From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Philosophy

2012

Phenomenal Concepts and the Mind-Body Problem

Chappell, Helen Yetter

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01pr76f3469

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

PHENOMENAL CONCEPTS
AND THE MIND-BODY PROBLEM

Helen Yetter Chappell

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

Advisors: Sarah-Jane Leslie, Frank Jackson

September 2012

© Copyright by Helen Yetter Chappell, 2012. All rights reserved.

To my husband and my love, Richard Yetter Chappell.

Table of Contents

Abstract …. iii

Acknowledgements …. iv

Part One: A Theory of Phenomenal Concepts

1 Introduction …. 1
- 1.1 Nature of Consciousness …. 2
- 1.2 Epistemic and Ontological Gaps …. 7
- 1.3 Thinking about Consciousness …. 12
- 1.4 Part One: A Theory of Phenomenal Concepts …. 16
- 1.5 Part Two: What Phenomenal Concepts Can (and Can't) do for You …. 18

2 Introducing Phenomenal Concepts: Indexical And Direct Reference Theories …. 21
- 2.1 Desiderata for a Successful Theory of Phenomenal Concepts …. 22
- 2.2 Two Unsuccessful Theories of Phenomenal Concepts …. 29
  - 2.2.1 Indexical Theories of Phenomenal Concepts …. 30
  - 2.2.2 Direct Reference Theories of Phenomenal Concepts …. 34

3 Constitutional Theory of Phenomenal Concepts and the Problem of No Experience …. 41
- 3.1 Accounting for the Desiderata …. 44
- 3.2 Two Challenges …. 51
- 3.3 Employing Concepts Without the Experience …. 51
  - 3.3.1 Phenomenal Concepts and Mental Imagery …. 52
  - 3.3.2 Two Kinds of Concepts …. 54
  - 3.3.3 Summary …. 60

4 Concept Individuation on the Constitutional Theory of Phenomenal Concepts …. 63
- 4.1 Dispositional Response …. 66
  - 4.1.1 Tye Against the Dispositional Response …. 67
  - 4.1.2 Problems for the Dispositional Approach …. 68
  - 4.1.3 Dualism and Dispositions …. 71
- 4.2 Attentional Response …. 73
  - 4.2.1 Attention and Dualism …. 79
- 4.3 Indeterminate Imagery Response …. 83
  - 4.3.1 Indeterminacy in Mental Imagery …. 89
  - 4.3.2 Indeterminacy in Individuating Phenomenal Concepts …. 93
  - 4.3.3 Psychological Restrictions on Indeterminacy …. 94
- 4.4 A Unified Account? …. 98
- 4.5 Summary …. 101
- 4.6 Echoes of Locke …. 105
- 4.7 Pointer Concepts and Tye's Direct Reference Theory …. 109

5 Phenomenal Concepts and the Mind-Body Problem …. 115

6 Circularity in the Conditional Analysis of Phenomenal Concepts …. 121
6.1 The Conditional Analysis of Phenomenal Concepts …. 122
6.2 Does the Conditional Analysis Yield the Right Intuition? …. 125
6.3 Fleshing Out the Analysis and Finding Circularity …. 132
6.3.1 Vicious Circularity …. 137
6.4 Can Circularity be Avoided? …. 143
6.4.1 Appeal to "the Folk" …. 144
6.4.2 Theories of Reference …. 144
6.4.3 Functionalist Analysis …. 147
6.5 Conclusion …. 153

7 A Psychological Gap: Against the Phenomenal Concept Strategy …. 155
7.1 A Clarification of the Phenomenal Concept Strategy …. 157
7.2 Constitutional Theories and the Phenomenal Concept Strategy …. 158
7.2.1 A Priori Derivation Without Associated Descriptions …. 159
7.2.2 An Alternative Psychology …. 162
7.2.3 A Priori Knowledge …. 169
7.3 Objections …. 176
7.3.1 Only a Problem for Functionalists …. 176
7.3.2 Constitutional Files Aren't Analogous to Phenomenal Concepts …. 179
7.3.3 Quasi-Phenomenal v. Phenomenal Concepts …. 183
7.3.4 An Explanatory Gap Remains …. 185
7.4 Extending the Argument …. 186
7.4.1 Indexical Phenomenal Concept Strategy …. 187
7.4.2 Direct Reference Phenomenal Concept Strategy …. 190
7.5 Conclusion …. 192

8 A Dialectical Stalemate: Making Progress on the Mind-Body Problem …. 196
8.1 A Defense of Epiphenomenalism …. 197
8.1.1 Objection from Introspection …. 199
8.1.2 Objection from Knowledge of Other Minds …. 200
8.1.3 Objection from Evolution …. 201
8.1.4 Psycho-physical Laws …. 203
8.1.5 Paradoxes of Phenomenal Judgment …. 204
8.1.6 Objection from Coincidence …. 212
8.1.7 The Zombie Dualist …. 214
8.1.8 Have Epiphenomenalists Debunked Themselves? …. 215
8.2 A Return to Mary* …. 218
8.3 Overcoming the Stalemate …. 222
8.4 Conclusion …. 228

Bibliography …. 230

ABSTRACT

This dissertation is an exploration of the nature of phenomenal concepts and the work phenomenal concepts can do in solving the mind-body problem. The first half of the dissertation is an attempt to formulate a successful theory of phenomenal concepts. I start by considering what we want out of a theory of phenomenal concepts. I argue that two popular theories of phenomenal concepts (indexical and direct reference theories) cannot satisfy these requirements. And I defend a version of the constitutional theory of phenomenal concepts, according to which our thoughts about conscious experiences literally involve the relevant experiences as constituents. I defend this theory from several hitherto underdiscussed objections: the problem of how we can think about conscious experiences in the absence of the relevant experiences, and a challenge for individuating phenomenal concepts.

The second half of the dissertation assesses the philosophical work that phenomenal concepts can do in solving the mind-body problem. I argue against a posteriori physicalism – the view that the phenomenal-physical truths are only knowable a posteriori. The dominant strategies for defending this view appeal to the special nature of phenomenal concepts. I show that such appeals are misguided, arguing against both the phenomenal concept strategy and the conditional analysis of phenomenal concepts. Out of these arguments emerges a new, and highly compelling way of using phenomenal concepts to defend a priori physicalism, which is sensitive to our own psychological limitations.

But just as the correct understanding of phenomenal concepts reveals a compelling version of physicalism that's immune to the standard anti-physicalist arguments, I also show that phenomenal concepts can help to defend epiphenomenalism. The upshot is that there are compelling and internally consistent version of both physicalism and dualism. The dissertation concludes by assessing where we should go from here: How can we make progress on the mind-body debate in light of this apparent stalemate?

ACKNOWLEDGEMENTS

I'd first like to thank my advisors Frank Jackson and Sarah-Jane Leslie for their constant support, and for helping me to shape and focus my work throughout this dissertation. I don't know how many drafts Frank has read of this work, but he's never seemed to tire of giving me feedback – which I am grateful for! His ability to always be positive and to focus as much on what I've done successfully as on what needs work has both helped me in crafting this dissertation and been a lesson in how to be a wonderful advisor. To this day, I get a thrill from the fact that the guy who created Mary is *my advisor!*

Sarah-Jane has constantly offered new perspectives on the problems I've engaged with (as well providing the valuable service of keeping my armchair psychology in check!) Sarah-Jane's philosophical creativity, both in her own work and what she's brought to my dissertation through her advising, has been an inspiration. In addition to being a wonderful advisor who's supported me philosophically, emotionally, and professionally throughout graduate school, she has also been an amazing role model. (Even my autocomplete tries to turn "amazing" into "amazing.Sarah-Jane"!)

I've also received great feedback through discussions, written comments, and Q&As from professors and colleagues. I'd especially like to thank David Chalmers, John Hawthorne, Mark Johnston, Corey Maley, and Galen Strawson. Each has given me substantial feedback throughout this dissertation and pushed me to think about the material in new ways. I feel privileged to have discussed my ideas with such an amazing group of philosophers.

Many wonderful philosophers have read and discussed various parts of this dissertation with me. I am also indebted to Ashley Atkins, Gilbert Harman, Mark Harris, Terry Horgan, Jimmy Martin, Angela Mendelovici, David Nowakowski, Kristin Primus, Derek Shiller, Jack Spencer, Jack Woods, and audiences at Princeton's Dissertation Seminar, the ANU, NYU, University of Mississippi, Bowling Green State University, Marquette University, Illinois State University, and Trinity College. Thanks, you guys!

