what one meant all along by a term or concept (Burgess and Rosen, 1997).[8] More precisely, such a distinction can only be made relative to a notion of derived content. Suppose that prior to giving the matter much thought, we tend to unpack BACHELOR as *unmarried man*. Then, after considering a few cases such as that of the Pope, we come to unpack BACHELOR as *available man*. Have we now discovered the true content of BACHELOR, or have we revolutionarily reconstructed our concept

[6]Since what makes this possible is that BACHELOR is not composed of its associated vehicles, this benefit is available to the complex content theorist as well.

[7]As an error theorist about color, I can vouch for this.

[8]This might help explain why theorists advocating the same theory about some concept sometimes disagree on whether they are offering a revolutionary or a hermeneutic reconstruction.

(perhaps so that its derived content is more in line with what BACHELOR tracks)? We can only offer an answer relative to a notion of derived content. If the circumstances relevant to determining derived content include ones in which you've considered a multitude of cases, such as that of the Pope, then you have uncovered more derived content. If they don't, then you've revolutionarily reconstructed your concept.

A natural attempt to draw a more principled distinction between revolutionary reconstruction and discovery of derived content is to require revolutionary reconstruction to involve structural changes in the vehicles of representation and/or their relations to one another. We might reason that revolutionary reconstruction usually involves learning or changing one's concepts in some other way, and that involves changes in the vehicles of representation and/or their connections with one another. For example, if thinking about the Pope induces a physical change in the one's brain that weakens the connection between BACHELOR and UNMARRIED MAN, and strengthens the connection between BACHELOR and UNAVAILABLE MAN, we might say that we have a revolutionary reconstruction on our hands. If no such change occurs, then we might say that we are discovering the concept's true derived content.

Although something like this appears fairly reasonable, the strategy does not yield clear-cut results. For one, some notions of derived content include circumstances that involve learning: for example, on some non-crazy notions of derived content, the derived content of my concept ELECTRON is determined by my dispositions to unpack in circumstances in which I have a physics textbook handy. The second obstacle to this way of forging the distinction is that regular unpacking and retrieval of derived content might involve structural changes in the vehicles of representation and/or their relations to one another as well. Still, depending on how the empirical facts pan out, such a strategy might have some success in allowing us to form an interesting, if not principled, distinction.

13.7 Conclusion

If all this is right, we learn something about conceptual analysis. We learn that it is not obvious, that there are predictable tensions and difficulties. We also learn that we can make sense of analytic/synthetic distinctions, but only relative to notions of derived content. This may seem like bad news for projects in conceptual analysis. However, I think it is actually good news. Conceptual analysis is sometimes considered to be in disrepute. However, by recognizing the limitations of conceptual analysis, we can isolate the features that give rise to difficulties, and thereby exculpate some of its other features. For instance, if we recognize that there are multiple analytic/synthetic distinctions, then we do not have to worry so much about disagreement on which are the analytic truths.

Additionally, we can distinguish good and bad uses of conceptual analysis. For instance, if our purposes are to provide a neutral starting point for metaphysical or other debates, then we can settle on a useful notion of derived content and analyze our concepts relative to that notion. This might be useful and legitimate given our purposes. However, if we are trying to find out metaphysical facts about the world from conceptual analysis alone, the in-principle limitations of conceptual analysis might be cause for concern.

This discussion also provides some additional support for the efficient concept view. The perhaps initially puzzling observations concerning conceptual analysis can be explained, and usually best explained, by the efficient concept view. To the extent to which we think the explanations are successful, we obtain further support for the view.

Chapter 14

Conclusion

We started off thinking of mental representation as a phenomenon we notice in ourselves. We notice that we see things, hear things, think things, and so on. The project, then, was to explain *that*. In Part I, I argued that mental representation is a production of the mind or brain, rather than a relation to something else. In Part II, I argued that the phenomenal-intentionality identity theory (PIIT) is true: mental representation is identical to consciousness. Although this theory is not reductionistic in the traditional sense (it does not reduce representation to a physical phenomenon in some unspecified physical theory), it does achieve an identification between consciousness and representation, and thereby reduces the problems in the philosophy of mind by exactly half. More importantly, it is well-motivated and fits the data, making it naturalistic in the sense that matters.