Finally, by far my biggest debt is to my husband, Richard Yetter Chappell. Richard has been my emotional anchor throughout graduate school, the dissertation process, and the job market. Your constant love, support, hugs, reassurance that there are no possums, washing of dishes, making of tea, listening to my angry rants about how stupid various theories are … is truly what has kept me going throughout graduate school. In addition to Richard's love and emotional support, Richard has been my biggest and best philosophical critic, reading every chapter of my dissertation probably as many times as I have, pointing out flaws and helping to talk me through solutions. I love you, Richard.

Chapter 1

Introduction

It's a rainy Saturday morning. You lie in bed, flickering between sleep and wake, listening to the rain against your window. You feel warm under your covers. The fabric of your sheets is soft beneath your touch. You consider getting up, but your vivid imagining of the chill of the air, and the hard, cold tiles beneath your feet quickly persuades you against it. Instead you sink deeper into your blankets and focus on the sound of the rain beating against the house in waves. Eventually, you catch a faint smell of fresh brewed coffee. *This is my favorite smell*, you think. Your whole body feels completely relaxed. It's the perfect lazy Saturday morning. You try to hold on to all the comforting sensations and emotions of the morning. Your mind briefly focuses on each sensation, thinking how perfect it is – how perfect this warmth and softness of your sheets is, how perfect this sound of the beating rain is. You wish you could capture this moment in time and hold it with yourself forever.

There's something that it's *like* to lie in that warm bed, feeling the sheets against your skin. There's something that it's like to emerge from your dark bedroom into a brightly lit kitchen. There's something that it's like to smack your shin into your coffee table as you fumble about, blinded by the brightness of your house. There's something that it's like to taste fresh, hot coffee and to feel its warmth spreading through your body. There was something that it was like to *imagine* leaving the warmth of your bed.

This is a dissertation about the nature of what-it's-likes, the way we think about

(or conceptualize) what-it's-likes, and how the way we conceptualize consciousness can inform our theorizing on the nature of consciousness.

What kinds of things are what-it's-likes? Is having a what-it's-like a matter of being in a certain physical or functional states – like being a fire hydrant, being grass, or being an ipad? Or are what-it's-likes something entirely different, something irreducible to physical (or other nonconscious) states? How do we think about consciousness? Could you have a concept of pain without ever having been in pain? Is our knowledge of phenomenal experiences integrated with our physical knowledge? What is it that we're consciously grasping when we think (seemingly substantial) thoughts about what finger nails on a blackboard sound like or what the cold weather outside feels like? And can these latter questions inform our answers to the first set of questions?

1.1 Nature of Consciousness

Let's begin with some background on the first problem: the nature of consciousness. Physicalists take qualia (what-it's-likes) to be reducible to physical-functional properties. According to the physicalist, once you specify the way the world is physically, you've specified the way that the world is simpliciter. [1] In particular, once you've specified the way that the world is physically – the arrangement of every particle of matter and the physical laws that govern their behavior – you've fixed the way that the world is phenomenally. Once you specify that my brain is made of gooey brain matter in

[1] More precisely, physicalism is the thesis that any minimal physical duplicate of our world – world physically like our own, *and with nothing else added* – is a duplicate simpliciter.

functional organization F, you've thereby specified that I'm having a smell-of-coffee experience, as opposed to an red experience or a smell-of-fresh-baked-bread experience.

While this view of qualia is appealing in its simplicity, there is a strong intuition that conscious experiences can't be reduced to arrangements of non-conscious components. Phenomenology seems something wholly different in kind from physical matter or functional arrangements. Intuitively, a fire hydrant or an ipad is a complex physical system, built up out of physical components. But it seems crazy to think that having qualia – smelling the aroma of fresh brewed coffee, hearing middle C, feeling pain – just is a matter of being in a complex physical state. A fire hydrant isn't conscious. It seems unfathomable that simply adding more of the same – more non-conscious physical components – could somehow result in consciousness.

This intuition that consciousness is something wholly different from physical reality is what David Papineau calls the "intuition of distinctness" (Papineau 2002, 161). This intuition is what motivates the competing view of the nature of consciousness. The dualist's ontology embraces the intuition of distinctness: Qualia are – as they seem – a wholly different kind of thing from ipads and chairs. No amount of added complexity of physical systems can yield consciousness from nonconscious physical phenomena. Thus the dualist's ontology holds that there are two distinct sorts of properties – phenomenal and physical – neither reducible to the other, both of which are essential to characterize the world we live in.[2]

[2] Or, alternatively, two irreducible sorts of substance. I'm not concerned with these ontological details. The relevant point is that phenomenology is irreducible.

While the physicalist has simplicity on her side, the dualist has an ontology that conforms to a powerful intuition. To bolster their case, dualists have used the intuition of distinctness to construct powerful arguments against the physicalist: the explanatory gap, the knowledge argument, and the conceivability argument.

Explanatory Gap:

Once upon a time, we did not know that water was H_2O. When we needed water for drinking or washing, we looked for the clear liquid that fell from the sky and flowed through rivers. We later discovered that this clear liquid – with all its familiar properties – had the chemical structure H_2O. This discovery did not make us cry out in disbelief. It was perfectly easy to believe that water was H_2O, because we had an intelligible story about why H_2O should exhibit the properties of water. There was a perfectly good story to be told about why gloms of H_2O molecules should be a clear liquid, potable, with certain boiling and freezing properties. And so we rest easy at night, not plagued by the question of how water could possibly be one and the same substance as H_2O.

But things are not so cheery for the physicalist about qualia. Suppose you think that having a smell-of-coffee experience just is a matter of being in functional state F. We now face the question: *Why* should being in functional state F feel like smelling coffee? Why shouldn't it feel like smelling a fresh baked bread? Or like seeing red? Why should being in functional state F feel like anything at all?

The problem is not just that we do not know precisely which story is the right one

about why functional state F feels like this, rather than like that. The problem is that there is no *conceivable* story that would illuminate this. It seems completely arbitrary why functional state F should feel like smelling coffee. It seems completely arbitrary why functional state F should feel like anything at all. (Levin 1983)

But it's plausible that if Q is a physical process, then it's *possible* to explain Q in terms of physical processes. If this is right, then consciousness – which defies physical explanation – cannot be a physical process. (Chalmers 2007)

Knowledge Argument:

Mary is a brilliant color scientist, living far in the future, who knows all of the physical truths pertaining to color and color experience. She knows all about how ambient light together and the surface reflectance properties of objects interact to affect our retina; she knows all there is to know about the workings of our eyes; she knows all the physical facts about how our brains process the signals they get from the eyes. But Mary has been raised from birth in a black and white room – even her body has been dyed black and white – so she has never had a color experience.

Imagine that one day, Mary's captors bring her a bright red tomato. Mary's reaction would surely be one of astonishment: "So *this* is what red looks like!" Mary has known all along about the different wavelengths of light, and about what was going on in people's brains when they had red experiences. But for the first time, Mary comes to know what *experience* normal people have when they see red. Since captive Mary knew

all of the relevant physical facts, this new fact that Mary learns upon being presented with the red tomato must be a *nonphysical* fact. (Jackson 1982)

So, the argument goes, the physical facts don't exhaust all the facts. In particular, the physical truths don't reveal or exhaust the truths about consciousness.

Conceivability Argument:

It's conceivable that there be a world physically just like our world – full of trees and buildings, H_2O and bugs, and fleshy bodies with brains that are physically and functionally identical to our own – but without consciousness.[3] Such a world would have a physical duplicate of you in all your detail, a physical duplicate of your wife and your dog and your house. But there would be nothing that it was *like* to be your twin. This twin you would feel no pain when he stubbed his toe (though he would curse and cringe in just the way you do). He would have no sensation of pleasure or of softness or warmth when he kissed his wife (though his eyes would look into hers and his mouth would vocalize "I love you" just as yours do). He would feel no tingling anxiety before giving a talk (though his voice would shake and his palms sweat just as yours do). Such a world is just like our world physically, only none of the lights are on, so to speak. Call such a world – physically like our own, but devoid of qualia – a zombie world, and call its nonconscious inhabitants 'zombies'.

When we think such a world is conceivable, we're not making a mistake due to

[3] More carefully, it's conceivable that there be a world that's a minimal physical duplicate of our world, but without consciousness, where a "minimal physical duplicate" is a world physically identical to our own that contains *nothing more* than it needs to to be a physical duplicate.

ignorance of the physical world. It's not the case that *if we only knew more*, we'd see that a zombie world is inconceivable. It seems that no matter how the physical story of the world is fleshed out, it won't rule the zombie world *inconceivable.* And our conviction that such a world is conceivable is intuitively not the result of our own irrationality – as an irrational person might think that it was conceivable that there be a triangle with angles adding up to 170 degrees. We don't seem to be making any rational errors when we imagine a world that is physically like our own, but phenomenally void. The physical way the world is just doesn't seem sufficient to specify the phenomenal way the world is. Thus, the zombie world seems not only conceivable, but *ideally* conceivable.

The conceivability argument then claims that the ideal conceivability of a zombie world entails its possibility. (More on this and necessary a posteriori truths in a moment.) If this is right, then it's metaphysically possible that there be a zombie world – a world physically just like our own, but without consciousness. But if such a world is possible, then being just like our world physically isn't sufficient to be just like our world simpliciter. Physicalism is false. (Chalmers 1996)

1.2 Epistemic and Ontological Gaps

The conceivability argument and the knowledge argument both take the following form: There's an ideal epistemic gap between the physical truths and the phenomenal truths. (A fully rational agent could know all the physical truths, without knowing the phenomenal truths. A fully rational agent could conceive of a world physically just like

our world, but phenomenally void.) Such an ideal epistemic gap entails an ontological gap between the phenomenal and the physical. The explanatory gap – when used as an argument against physicalism – takes a similar form, arguing that the ideal explanatory gap between the physical and the phenomenal entails an ontological gap.

Physicalists have two ways of avoiding the above anti-physicalist arguments: First, they can deny that there is an ideal epistemic/explanatory gap – getting off the boat at the start. The phenomenal-physical truths are knowable a priori, and so a fully rational agent with complete physical information would be able to work out the phenomenal truths a priori. Were such an agent in Mary's position, she would not learn anything (or at least, wouldn't gain any new propositional knowledge) on seeing red for the first time. And such an agent would not find zombies conceivable. This view is known as *a priori physicalism* or *type-A physicalism.* This view typically explains away our intuitions as the product of our own ignorance about the physical world. After all, the phenomenal truths can't be expected to follow from the incomplete subset of the physical truths we happen to know. (A type-A physicalist could also explain them away as the product of our own irrationality. The key point is that a fully rational agent – someone with both ideal rationality and all relevant physical information – would not be ignorant of the phenomenal truths.)

But it's difficult to believe that if we only knew more, we'd somehow see the contradiction in zombies, or to insist that Mary would be unphased when she saw red for the first time and wouldn't gain any propositional knowledge. And we don't seem to be

making any rational errors when we imagine a zombie world. We are not like a mathematical lunatic, who believes she can imagine triangles with angles summing to 170 degrees.

The intuition of distinctness is such a powerful intuition that even many physicalists believe cannot be dismissed so easily. Because of this, most physicalists today take a different line of response[4]: They accept that there is an ideal epistemic and explanatory gap between the phenomenal truths and the physical truths, but they deny that this ideal epistemic/explanatory gap entails an ontological gap. While the phenomenal-physical truths (the entailments from the physical to the phenomenal)[5] are necessary, they are only knowable a posteriori. So Mary, despite knowing all the physical truths, could be ignorant of what it was like to see red. And a zombie world is ideally conceivable. But this does not mean that there are any facts over and above the physical facts or that being physically just like our world isn't enough to fix the way the world is simpliciter. This view is known as *a posteriori physicalism* or *type-B physicalism.*

The type-B strategy may seem the obviously right response for the physicalist. The critic of physicalism says "It's conceivable that there be a world like our world physically, but not like our world phenomenally. Therefore, it's possible that there be a world that's physically identical to our own, but phenomenally different!" But on the face of it, this is a terrible argument. Haven't we all learned from Kripke that conceivability

[4] Of philosophers of mind who take zombies to be metaphysically impossible, 64% took zombies to be conceivable, according to the 2009 philpapers survey: http://philpapers.org/surveys

[5] Throughout the dissertation I use the phrase "phenomenal-physical truths" to refer to what the physicalist takes to be the entailments from the way the world is physically to the way that it is phenomenally. Note that the physicalist is not committed to there being connections the other way around from the phenomenal truths to the physical truths.

doesn't entail possibility? It's conceivable that something be water and not be H_2O, but we don't think that this is possible. It's tempting to think that we could explain phenomenal-physical truths in the same way as these standard cases of the necessary a posteriori: The phenomenal-physical truths are just another standard case of necessary a posteriori truths, to be added to water and H_2O.

Unfortunately, response for the type-B physicalist is too quick. It's commonly accepted that this appeal to familiar cases of the necessary a posteriori will not do, for three reasons. First, there are Kripkean worries. I pick out water via a contingent mode of presentation, describing the water role: "clear, odorless, ...liquid around here". When I conceive of a world in which water is not H_2O, I imagine that the stuff that plays the water role is not H_2O. This is a genuine possibility. (Twin Earth is such a world.) Plausibly, it's the existence of this genuine possibility, in which the watery stuff is not H_2O, that explains the a posteriori status of 'water=H_2O'. But the same explanation is not available for phenomenal-physical identities. Suppose the physicalist holds that pain *just is* physical-functional property P, but wants to accept that it's *ideally conceivable* that there be a zombie world, in which there's physical-functional property P, but no pain. To give a parallel justification to the water-H_2O case, she'd need to accept that there's a genuinely possible world in which there's P, but nothing plays the pain role (the role of *hurting*, of *feeling painy)*. But to play the pain role *just is* to be pain. 'Pain' unlike 'water' doesn't refer via a contingently associated role. Its role is a necessary one: Feeling painy, hurting, just is what it takes to be pain. So if there's a genuinely possible world in which

there's P, but no painy feelings, then there's a genuinely possible world in which there's P, but no pain. And the physicalist cannot accept this.

Second, and relatedly: Though I can conceive of a world in which water is not H_2O, it's plausible that if I knew all the physical facts about the actual world, I would not be able to imagine this (compatible with my actual knowledge). If I knew all of the physical facts, I would be in a position to determine a priori the macroscopic behavior of bodies of H_2O, and so to conclude that H_2O plays the water role and is water. By contrast, if there is an epistemic gap, I could know all of the physical facts and yet still be able to conceive of a zombie world (compatible with my actual knowledge).[6]

Third, though there is an epistemic gap between water and H_2O, there is no explanatory gap. Once we learn that water is H_2O, we are not left with a nagging feeling that the identity is somehow arbitrary. We do not continue to wonder *why* water is H_2O, *how* they could possibly be identical. But, as Levine has made familiar, we do wonder these things about phenomenal-physical identities (Levine 1983).

So the epistemic and explanatory gaps between the phenomenal and the physical can't be explained in the same way as standard cases of necessary a posteriori truths. If the phenomenal-physical truths are necessary but only knowable a posteriori, we need a *distinctive* explanation of why they have this status. What is it about the way we think about conscious experiences that results in this special epistemic gap? Since this story is to be unique to phenomenal-physical truths, it must focus on the different ways we think about conscious experiences, aiming to give a physicalistically acceptable account of the

[6] See: Jackson *From Metaphysics to Ethics* and Chalmers "Consciousness and Its Place in Nature"

special epistemic gap we find in the case of consciousness.

This strategy – for giving a physicalistic account of the special epistemic gap that we find in the case of consciousness – is known as *the phenomenal concept strategy* (PCS). According to PCS, we have two very different ways of thinking about conscious experiences: using phenomenal concepts (e.g. what it's like to see red, what it's like to feel pain) and using physical concepts (e.g. neural state 438921). Because phenomenal concepts are so radically different from physical concepts, even an ideally rational agent with complete physical information could not work out the phenomenal-physical entailments a priori. Despite this, phenomenal concepts and physical concepts still *refer* to the same states. So we have a *conceptual* dualism between the phenomenal and the physical, but an ontological monism. (This basic story is fleshed out with different physicalistically acceptable accounts of phenomenal concepts.) While this strategy cannot *close* the epistemic and explanatory gaps, it aims to do the next best thing: to give a physicalistically respectable story of why the gaps arise.