Part III of this dissertation argues for the efficient concept view and discusses some of its applications. Although the efficient concept view, the production view, and PIIT go hand in hand, the efficient concept view does not require PIIT or the production view, and can be independently motivated. According to the efficient concept view, mental representation is scarcer than we may have previously thought. Rather than representing full-blown contents that themselves manage to specify a state's intuitive conditions of truth and reference, concepts represent contents that are fairly impoverished, impoverished enough to match the phenomenal character of thought. Apart from extending PIIT to thought, Part III offers an extended discussion of the applications and implications of the efficient concept view in other areas of philosophy, which may be of interest even to those who do not endorse PIIT or the production view.

14.1 Return to other ways of approaching the problem of mental representation

Although the focus of this dissertation is mental representation, we have shed light on related (and, as I've argued, often conflated) phenomena along the

way. Thus, a major theme of this dissertation is that mental representation does not play all the roles it is sometimes thought to play. Chapter 2 briefly overviews various possible starting points for an inquiry on mental representation and offers reasons for preferring my fairly minimal starting point. The reasons are more or less the same: Mental representation might fail to have the features deemed essential by these alternative starting points. We can now go back and put the problem slightly differently: These alternative starting points define roles that we take something to play. However, what plays these roles are often something other than what we notice in ourselves when we notice that we represent. Let us consider these approaches in turn to see what has become of their favored features.

14.1.1 Folk content

We attribute to ourselves and others contentful states, such as beliefs, desires, and perceptions. These attributions allow us to make sense of and predict each other's behavior, and with surprising success. One approach to mental representation, then, is to take representational content to be **folk content**, the content posited by folk psychology.

In Chapter 2, we dismissed folk psychology as a starting point of inquiry on mental representation, but we can still ask whether anything plays the role of folk content. Having set aside the question of what the notion of folk content represents, our discussion has reached the following account of what it tracks: it tracks a combination of source content, derived content, and referents of derived content. It turns out that this is a particularly useful combination of things to track if what we want to do is predict behavior. We can thus think of folk psychology somewhat as the Churchlands urge: as analogous to folk physics (Churchland, 1981). In most ordinary circumstances, it works, but it does not really accurately capture what's going on in the head. Although some might take this to be disparaging news to folk psychology (and perhaps it should be for those relying on it to approach problems of mental representation), I would instead recommend a more positive attitude. Given our limited access to each other's occurrent thoughts and dispositions to have other occurrent thoughts, we may have reached a near-optimal strategy of predicting the behavior of others. We've conflated various distinct factors, only some of which are genuinely mental, to yield a notion of content that might be close to ideal for the purposes of navigating an incredibly complex social world. In other words, the upshot of our discussion on folk content should not be that we set this notion completely aside, but rather that it has been partly vindicated by our better understanding of it and the success and limitations of the predictions and explanations it allows.

A proponent of folk psychology might object that now we are studying folk psychology itself, not using folk psychology to study something else, which is perhaps what she might prefer. While this is partly true, it is not the whole story. I really intend to be taken seriously when I say that

we should be impressed by folk psychology, despite its shortcomings and limitations. Folk psychology is a good solution to the very real problem of navigating a social world with incredibly limited access to crucial information. Indeed, just as folk psychology is good and useful for certain purposes, other notions of content, perhaps more rigorously defined, might be good and useful for other projects, such as projects in ethics, decision theory, and social psychology, where similar limitations and constraints are in play. These are all areas where we might not be interested in exactly what contents a subject is entertaining at any given moment, but instead would like to attribute more robust contents, perhaps only to make our theories more workable. In some way or another, folk psychology has developed notions of contentful states that respond to similar challenges under similar constraints, and so those wishing to develop notions of derived content for ethics and other domains might have something to learn from the solutions of folk psychology.