1.3 Thinking about Consciousness

Most physicalists today acknowledge that the intuition of distinctness cannot be simply ignored. A compelling physicalistic story must either respect the intuition or give a satisfactory account of why we find the intuition so compelling. This requirement has lead to an increasing interest in how we *think about* consciousness – and, in particular, in the question of whether we should expect to find a conceptual gap between our physical

and phenomenal ways of thinking about the world. This heightened interest in how we think about consciousness has illuminated a host of other questions concerning how we conceptualize consciousness: What's the connection between our past experiences and our ability to think about experiences? What is it that you're *consciously grasping* when you think about the delicious warmth of your bed? Is there a connection between the capacity for mental imagery and our ability to think about phenomenology? How do thoughts about phenomenology behave when thinking about counterfactual and counteractual worlds?

These are the questions this dissertation seeks to answer. The dissertation has two parts: In the first, I explore what we want out of a theory of how we think about consciousness, and try to develop an adequate account. In the second, I turn to the philosophical implications of how we think about consciousness, and in particular to its implications for the mind-body debate. I argue against type-B physicalism, showing that appeals to how *we* think about consciousness cannot defend the claim of an *ideal* epistemic gap. Out of these arguments against type-B physicalism emerges a new and highly compelling way of being a type-A physicalist, which aims to explain away the intuition of distinctness as a product of how beings psychologically like *us* think about consciousness. In addition to showing how phenomenal concepts can support a compelling version of physicalism, I also argue that phenomenal concepts can help to defend dualism (an in particular, epiphenomenalism) from intuitive and epistemological objections. While I lean towards dualism, I argue that neither physicalists nor dualists will be able to get traction against their opponents on grounds that their opponents will

accept. I conclude by considering how to make progress on the mind-body problem in light of this apparent stalemate.

Two Concepts of Pain:

What is it that we are asking when we ask how we think about consciousness? We have two ways of thinking about conscious experiences, which are intuitively very different. For example, we can think about pain in terms of what it feels like (using phenomenal concepts), or we can think about pain in terms of tissue damage, physical behavior, and brain processes (using physical concepts). The knowledge argument gives a nice illustration of this difference.

By hypothesis, Mary knows all the physical truths about color: She knows how ambient light together with the surface reflectance properties of objects interact to affect our retina, how our brains process the signals they get from the eyes, and so on. Despite having never had a red experience, Mary can still *refer* to such experiences, using her physical knowledge. She might, for instance, think of 'the experience caused in normal observers by objects with thus-and-so surface reflectance properties in thus-and-so ambient light', or of 'the experience normal observers have when in thus-and-so brain state'. These concepts – these chunks of thought she uses to think about redness – are *physical concepts*.

But when Mary's captors bring her a bright red tomato for the first time, Mary will not only have an experience she's never had before; intuitively, Mary gains a new way of *thinking about* red experiences. Prior to this point, Mary could only think about

red experiences relationally, by way of some associated role. But once Mary has had an experience of red, she can also think about phenomenal redness by latching onto its distinctive character; she can think about phenomenal redness in terms of what it's like to experience it: "*This experience* is something I have never had before!" This new way of thinking about red experiences that Mary gains on seeing red for the first time is a *phenomenal concept*. It doesn't merely refer to phenomenal experiences, but categorizes these experiences in terms of what it's like to have them.

While discussion of "phenomenal concepts" naturally leads to questions of the nature of concepts, we can set this question aside for purposes of this work. Our focus here is not how we carve thought up into concepts, but simply the fact that we can have thoughts about phenomenal experiences that differ in nature from the thoughts Mary is able to have before and after leaving the black and white room. The question "How should we account for phenomenal concepts?" can be interpreted as equivalently to "How can we account for the cognitive difference that Mary undergoes when leaving the black and white room?"

Using this terminology, the aim of this work can be stated as follows: To defend an adequate account of phenomenal concepts and to assess what philosophical work can be done by the correct theory of phenomenal concepts. In the remainder of this chapter, I offer a sketch of the dissertation.

1.4 Part One: A Theory of Phenomenal Concepts

In **chapter two**, I explain what we should want out of a theory of phenomenal concepts, and argue against two popular theories of phenomenal concepts. The first half of the chapter lays out a list of eight desiderata that any successful theory of phenomenal concepts should be able to meet. Many of these requirements focus on respecting special epistemic features of phenomenal beliefs: our apparent infallibility regarding certain phenomenal beliefs, the substantialness of phenomenal knowledge, the unique relationship of acquaintance that phenomenal concept possession seems to require with its referents, and so on. The final desideratum concerns the need for a theory of phenomenal concepts to either respect or explain away our dualist epistemic intuitions (the intuition that there is an epistemic and explanatory gap between the phenomenal truths and the physical truths). Whether physicalists can satisfy this last desideratum is the focus of much of the second half of the dissertation.

In the second half of the chapter, I turn my attention to two popular theories of phenomenal concepts: indexical theories (Perry 2001, Bigelow & Pargetter 2004) and direct reference theories (Levin 2007, Schroer 2010). I argue that are both unable to satisfy the desiderata.

Chapter three introduces the constitutional theory of phenomenal concepts, defended recently by Katalin Balog (2012a) and David Papineau (2002). According to the constitutional theory, when captive Mary sees the red paper and attends to her new experience, the experience is not merely the referent of her phenomenal concept. It is a

constituent of the concept, serving as the concept's cognitive significance and fixing the concept's reference. Phenomenal concepts might take the form: "the experience __" where the blank is filled in by a token of the relevant type of experience. As Mary stares at the paper, she thinks to herself "*This experience *red42** is something I've never had before." The red experience is not detached from her thoughts, but is a part of them.

I take this to be a highly attractive theory, which has the potential to give us what we want out of a theory of phenomenal concepts. After motivating this theory and showing that it has the potential to meet all of the desiderata, I turn to consider two serious challenges to the theory: the problem of how to individuate phenomenal concepts, and the problem of how we can employ phenomenal concepts in the absence of the relevant phenomenal experiences. Chapter three addresses the easier challenge of explaining how we can employ phenomenal concepts in the absence of the relevant experiences.

Chapter four tackles the much more challenging question of how to individuate phenomenal concepts. I consider three possibilities: a dispositional approach, an attentional approach, and an approach that relies on the possibility of indeterminacy in mental imagery. While I argue that none of these approaches can do the required individuating work by themselves, they each have an important role to play in phenomenal concept individuation, and in tandem, they provide a satisfactory solution. This chapter involves a lengthy discussion of indeterminacy in mental imagery. I argue that mental imagery can be indeterminate in quite radical ways, such that we might have

a mental image that's, e.g., determinately red, but not determinately any particular shade of red. I think this is an interesting result independently of its role in individuating phenomenal concepts.

1.5 Part Two: What Phenomenal Concepts Can (and Can't) do for You:

Chapter five returns to the potential uses of phenomenal concepts for defending type-B physicalism, and spells out two ways of doing this in detail. The first is the phenomenal concept strategy, and the second is the conditional analysis of phenomenal concepts. The phenomenal concept strategy takes phenomenal concepts to be so different from other concepts in structure that an a priori derivation of the phenomenal-physical truths is impossible. It's argued that this *conceptual* dualism is compatible with physicalism. The conditional analysis of phenomenal concepts posits that phenomenal concepts have a *conditional* structure: referring to nonphysical properties if there are such properties, and to physical properties otherwise. This is alleged to account for the a posteriori status of phenomenal-physical truths, as we can't rule out the possibility of nonphysical properties a priori. The following two chapters argue that neither of these strategies can succeed.

Chapter six develops a novel argument against the conditional analysis, arguing that when fully fleshed out, the analysis either becomes viciously circular or collapses into a particularly unattractive version of analytic functionalism.

In **chapter seven**, I turn my attention to the phenomenal concept strategy. I argue that no version of this strategy can succeed. According to the phenomenal concept

strategy, phenomenal concepts are so different in nature from physical concepts that a derivation of the phenomenal-physical truths is impossible. This is not just supposed to be impossible for creatures like us, but for any creature that has these phenomenal concepts. Even an *ideal* creature with the relevant concepts would be incapable of such a derivation. This accounts for the ideal conceivability of zombies, for Mary's learning something, and for the explanatory gap. I argue that no account of phenomenal concepts is sufficient to generate an *ideal* epistemic gap. The strategy is to design a creature's psychology from the ground-up, such that if it possessed phenomenal concepts of the appropriate structure, and the relevant physical information, it would be in a position to directly work out the phenomenal-physical truths. I then argue that this sort of derivation would constitute a priori knowledge. The result is that no story of phenomenal concepts can yield an *ideal* epistemic gap (since creatures of the appropriate psychology would not find the gap). But since *we* are clearly not psychologically wired as the creature I've devised, we also find an explanation of why *creatures psychologically like us* find an epistemic and explanatory gap. This yields a highly compelling way to be a type-A physicalist, which I argue respects our dualist intuitions.

If this is right, it should shift the focus of the mind-body debate: Most of the literature responding to anti-physicalist arguments has centered around the question of whether there's an *ideal* epistemic gap. This question has been thought of as another way of getting at the question of whether we should accommodate or simply reject the epistemic intuitions that ground anti-physicalist arguments. Because of this physicalists

have gone in one of two ways: (1) insist that there is no ideal epistemic gap and reject the dualist epistemic intuitions, or (2) accommodate the dualist epistemic intuitions by holding that there's an ideal epistemic gap. The first strategy faces the obvious challenge that it just seems crazy to deny these firmly held epistemic intuitions—we should at very least be able to explain them away! The second strategy faces the challenge of defending an ideal epistemic gap, without collapsing into dualism. This has generated a huge amount of debate over the relationship between ideal conceivability and possibility.

If my argument is right, the physicalist can walk a middle ground between these two positions, retaining the virtues of each without the problems. The physicalist can accommodate our dualist epistemic intuitions by holding *not* that there's an ideal epistemic gap between the phenomenal and the physical, but that there's an unbridgeable *psychological* gap. Thus they can both respect our most firmly held intuitions and avoid messy debates about the relationship between conceivability and possibility.

We have seen that the correct understanding of how we think about consciousness can reveal a compelling version of physicalism, with the ability to answer the dualist's challenges. In this **chapter eight**, I argue that it can also help to reveal a coherent and compelling version of dualism. The first part of this chapter offers a defense of epiphenomenalism from intuitive and epistemological problems. I argue that dualists and the physicalists of the right varieties will not be able to get any traction against one another, and consider how (and whether) we can make progress on the mind-body problem in light of this.

Chapter 2

Introducing Phenomenal Concepts: Indexical and Direct Reference Theories

In this chapter we begin our exploration of what we want out of a theory of phenomenal concepts and our search for an adequate theory. We have already given a gloss as to what phenomenal concept is as the sort of concept Mary gains upon seeing red for the first time, which enables her to think about redness directly, not in terms of an associated role. As noted, my focus is not to adjudicate the terminological question of what deserves title 'phenomenal concept' – perhaps there are several related concepts, some sparser, some more substanative that all have a claim on the title – but rather to give an account of how Mary comes to change cognitively on leaving the black and white room.

We have already seen two unique features of Mary's cognitive change on leaving the black and white room: (1) She becomes able to think about what red is like *directly*, whereas before she could only think of it relationally, and (2) there's an intuition that Mary would not be able to integrate this new way of thinking about what red looks like with her old ways of thinking about it. The first half of this chapter fleshes out this list of desiderata for an adequate theory of phenomenal concepts. The second half of this chapter looks at two rival theories of phenomenal concepts: the indexical theory and the direct reference theory. After motivating and explaining each theory, I argue that neither is capable of meeting the desiderata laid out in the first half of the chapter.

2.1 Desiderata for a Successful Theory of Phenomenal Concepts[7]

Experience Required: Intuitively, only subjects who have had or are currently having a phenomenal experience of the appropriate type can have the corresponding phenomenal concept. Captive Mary cannot have a phenomenal concept of redness because she has never experienced phenomenal redness. Phenomenal concepts are unique in this respect: Having a concept of high blood pressure or red hair doesn't require that the subject have had high blood pressure or red hair.

It should be noted that this intuition might not be quite accurate. Presumably quasi-memories of phenomenal redness, implanted by a scientist, could suffice for a phenomenal concept of redness. Similarly, a different sort of creature might have innate phenomenal concepts, as though it were born with quasi-memories of phenomenal experiences. It's not obvious that these serve as counter-examples to the intuition, as it seems plausible that until the said quasi-memories or innate concepts triggered the appropriate sort of mental imagery, the person would not have the relevant concept. (Imagine a person with an implanted quasi-memory of phenomenal redness, who's never recognized an instance of phenomenal redness – as that would require a red experience – or imagined what it would be like to have such an experience – as that would again involve something qualitatively akin to the experience[8]. If they ever are confronted with a red experience, they will appropriately classify it as such. But is it really correct to say that they have the concept prior to this time? It's far from clear.) What is clear is that in

---

[7] These criterion are not new—they are a synthesis of criteria discussed throughout the phenomenal concepts and consciousness literature. They have been particularly informed by Katalin Balog's attempt to formulate a similar list of desiderata (2012a), with some important additions.

[8] I'll argue that mental imagery is qualitatively similar to perceptual imagery in chapter 3. Suffice it to say for now that there's something awfully *reddish* about my mental image of a firetruck.

normal human cases, like the case of Mary, having had an experience of the relevant sort is essential for having a phenomenal concept.

Ability to Imagine and Recognize: Possession of phenomenal concepts seems to have a tight connection to the ability to imaginatively recreate phenomenal experiences and to recognize instances of the experience. If I can't think about what it's like to have a red experience, then I'm not in a position to generate red mental imagery or to recognize phenomenal redness. Once I can do both these things, I'm doubtless able to think about red experiences in terms of what it's like to have them. This relationship between imaginative capacity and phenomenal concepts may simply be grounded in the fact that both require the subject to have had the relevant experience, or there may be a more direct connection. For instance, imaginative recreations of phenomenal experiences might play a reference-fixing role for the concepts. If this were the case, the ability to imaginatively generate red experiences would be a crucial part of the possession of a red concept.

While it's clear that phenomenal concept possession has a tight connection to the ability to imagine and recognize the appropriate phenomenal experiences, our theory of phenomenal concepts should not take this commitment too far. It's obvious that we can employ phenomenal concepts even when we are not actively employing either of these abilities. For example, I can truly think "I am not in pain or imagining being in pain." When I do this, I am neither imagining pain nor am I actively recognizing an instance of pain.

Epistemic Features: Call beliefs in which a subject applies phenomenal concepts to her own current state 'first-personal phenomenal judgments'. These beliefs seem to have distinctive epistemic features. First, when someone else is in pain, I can only know this by inference (from their behavior, a brain scan, etc). When I am in pain, I can know this directly. I no more have to infer that I'm in pain than I need to infer that I'm thinking P. This stems from the special immediate sort of acquaintance I have with my own experiences, which I'll return to later.

Second, we at least *seem* to be infallible in our first-personal phenomenal judgments. It doesn't seem that any third-personal beliefs can coherently override your belief e.g. that you're currently in pain. There are apparent counterexamples to this infallibility intuition—I might misclassify an intense cold as a burning sensation; my phenomenal experiences might be changing so rapidly that my judgment lagged behind my experience, causing me to misidentify a current experience as the kind experiences a moment earlier—but a successful theory of phenomenal concepts should at least aim to explain why we have this compelling intuition. Perhaps the intuition needs to be qualified in certain respects; it certainly shouldn't be ignored.

Acquaintance: The unique epistemic features of phenomenal belief seem to stem from the unique relationship we stand in to our phenomenal experiences. Phenomenal experiences are not detached from me. I don't "stand at a distance" from my phenomenal experiences—as Chalmers puts it—but am intimately connected to them. As such, I don't

have to rely on *causal* connections to phenomenal experiences in order to form phenomenal judgments. I can form them directly. We might call this unique relationship *acquaintance* (Balog 2012a, 16; Chalmers 2003, 27-35). This relationship of acquaintance may help explain the immediacy and apparent infallibility of our phenomenal judgments, and (as we shall see in the next desideratum) with the *substantialness* of our knowledge of phenomenal experiences.