14.1.2 Computational content

Computational content is the content that projects in computational cognitive science attribute to "mental" states. One way of approaching the problem of mental representation is as really being the problem of understanding computational content. Although I have dismissed this as an approach to thinking about mental representation, there is still the question of what cognitive scientists are doing when they posit computational content. I have not said much about the extent to which we should be realists about computational content. However, what our discussion in Chapter 7 reveals is that, if there are computational contents, then they are not the same thing as genuine mental representational contents. This is because non-conscious states do not have mental contents, but they might have computational contents. Further, computational contents are generally thought to be a matter of conceptual role, functional role, tracking, and/or the genuine mental contents they give rise to. I have argued that using all these factors for computational content attribution yields conflicting results (see Chapter 7, §7.3.2), but this does not mean that there are not useful notions of computational content, perhaps in limited domains.

14.1.3 Getting around the world

Another approach to mental representation is to think of it as an internal model of the world that we use for the purposes of getting by. We might be moved towards such an approach by our somewhat impressive ability to get what we need, avoid the things that are harmful to us, plan sophisticated actions well in advance of the time for action, and in general survive in a complex world. I have argued that thinking of mental representation as a posit in an explanatory theory of successful behavior risks leaving out the semantic aspects of mental representation, instead focusing on its physical or

syntactic aspects. In our discussion, we have not made any headway on the important question of whether the semantic features of mental representation are causally potent or instead merely correlated with behavior. However, we have shed some light on what it takes to generate successful behavior.

Apart from mental representation itself, one important factor in the generation of successful behavior is the bearing of tracking relations with the world. We manage to get by not only by representing things, but also by our representations tracking important things in the world. The importance of tracking is underscored in mismatch cases, cases where representation and tracking come apart. I have argued that there are many cases like this, and that it might indeed be optimal given certain constraints for tracking and representation to come apart in these cases. We gain no additional benefit when it comes to the successful generation of behavior by representing what it is that we track, rather than representing something else with the very same tracking relations in place. And in certain cases, it might be cheaper, easier, or more efficient to represent something other than what we track, for instance, when what we track is something fairly complex, but whose constituent parts we do not need to keep track of. This might be the case with surface reflectance properties. A simpler example is that of heaviness: The property that heaviness-experiences track might be a relational property holding between objects and (mainly) the earth. However, since the earth's involvement is a constant, we do not need to separately represent it, and so it is sufficient for us to represent something other than this complex relational property, say, heaviness *simpliciter*, even though it is the complex relational property that we track.

Another additional factor that is important for the generation of behavior is unpacking relations. We behave not only based on our occurrent thoughts, but also based on the thoughts they are likely to give rise to, which might merely be a matter of the way we are wired up, as opposed to some genuinely mental features or processes. This proposal might end up being more radical than might at first appear. The process of unpacking is (at least usually) not an intentional, conscious, personal-level, or otherwise mental-like process. Rather, it is more like a mere "stab in the dark," at best allowing for a computational explanation (but, as we already said, computational content is not mental representational content, so this possibility is not relevant in what follows). In other words, we are just wired up in such a way so as to unpack in certain ways, and there could be nothing else interesting that could be said about that. It's not as if the content of one's occurrent state *dictates* or *determines* how a concept is to be unpacked. It is unpacked automatically, unintentionally, and *without the use of representational content.* This means that representational content has at most a limited type of psychological involvement. Much of the work is being done by non-representational states and processes.

Perhaps this might be disappointing. We like to think of ourselves as exhibiting the full-blooded agency that action theory often presupposes: We

like to think that we have full-blown intentions, desires, beliefs, and other such states, whose contents directly interact with one another and lead to the generation of behavior. Action theorists sometimes describe themselves as doing *a priori* psychology, the kind of psychology that would apply to any "agent." Physical processes in the brain might be necessary for our having psychological lives, but they merely play a facilitating or realizing role. If the efficient concept view is correct, this, perhaps comforting, action theoretic approach to understanding human psychology is mistaken.[1,2]

14.1.4 Truth and reference

Another way to approach mental representation is from a theory of truth and reference. We can ask what it takes for a state to be true or to refer and then look for something in us that plays this role of allowing for truth or reference. I dismissed this approach for various reasons, such as that there could be representational states that are not assessable for truth or reference, but that are otherwise the same type of beast as representational states that are thus assessable.