Substantialness of Phenomenal Knowledge: Intuitively, when Mary's captors show her a red piece of paper for the first time, Mary learns something. It's open to philosophers to deny this, or to try to explain the intuition away. But it should be pointed out that along with this intuition comes another intuition: the intuition that the knowledge Mary gains is cognitively *substantial*. Intuitively, Mary's new intimate acquaintance with phenomenal redness doesn't merely give her a new mental tag or way of pointing at phenomenal redness. It gives here a new and rich way of thinking about phenomenal redness. Just as philosophers can deny that Mary gains propositional knowledge when she sees red for the first time, they can deny that this knowledge is cognitively substantial. (This seems a strange position to take, since the motivation for granting Mary new propositional knowledge is presumably grounded in a respect for our intuitions. Nevertheless it's a coherent position.) But if the intuition is not respected, it must at least be explained away.

Super-Rigidity: When I introduced the notion of a phenomenal concept, I wrote

that even captive Mary could refer to red experiences, using a concept like 'the experience caused in normal observers by objects with thus-and-so surface reflectance properties in thus-and-so ambient light'. We might imagine that this concept is rigidified: 'the experience actually caused in normal observers by objects with thus-and-so surface reflectance properties...' Thus, in every counterfactual world, captive Mary's concept will refer red experiences. But if we imagine that Mary was actually in a world where normal observers have a different experience (say, a green experience) as a result of these surface-reflectance properties, Mary's concept would refer (rigidly) to green experiences.[9] This is just to say that captive Mary's concept is rigid in the familiar way that 'water' and 'gold' are rigid: It refers to the same thing in every counterfactual worlds, but its reference shifts across counteractual worlds.

Now think about Mary's *phenomenal* concept of redness: the concept that characterizes red experiences in terms of what it's like to have them, by latching onto their distinctive character. This concept is also rigid: In a counterfactual world where my counterpart was having a green experience, *my counterpart* might associate the word 'red' with her experience. But it's clear that *my* concept of phenomenal redness does not apply to this experience. (Similarly, my twin on Twin Earth might call XYZ water, but this doesn't mean that *my concept* 'water' applies to it: It doesn't.)

But many philosophers think that my phenomenal concept is not merely rigid in

[9] Suppose captive Mary has just been shown two pieces of paper, one red and one blue, but has not been told what colors the papers are. Mary calls the experience she has looking at the red paper 'A', and the experience she has looking at the blue paper 'B'. Mary would find it epistemically possible that 'the experience caused in normal observers by objects with thus-and-so surface reflectance properties' is A. She would also find it epistemically possible that this experience was B. When we consider as counteractual the epistemically possible world in which A is the relevant experience, A is the referent of the concept. When we consider as counteractual the epistemically possible world in which B is the relevant experience, B is the referent of the concept. Reference across counteractual worlds varies.

the familiar way. Consider Mary's thought "*This experience* (phenomenal redness) is something I've never experienced before". In this thought, Mary characterizes *this experience* in terms of what it's like to have it, by latching onto its distinctive nature. In doing so, she seems to rule out as an epistemic possibility '*this experience* having a different qualitative nature: It is not epistemically possible that *this experience* she is latching onto is phenomenal greenness. Any epistemically possible world that Mary could consider as counteractual is a world at which *this experience* is phenomenal redness. Like 'water' and 'gold', phenomenal concepts have stable referents across counterfactual worlds. But unlike these natural kind terms, they also have stable referents across counter*actual* worlds. In 2-D terminology, they have stable secondary and *primary* intensions. In this way, phenomenal concepts are like shape concepts. (In every epistemically possible world, my concept 'circle' refers to circles.) Chalmers and others have called this feature 'super-rigidity'.

It should be noted that this desideratum is not wholly uncontentious. In particular proponents of a conditional analysis of phenomenal concepts argue that phenomenal concepts have a conditional structure that makes their reference dependent on the nature of the actual world. These philosophers don't dispute that the referents of our phenomenal concepts are stable across counteractual worlds *in respect of their introspectively accessible natures*. But they hold that the underlying metaphysical nature of our phenomenal experiences (whether it is physical or nonphysical) varies across epistemically possible worlds. I'll argue against this view in chapter 6. But it's worth noting here that there's a sense in which even these philosophers don't dispute the claim

of super-rigidity, namely, that these concepts are stable across counteractual worlds in respect of introspectively accessible natures.

Vagueness: Many phenomenal concepts seem to be vague. 'Red', 'soft', 'salty', 'bright': each of these terms presumably inherits its vagueness from the corresponding phenomenal concepts. But not all phenomenal concepts are vague. Consider: VISUAL EXPERIENCE, COLOR EXPERIENCE, PHENOMENAL EXPERIENCE. None of these concepts seem to admit borderline cases. An experience either is a visual experience or it isn't. A full theory of phenomenal concepts should accommodate these facts.

Respecting or Explaining Away Dualist Intuitions: Finally, as we have seen, the main reason that phenomenal concepts recently have generated such interest from philosophers of mind is that many physicalists think that they can offer a way to accept the compelling epistemological claims that dualist arguments are based on, while blocking the move to dualistic metaphysical conclusions.

I'll explore whether an account of phenomenal concepts can meet this task of showing physicalism to be compatible with an ideal epistemic gap in chapter 7, concluding that it cannot. But even if an account of phenomenal concepts cannot give a physicalistically acceptable explanation of why there is an ideal epistemic/explanatory gap, as type-B physicalists require, philosophers should all be able to agree that an adequate theory of phenomenal concepts needs to *either* be compatible with respecting dualist intuitions, *or* be compatible with a debunking explanation of why we find these

intuitions compelling. Given the uniqueness of our dualist intuitions in philosophy of mind (we don't find corresponding intuitions of epistemic/explanatory gaps compelling in other domains, save perhaps the moral domain), it's plausible to think that the distinctive nature of our phenomenal concepts will have an important role to play in either explaining or explaining away these intuitions. This brings us to the final desideratum for an adequate theory of phenomenal concepts: It must either respect or help explain away the intuition that there is an ideal epistemic and an ideal explanatory gap between the phenomenal and the physical truths.

2.2 Two Unsuccessful Theories of Phenomenal Concepts

Now that we've seen what a successful theory of phenomenal concepts will need to account for, we can turn to evaluate whether any of the current theories of phenomenal concepts can meet the task. Three general classes of theories have emerged: indexical theories, direct reference theories, and constitutional theories. Each accepts that phenomenal concepts don't refer via contingent mode of presentation—avoiding the problems faced by the water/H_2O model—and aims to give an alternative physicalistically acceptable explanation of the gaps.

In this chapter, I'll present the indexical and direct reference theories of phenomenal concepts. I'll argue that neither theory can meet the desiderata for an adequate theory of phenomenal concepts. In the following two chapters I'll present and defend a version of the constitutional theory of phenomenal concepts, arguing that it can successfully give us everything we want out of such a theory.

2.2.1 Indexical Theories of Phenomenal Concepts

Indexical theories take phenomenal concepts to function as internal demonstratives. According to these theories, when I have an experience of phenomenal redness, to bring the experience under a phenomenal concept is to form an internal demonstrative pointing to *that experience (whatever it is)*. Reference to the experience is rigid, picking out the same experience in each counterfactual world (just as reference to "that (pointing at a cup)" is rigid). But, had I *actually* been demonstrating a different experience, the concept would have referred to a different experience (just as "that" would refer to a bowl had I actually been pointing at a bowl).

Once we have rejected the analogy to standard cases of the necessary a posteriori, the indexical theory looks quite attractive. Mary knows all the physical facts, but doesn't know what it's like to see red. John Perry's amnesiac, lost in the Stanford library, knows all the physical facts, but doesn't know "I am in the Stanford library".[10] Self-locating knowledge doesn't pose a problem for physicalism. So it's natural to suppose that Mary's ignorance might be captured in an analogous way. According to the indexical theory, when Mary sees red for the first time and learns what phenomenal redness (which she's had a relational concept of all along) is like, she's learning something analogous to the amnesiac learning "I am in the Stanford library" or to someone who knows all the

[10] Imagine that there are two different amnesiacs. One is lost in the Stanford library, the other in the Harvard library. Neither one knows who he is or which library he is lost in. They might both, at the same time, find books giving the complete physical description of the world. The books describe two amnesiacs, one lost in this library, one lost in that. It describes their surroundings, and how they've both just found books with the complete physical description of the world. They might read the whole book, and learn that So-and-so is in the Stanford library and So-and-So is in the Harvard library. Despite all this, they could still fail to know that *I am in the Stanford (or Harvard) library*. The complete physical description alone doesn't tell you where *you* are located within that description.

physical facts about Fred Dretske learning "*That man* is Fred Dretske".