Still, even if this is not our starting point, we might wonder whether and where truth, reference, and truth-conditions enter into the picture, on my view. The short answer is that we have made little headway on these important questions, but this is not a problem for the view.

First, let me say how much headway we have made. We have intuitions concerning the truth-values and truth-conditions of thoughts and the referents or conditions of reference of concepts. From our discussion of derived representation and folk content in particular, we see that these intuitions more closely track what we should say are the truth-values, truth-conditions, referents, and conditions of reference of derived contents or folk contents. The contents occurrently running through our heads are either insufficient to determine conditions of truth and reference (just as *blue* is insufficient to determine truth-conditions), or, if they do determine conditions of truth and reference, they are often the wrong ones (for instance, the truth-conditions of the thought john is a MODAL REALIST might be that some guy has some view). So, in short, the truth-values, referents, and conditions of truth and reference we care about are those determined by the folk contents of mental states.

However, there is a general worry about truth that I have not touched upon, and that may be perceived by some to be a problem for production views generally. This worry has to do with something it seems most of us quite desperately want, an **epistemic connection with the world**. I will deliberately leave this notion vague, as I am not sure just what could satisfy

[1]It is an open question, of course, whether there are other uses of action theory, perhaps, as is sometimes claimed, for the purposes of ethics, that do not require any strong assumptions about our being agents.

[2]Thanks to Derek Baker for endless discussions on what action theory is.

this intense desire apart from metaphysically rubbing our minds all over objects and property instances. However, the general idea is that somehow the states in our head should in some epistemically important way pick out actual or at least possible or potential states of the world that are to count as making those states true or to which they refer. Not just any old relation will do, which is why this relation has to be epistemically important, whatever that amounts to (maybe EPISTEMICALLY IMPORTANT RELATION is a recognitional concept and we'll know it when we see it). Now, consider the situation for relation views of mental representation: They place the mind-world interface between mental states (that is, vehicles) on the one hand and content on the other hand. Once we have a theory of content, an epistemic connection of some sort with the world has been made. Truth and reference fall easily out of such a view: it's more or less just a matter of reading truth- and reference-conditions off of the already-world-involving content of mental states.[3] Production views, on the other hand, locate the mind-world interface between content on the one hand and truth-conditions and reference on the other. So getting a theory of content does not automatically suffice to provide us with an epistemic connection with the world. The connection with the world must instead be made when it comes to truth and reference.

This is sad, since, as I said, most of us want an epistemic connection with the world, and it would have been nice if a theory of mental representation could deliver one. However, here's why this sadness should not sway us from the production view: As argued in Part I, relation views fail: The worldly entities they connect us to are not the ones that we represent.[4] This is another way of putting the mismatch problem. Figuratively put, relation views connect us to the world, but only by disconnecting us from the mind. And so, they do not close the dreadful epistemic gap between mind and world, since they operate only on one side of the gap, the worldly side. They are neither successful as theories of the mind, nor successful as theories of a mind-world epistemic connection. Production views don't close the mind-world gap either, but at least they operate on the side appropriate for a theory of mental representation: the mind side. In other words, they too do not provide theories of a mind-world epistemic connection, but at least they provide theories of the mind. They place the mind-world connection in the right place, that is, in between content and the world, not in between mental states and content, as does the relation view.

[3]Okay, its not so easy, and complications arise, but, setting aside the (I think) insurmountable problem of intentional inexistents, all the necessary ingredients for truth and reference are more or less there.

[4]Relation views might connect us to the world by, e.g. connecting our blue-representation to reflectance property P. However, as I've argued, reflectance property P is not what our blue-representation represents.