Among the philosophers who have endorsed indexical theories of phenomenal concepts are John Bigelow and Robert Pargetter (1990), and John Perry (2002) who offers a sophisticated account of the nature of our indexical knowledge.[11]

Problems for the Indexical Theory:

The indexical theory faces three problems: a semantic problem that on this model phenomenal concepts are not super-rigid, an inability (for physicalists) to explain away the explanatory gap, and an individuation challenge.

The main problem for the indexical theory of phenomenal concepts is that phenomenal concepts don't seem to function as indexicals. This is argued compellingly by David Chalmers (Chalmers 2003). If phenomenal concepts are indexical in nature, then when Mary is released, what she learns is simply that whatever experience she's currently having and attending to is the experience she read about in textbooks. If Mary were actually attending to a different experience, the concept would refer to this experience, not to phenomenal redness. But intuitively, Mary isn't learning that *whatever it is* that she happens to be internally demonstrating to is red; she's latching onto the particular qualitative experience that is (and would be regardless of what the world is like) red, and learning that *that* is the thing she read about in textbooks.

Chalmers suggests analogy to bring the problem out: I might point to a shape and think "This is a circle". This thought is cognitively significant. The character of 'this' is such that it refers rigidly to whatever is actually being pointed at. If I were actually

[11] Perry's theory is described in some detail in chapter 7, section 7.4.1.

pointing at a different shape, 'this' would refer to something different. By contrast, 'circle' characterizes the shape by the kind of shape it is. 'Circle' picks out the same sort of thing regardless of what the actual world is like. Similarly, when Mary sees red for the first time, she might think "This (pointing inward) is like such-and-such". In thinking this, she's thinking something cognitively significant. She's thinking that the experience she's pointing to is like such-and-such, where the "such-and-such" characterizes the experience as the kind of experience it is. It's these concepts that latch onto the qualitative character of experiences and characterize the experiences *as the kinds of experiences they are* that are at work in the anti-physicalist arguments. I can latch onto the qualitative character of my experiences and imagine a creature physically like me, but lacking the such-and-such. Mary had complete physical knowledge but didn't know that red was like such-and-such. (Chalmers 2003) But this is not what the indexical theory is giving us. It seems to be missing the point. Another way to put this point is that the indexical theory runs afoul of super-rigidity, as indexicals don't have stable reference across counteractual worlds.

I'll address the question of whether this—or any—theory of phenomenal concepts can explain away the epistemic gap in chapter 7. But note that even if phenomenal concepts having an indexical nature could explain why there's an *epistemic* gap between the phenomenal and the physical truths, it can't explain why we additionally find an *explanatory gap*. Even once we know all the phenomenal-physical truths, intuitively, it will still be mysterious *why* this brain-state feels the way it does. By comparison, in standard cases of indexical knowledge, we find no epistemic gap: Though I cannot know

a priori that "This (pointing) is a cup," once I learn that this is a cup, I don't wonder *why* this is a cup or *how* it could possibly be a cup. There's no mystery, no feeling of arbitrariness. But there is a mystery and a deep feeling of arbitrariness in the phenomenal-physical case. So the the explanatory gap cannot be explained by phenomenal concepts having an indexical nature.

Finally, the indexical theory of phenomenal concepts faces an individuation challenge. According to the indexical theory, the phenomenal concept RED is formed by using an internal *that experience (whatever it is)* to point to a phenomenal red experience. But any red experience that we could internally point to will be of many different types (a particular shade of red, a colored experience, a visual experience). We can now ask, "In virtue of what is this concept a red concept, as opposed to a colored concept or a brick red concept?" We might insist that it's the *redness* that I'm internally pointing to. But how is it that I'm able to point to the *redness* in particular? We can't rely on the internal demonstrative *the reddish aspect of that experience*, for this uses the very phenomenal concept we're trying to explain as part of the demonstrative!

This is a version of the individuation problem that Michael Tye raises for the constitutional theory of phenomenal concepts. I'll discuss this objection at length in chapter 4 (focusing on the constitutional theory, which I take to be a more promising theory over all). There I'll argue that this particular objection is not as devastating as it initially appears.

But while the individuation objection may not prove fatal to the indexical theory, I take the other objections we've discussed—and in particular the lack of super-rigidity—

to show that phenomenal concepts are not indexical in nature.

2.2.2 Direct Reference Theories of Phenomenal Concepts

Direct reference theories take phenomenal concepts to be recognitional concepts, which refer to phenomenal experiences directly, without employing any modes of presentation at all. After seeing several cacti of a certain variety, I might form a recognitional concept "that sort of cactus" or "one of those". Similarly, according to direct reference theories, phenomenal concepts involve the ability to reliably identify experiences as "another one of those." But phenomenal concepts on this view are importantly different from standard recognitional concepts: My recognitional concept of the cactus crucially involves a contingent mode of presentation (the way the cactus looks) which facilitates my recognition of the cactus as "one of those" and gives the concept its cognitive significance. Phenomenal concepts, by contrast, involve no modes of presentation. (It is not that they latch onto their referents using an *essential* mode of presentation—the way the phenomenal experience feels—whereas standard recognitional concepts use a contingent one. Phenomenal concepts on this view have *no modes of presentation whatsoever.*[12])

And though direct reference phenomenal concepts are sometimes described as demonstratives, they do not function as demonstratives standardly do: Phenomenal concepts on this view are individuated by their referents. Hence, unlike on the indexical

[12] I'm not sure what sense can be made of these theories. How can I reliably discriminate without any mode of presentation whatsoever? Surely *something* must be guiding my recognitional abilities. Perhaps the way to make sense of direct reference theories is to take them as simply denying that the basis for my ability to discriminate is *consciously accessible.* We might think of this as a subconscious mode of presentation guiding my recognitional capacity.

view, direct reference phenomenal concepts refer to the same experience in every world considered as actual. Janet Levin, a prominent defender of the direct reference theory, is very clear on this: Phenomenal concepts "have *no* reference-fixing 'modes of presentation' *or* Kaplanian 'characters' that can change reference from world to world" (Levin, 89).

Having rejected the indexical theory for making phenomenal concepts' referents unstable across counteractual worlds, the direct reference theory is a natural next step. Tweak the indexical theory by simply stipulating that phenomenal concepts are super-rigid, and the direct reference theory is a natural interpretation of what results.

Because phenomenal concepts have no modes of presentation, there is supposed to be no possibility of an a priori derivation of the phenomenal from the physical, generating an epistemic gap. There's just no substance to be found in phenomenal concepts that could be used in an a priori derivation.

Many philosophers have defended versions of this theory. Janet Levin (2007) defends a view that is unequivocally a direct reference theory. Robert Schroer (2011) defends an extension of Levin's direct reference theory. And Michael Tye (2003) formerly held a theory that might be appropriately characterized as a direct reference view. A number of philosophers have attributed a direct reference theory to Brian Loar (1990), based on his argument that phenomenal concepts don't refer via contingent mode of presentation. I think this is a mistake. I think it is more plausible to read Loar as rejecting contingent modes of presentation (in favor of essential modes of presentation, along the lines of the constitutional view) than to interpret him as rejecting modes of

presentation.

Problems for the Direct Reference Theory:

The direct reference theory, like the indexical theory, seems unable to account for the explanatory gap. The difference between this and the indexical view is that on this view, phenomenal concepts are a strange sort of demonstrative that lacks a Kaplanian character, referring to the same type of experience regardless of what the actual world is like. But it's hard to see how this could generate the puzzlement as to how phenomenal and physical concepts could possible corefer or the feeling that phenomenal-physical identities are arbitrary.