14.2 Concluding remarks

On my view, representation just is identical to consciousness. It has many of the features we previously thought consciousness had and lacks a few of the features we previously thought representation had. At the same time, I acknowledge that there are tracking relations, that we might be able to sensibly attribute computational contents, and that there are folk contents that can be predictive of behavior. Thus, I anticipate the following objection: "You are just shuffling definitions around. You use 'representation' to refer to consciousness, and you call other phenomena that we call 'representation' by other names (e.g. 'tracking')."

Here is what is right about this objection: Recall that I initially defined the notion of "representation" ostensively. I took representation to be a phenomenon that we at least sometimes notice in ourselves. The project was to explain that phenomenon. On this way of setting things up, it does turn out that representation doesn't have many features we may have initially thought it had, such as determining the intuitive truth conditions of thoughts. And it also turns out that other phenomena have those features instead. And this means that if we had started off using the word "representation" to refer to whatever has those features, then our conclusion would have been stated differently, and perhaps in a way more agreeable to the objector I have in mind. All this is true, and I acknowledge that some disagreements with certain opponents are at least partly terminological.

However, this does not affect my main argument. Here is a way to state my main claims that makes clear that they extend beyond mere terminological suggestions: Many theorists take it that there is one phenomenon that has many or all of these features: It accounts for how we get around in the world, determines the truth-conditions of our thoughts and other states, is what is attributed by folk psychological explanations and explanations in cognitive science, and is at least sometimes introspectively accessible. My claim is that there is no one type of thing that has all these features. Rather, these various features are features of different, though in some cases related, phenomena. Although I reserve the term "representation" for whatever has the last of these features, this is an inessential part of the big picture. We could have just as easily called what's available introspectively by some other name and reserved "representation" for something else. This does not affect what might be seen as the main contribution of this work, which is to provide a series of arguments supporting the distinctness of the several often conflated phenomena, as well as an account of how they hang together, with a focus on one of them, the one that I take to be most central and most closely related to each of the others and with which what is thought to be the largest mystery of the mind (consciousness) can be identified.

Bibliography

Akins, K. (1996). Of sensory systems and the "aboutness" of mental states. *The Journal of Philosophy*, 93(7):337–372.

Allard, F., Graham, S., and Paarsalu, M. E. (1980). Perception in sport: Basketball. *Journal of Sport Psychology*, 2:14–21.

Allard, F. and Starkes, J. L. (1990). Motor-skill experts in sports, dance, and other domains. *Toward a General Theory of Expertise: Prospects and Limits*, pages 126–152.

Atlas, J. (2005). *Logic, Meaning, and Conversation*. Oxford University Press, Oxford.

Ayer, A. J. (1956). *The Problem of Knowledge*. Macmillan, London.

Bain, D. (2003). Intentionalism and pain. *Philosophical Quarterly*, 53.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–660.

Beal, A. L. (1985). The skill of recognizing musical structures. *Memory and Cognition*, 13:405–412.

Bermúdez, J. L. (2009). The distinction between conceptual and nonconceptual content. In McLaughlin, B. and Beckermann, A., editors, *Oxford Handbook to Philosophy of Mind*. Oxford University Press, Oxford.

Block, N. (1995). On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 18.

Boghossian, P. and Velleman, D. (1989). Color as a secondary quality. *Mind*, 98:81–103.

Boghossian, P. and Velleman, D. (1991). Physicalist theories of color. *Philosophical Review*, 100:67–106.

Bourget, D. (2010). Consciousness is underived intentionality. *Noûs*, 44(1):32–58.

Braddon-Mitchell, D. and Jackson, F. (1996). *Philosophy of Mind and Cognition*. Oxford Blackwell, Oxford.

Braun, K. A., Ellis, R., and Loftus, E. F. (2002). Make my memory: How advertising can change our memories of the past. *Psychology and Marketing*, 19(1):1–23.

Brentano, F. (1973/1874). *Psychology from an Empirical Standpoint*. Routledge and Kegan Paul, London.

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

Burge, T. (1988). Individualism and self-knowledge. *The Journal of Philosophy*, 85:649–663.

Burgess, J. P. and Rosen, G. (1997). *A Subject with no Object: Strategies for Nominalistic Interpretations of Mathematics*. Clarendon Press, Oxford.

Byrne, A. (2001). Intentionalism defended. *Philosophical Review*, 110(2):199–240.

Chalmers, D. (2004a). Epistemic two-dimensional semantics. *Philosophical Studies*, 18:153–226.

Chalmers, D. (2004b). The representational character of experience. In Leiter, B., editor, *The Future of Philosophy*, pages 153–181. Oxford University Press, Oxford.

Charness, N. (1979). Components of skill in bridge. *Canadian Journal of Psychology*, 33:1–16.

Chase, W. G. and Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4.

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

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge University Press, Cambridge.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78:67–90.

Churchland, P. M. (1989). On the nature of theories: A neurocomputational perspective. In Savage, W., editor, *Scientific Theories: Minnesota Studies in the Philosophy of Science*, volume 14, pages 59–101. University of Minnesota Press, Minneapolis.

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

Crane, T. (2008). Intentionalism. *Oxford Handbook to the Philosophy of Mind*, pages 474–493.

Cummins, R. (1994). Interpretational semantics. In Steven, P. S. and Ted, A. W., editors, *Mental Representation: A Reader*, pages 278–301. Blackwell, Oxford.

Cummins, R. (1998). Reflections on reflective equilibrium. In DePaul, M. and Ramsey, W., editors, *Rethinking Intuition*, pages 113–127. Rowman and Littlefield.

Cytowic, R. E. and Wood, F. B. (1982). Synesthesia: I. a review of major theories and their brain basis. *Brain and Cognition*, 1:23–35.

De Groot, A. (1966). *Thought and Choice in Chess*. Mouton Publishers, The Hague.

Dennett, D. C. (2004). *Consciousness Explained*. Gardners Books.

Dretske, F. (1995). *Naturalizing the Mind*. MIT Press, Cambridge.

Ducasse, C. J. (1942). Moore's refutation of idealism. In Schilpp, P. A., editor, *The Philosophy of G. E. Moore*, pages 225–251. Tudor Publishing Company, New York.

Dulany, D. E. (1997). Consciousness in the explicit (deliberative) and implicit (evocative). In Cohen, J. and Schooler, J., editors, *Scientific Approaches to Consciousness*, pages 179–211. Lawrence Erlbaum Associates, Mahwah, NJ.

Dumitru, M. (Ms.). *The Phenomenology of Thought*. PhD thesis, University of Oxford.

Egan, D. E. and Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory and Cognition*, pages 149–158.

Engle, R. W. and Bukstel, L. (1978). Memory processes among bridge players of differing expertise. *American Journal of Psychology*, 91.

Farkas, K. (2008). Phenomenal intentionality without compromise. *The Monist*, 91(2):273–293.

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

Fodor, J. A. (1995). Concepts: A potboiler. *Philosophical Issues*, 6:1–24.

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

Fodor, J. A. (2008). *LOT 2: The Language of Thought Revisited*. Oxford University Press, Oxford.

Foster, J. (2000). *The Nature of Perception*. Oxford University Press, Oxford.

Georgalis, N. (2006). Representation and the first-person perspective. *Synthese*, 150:281–325.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23:121–123.

Gilhooly, K. J., Wood, M., Kinnear, P. R., and Green, C. (1988). Skill in map reading and memory for maps. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 40:87–107.

Greene, J. (2008). The secret joke of kant's soul. In Sinnott-Armstrong, W., editor, *Moral Psychology: The Neuroscience of Morality, vol. 3*, pages 35–79.

Harman, G. (1990). The intrinsic quality of experience. *Philosophical Perspectives*, 4:31–52.

Hebb, D. O. (1949). *The Organization of Behavior*. Wiley, New York.

Heck, R. (2000). Nonconceptual content and the "space of reasons". *The Philosophical Review*, 109:483–523.

Holman, E. L. (2002). Color eliminativism and color experience. *Pacific Philosophical Quarterly*, 83(1):38–56.

Horgan, T. and Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In Chalmers, D. J., editor, *Philosophy of Mind: Classical and Contemporary Readings*, pages 520–533. Oxford University Press, Oxford.

Horst, S. (2008). Naturalisms in philosophy of mind. *Philosophical Compass*, 4(1):219–254.

Husserl, E. (2001/1900a). *Logical Investigations I*. Routledge Press, Trans. J. N. Findley. London.

Husserl, E. (2001/1900b). *Logical Investigations II*. Routledge Press, Trans. J. N. Findley. London.

Jackson, F. (1977). *Perception: A Representative Theory*. Cambridge University Press, Cambridge.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford University Press, Oxford.

Jackson, F. (2004). Representation and experience. *Representation in Mind: New Approaches to Mental Representation*, pages 107–124.

Jackson, F. (2005). Consciousness. *The Oxford Handbook of Contemporary Philosophy*, pages 310–333.

Jessup, R. K. and O'Doherty, J. P. (2009). It was nice not seeing you: Perceptual learning with rewards in the absence of awareness. *Neuron*, 61:649–650.

Johnston, M. (1992). How to speak of the colors. *Philosophical Studies*, 68:221–263.

Johnston, M. (2004). The obscure object of hallucination. *Philosophical Studies*, 120:113–183.

Johnston, M. (2009). *Surviving Death*. Princeton University Press.

Kouider, S. and Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society*, 362:857–875.

Kriegel, U. (2002). Panic theory and the prospects for a representational theory of phenomenal consciousness. *Philosophical Psychology*, 15, No. 1:56–64.

Kriegel, U. (2003). Is intentionality dependent upon consciousness? *Philosophical Studies*, 116:271–307.

Kriegel, U. (2007). Intentional inexistence and phenomenal intentionality. *Philosophical Perspectives*, 21(1):307–340.

Kriegel, U. (forthcoming a). Cognitive phenomenology as the basis of unconscious content. In Bayne, T. and Montague, M., editors, *Cognitive Phenomenology*. Oxford University Press.

Kriegel, U. (forthcoming b). Intentionality and normativity. *Philosophical Issues*, 20.

Kriegel, U. and Horgan, T. (forthcoming). The phenomenal intentionality research program. In *Phenomenal Intentionality*. Oxford University Press.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50:249–258.

Lewis, D. (1974). Radical interpretation. *Synthese*, 23.

Loar, B. (2003). Phenomenal intentionality as the basis of mental content. *Reflections and Replies: Essays on the Philosophy of Tyler Burge*.

Loftus, E. and Palmer, J. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Behavior*, 13:585–589.

Loftus, E. F. and Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*, 25:720–725.

Lowe, E. J. (1996). *Subjects of Experience*. Cambridge University Press, New York.

Lycan, W. (1996). *Consciousness and Experience*. MIT Press, Bradford Books, Cambridge.

Lycan, W. (2001). The case for phenomenal externalism. *Philosophical Perspectives*, 15:17–35.

Machery, E. (2009). *Doing Without Concepts*. Oxford University Press, New York.

Mackie, J. L. (1977). *The Subjectivity of Values*. Harmondsworth, London.

Maddy, P. (2001). Naturalism: Friends and foes. *Philosophical Perspectives*, 15, Metaphysics:37–67.

Maier, J. (2008). *The Possibility of Freedom*. PhD thesis, Princeton University.

Malt, B. C. (1994). Water is not H_2O. *Cognitive Psychology*, 27.

McGinn, C. (1988). Consciousness and content. *Proceedings of the British Academy*, 74:219–239.

McKeithen, K. B., Reitman, J. S., Reuter, H. H., and Hirtle, S. C. (1981). Knowledge organization and skill differences in computer programmers. *Cognitive Psychology*, 13:307–325.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85:207–238.

Mendoza, E. (1961). A sketch for a history of early thermodynamics. *Physics Today*, 14(2):32–42.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63:81–97.

Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy*, 86:281–297.

Moore, G. E. (1953). *Some Main Problems of Philosophy*. Allen & Unwin.

Murphy, G. and Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92:289–316.

Murphy, G. L. (2004). *The Big Book of Concepts*. MIT Bradford, Cambridge.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4):435–450.

Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84.

Papineau, D. (2007). Naturalism. *The Stanford Encyclopedia of Philosophy*, Spring 2007 Edition `http://plato.stanford.edu/archives/spr2009/entries/naturalism`.

Parfit, D. (1984). *Reasons and Persons*. Clarendon Press, Oxford.

Pautz, A. (2007). Intentionalism and perceptual presence. *Philosophical Perspectives*, 21(1):495–530.

Penco, C., Beaney, M., and Massimiliano, V. (2007). *Explaining the Mind: Naturalist and Non-Naturalist Approaches to Mental Acts and Processes*. Cambridge Scholars Publishing, Cambridge.

Perruchet, P. and Vinter, A. (2002). The self-organizing consciousness. *Behavioral and Brain Sciences*, 25:297–388.

Pitt, D. (1999). In defense of definitions. *Philosophical Psychology*, 12(2):139–156.

Pitt, D. (2004). The phenomenology of cognition or what is it like to think that P? *Philosophy and Phenomenological Research*, 69(1):1–36.

Pitt, D. (Ms.). Unconscious intentionality.

Putnam, H. (1975). *The Meaning of "Meaning"*, pages 215–271. Cambridge University Press, Cambridge.

Putnam, H. (1981). *Two Philosophical Perspectives*, pages 49–74. Cambridge University Press, Cambridge.

Putnam, H. (1983). *Why There Isn't a Ready-Made World*, pages 205–228. Cambridge University Press, Cambridge.

Rey, G. (1998). A narrow representationalist account of qualitative experience. *Noûs*, 32(12):435–457.

Robinson, H. (1994). *Perception*. Routledge, London.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233.

Rosch, E. and Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605.

Schellenberg, S. (2010). The particularity and phenomenology of perceptual experience. *Philosophical Studies*.

Searle, J. (1990). Consciousness, explanatory inversion and cognitive science. *Behavioral and Brain Sciences*, 13:585–642.

Searle, J. (1991). Consciousness, unconsciousness and intentionality. *Philosophical Issues*, 1 (Consciousness):45–66.

Searle, J. (1992). *The Rediscovery of Mind*. MIT Press, Cambridge.

Shapiro, L. A. (1997). The nature of nature: Rethinking naturalistic theories of intentionality. *Philosophical Psychology*, 10(3):309–323.

Shoemaker, S. (1994). Phenomenal character. *Noûs*, 28:21–38.

Shoemaker, S. (2003). Content, character and color. *Philosophical Issues*, 13, Philosophy of Mind:253–278.

Siegel, S. (2005). Which properties are represented in perception? In Szabo Gendler, T. and Hawthorne, J., editors, *Perceptual Experience*. Oxford University Press, Oxford.

Siewert, C. (1998). *The Significance of Consciousness*. Princeton University Press, Princeton.

Sloboda, J. (1976). Visual perception of musical notation: Registering pitch symbols in memory. *Quarterly Journal of Experimental Psychology*, 28:1–16.

Smith, E. E. and Medin, D. L. (1981). *Categories and Concepts*. Harvard University Press, Cambridge.

Strawson, G. (1994). *Mental Reality*. MIT Press, Cambridge.

Treisman, A. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136.

Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press, Cambridge.

Tye, M. (2000). *Consciousness, Color, and Content*. MIT Press, Cambridge.

Tye, M. (2009). *Consciousness Revisited: Materialism without Phenomenal Concepts*. MIT Press, Cambridge.

Wickelgren, W. A. (1979). Chunking and consolidation: A theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychological Review*, 86(1):44–60.

Wickelgren, W. A. (1992). Webs, cell assemblies, and chunking in neural nets. *Concepts in Neuroscience*, 3(1):1–53.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford University Press, Oxford.

Winkielman, P., Berridge, K. C., and Wilbarger, J. L. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality and Social Psychology Bulletin*, 31:121–135.

Index