The direct reference theory also runs afoul of the substantialness desideratum. On the direct reference theory, phenomenal concepts are "thin": Having no modes of presentation, they are cognitively insignificant, simply serving as pointers directed at (and reliably caused by) a certain type of experience. It is this thinness that's supposed to generate the epistemic gap: If phenomenal concepts are devoid of cognitive content, there is no hope of deriving them a priori from the physical. But if phenomenal concepts are to be thin enough to eliminate the possibility of a priori derivation, they seem too thin to explain the substantialness of our phenomenal knowledge. Recognitional concepts are typically cognitively significant only because they refer via some modes of presentation. If, as on the direct reference view, phenomenal concepts have no modes of presentation, they seem to lack any cognitive significance. But surely if anything we might learn is rich and cognitively substantial, it's what phenomenal redness is like!

Levin suggests that the ability to reliably point to experiences of a certain type (and so to have a phenomenal concept) typically involves a vast body of knowledge about how the phenomenal experience relates to other phenomenal experiences. It's this body of knowledge that, on her view, gives the substantialness to phenomenal knowledge:

> In at least *some* (and maybe most) cases, one has to know a lot about the properties in question, in particular their interrelations with others in the relevant quality space, to acquire the concepts—a phenomenon that accounts for the intuition that what one gains in knowing what it's like to see red or feel pain is interesting and substantial. (Levin, 100)

For example, if I am in a position to reliably pick out scarlet experiences, I will tend to have knowledge of how scarlet experiences relate to crimson experiences, red experiences, and so on.

I don't think this response is successful for two reasons. First, not all phenomenal concepts involve knowledge of the experience's relation to other phenomenal experiences. And knowledge involving these phenomenal concepts seem just as robust as Mary's knowledge. Imagine that Sally has never tasted salt before. When she tastes salt for the first time, she might thereby acquire the ability to reliably pick out salty tastes. But this doesn't require her to have knowledge of the relationship between salty experiences and other experiences. (I can't even imagine what such a relationship would be.) Nevertheless, she seems to learn something substantial. She doesn't merely come to point to salty experiences or to reliably point to them when she has them; she learns *what they are like*. The concept she acquires has a mode of presentation, it has cognitive significance: the what-it's-like of tasting salt.

Second, even in cases where we do have vast knowledge of how the phenomenal

property is related to other phenomenal properties, Levin's response fails to explain why our knowledge is substantial. Knowledge of phenomenal properties involves phenomenal concepts. If phenomenal concepts are cognitively insignificant, then knowledge of how different phenomenal properties relate to one another seems to involve relating the referent of one cognitively insignificant concept to the referent of another cognitively insignificant concept along a dimension that is also picked out in a cognitively insignificant manner. For example, knowing that maroon is darker and more purple than firetruck red involves the (cognitively insignificant) concept MAROON being related to the (cognitively insignificant) concept FIRETRUCK RED, by way of the (cognitively insignificant) concepts of DARKER and PURPLE. If it's accepted that these concepts are cognitively insignificant, how could this give us the robust sort of knowledge that Mary seems to gain?

Robert Schroer offers an extension of Levin's defense of substantialness, designed to counter the potential worry that the direct reference theory cannot account for the fact that phenomenal concepts reveal something of the *intrinsic natures* of their referents. While his extension seems unproblematic as far as it goes, it cannot avoid the charges I've leveled against the direct reference theory.

Shroer suggests that phenomenal colors are structurally complex. For example, a particular shade of red might include different portions of red, yellow, blue, and green, and different degrees of saturation and lightness. Phenomenal concepts, on his view, are similarly structurally complex: Each component of the color is referred to demonstratively (a la Levin), but the concept also specifies the quantity of each of these

color components. So my concept of the shade of blue that the sky is right now might be given by: 100% $this_1$ [blue hue], 60% $this_2$ [saturation], 70% $this_3$ [lightness]. Because colors are structurally complex, knowledge of the relationships of different color components to one another will give us knowledge of *intrinsic* features of colors.

But Schroer's view faces the very same problems I raised for Levin. His view, like Levin's, is unable to account for the substantialness of phenomenal knowledge that doesn't provide us with relational/structural knowledge, like our knowledge of what saltiness tastes like. Schroer simply denies that we have substantial knowledge of certain basic phenomenal properties. But this seems implausible. And he too faces the problem that building cognitively insignificant elements into a complex structure does not somehow yield cognitive significance. If RED, BLUE, GREEN, and YELLOW are all cognitively insignificant, then "50% green, 50% yellow" will be cognitively very sparse: too sparse to distinguish it from "50% red, 50% blue". But it's obvious that we have a substantial grasp of yellow-green experiences that distinguishes them from purple ones.

So simple direct reference views like Levin's and Schroer's are unable to account for the substantialness of our phenomenal knowledge or certain dualist epistemic intuitions.

Tye (2003) formerly held a view that might be classified as a direct reference view, but which is more complex the simplistic view I've described, and which seems to avoid these charges. Tye argues that we must distinguish the question "What makes concept C a concept of e.g. phenomenal redness?" from the question "What makes concept C a *phenomenal* concept?" He thinks that concept C is a concept of phenomenal

redness, just in case C is reliably triggered by (and because of) phenomenal redness. But this isn't sufficient to characterize what it takes to be a phenomenal concept. To be a phenomenal concept, C must be laid down in memory as the result of a phenomenal experience and must tend to trigger mental images in response to certain mental activities.

In response to the charge that direct reference theories make phenomenal knowledge cognitively insignificant, Tye can appeal to the mental images that phenomenal concepts tend to trigger: These mental images make it seem that phenomenal concepts carry a rich phenomenology with them. But Tye's view still seems unable to explain why we find an explanatory gap.

Insofar as Tye's theory appeals to the phenomenology of having/imaginatively recreating phenomenal experiences in order to explain the cognitive significance of phenomenal knowledge, it seems more akin to the constitutional theory discussed in the following chapter than to Levin's direct reference theory. But it differs from the constitutional theory in whether phenomenal experiences are assigned a reference-fixing role. Tye thinks it's crucial that experiences not play a reference-fixing role to get around (1) individuation problems and (2) worries that we can employ phenomenal concepts when not having experiences of the relevant sort. While I favor the constitutional theory of phenomenal concepts, introduced in the next chapter, I'll argue in chapter 4 that Tye's view and the constitutional theory have more in common than may be initially apparent.

Chapter 3

Constitutional Theory of Phenomenal Concepts and the Problem of No Experience

According to the constitutional theory, phenomenal concepts are partly constituted by the experiences they refer to. To see the motivation for such a view, consider what happens to captive Mary when she's shown a red piece of paper for the first time. Mary has a new experience that she's never had before. And intuitively, Mary also becomes able to think about red experiences in a new way. Whereas before she could only think about them relationally, e.g. as "the experience that's caused by thus-and-so," she now seems able to latch onto red experiences directly. Mary can now think "*This experience* is something I've never had before." When she does this, she seems to pull the distinctive nature of the red experience into her thought: she latches onto her red phenomenology and uses it to think about her new experience.

The constitutional theory of phenomenal concepts takes this intuition very seriously. It literally makes phenomenal experiences components of our thoughts. According to the constitutional theory, when captive Mary sees the red paper and attends to her new experience, the experience is not merely the referent of her phenomenal concept. It is a *constituent* of the concept, serving as the concept's cognitive significance and fixing the concept's reference. The most popular version of this theory, proposed by David Papineau and Katalin Balog, holds that phenomenal concepts have a quotational structure "the experience __" where the blank is filled in by a token of the relevant type of experience.[13] As Mary stares at the paper, she thinks to herself "*This experience*

[13] For more on the constitutional theory, see: Papineau's *Thinking About Consciousness* and Balog's "Acquaintance and the Mind-Body Problem."

red42 is something I've never had before." The red experience is not detached from her thoughts, but is a part of them.

The "the experience" operator functions to point to an instance of the relevant type of experience (e.g. red42). This concept is (partly) individuated by the experience that is pointed to. We might think of this on the model of quotation marks: Much as ' "dog" ' refers to the word of the type between the quotation marks, so the phenomenal concept will refer to the experience of the type presented between the '*'s of the "the experience" operator.

Finally, because the phenomenal experience is a constituent of the concept, the phenomenal concept carries with it the phenomenology of the experience. Knowledge involving these experiences is supposed to give us unmediated insight into the "essence" of the phenomenal experience. (Balog 2012a, 16).

The constitutional theory is a natural move from the direct reference theory. Initially, it seemed plausible that we could accommodate the necessary a posteriori status of phenomenal-physical truths on the model of water/H_2O. But this did not work, as the a posteriori status of these truths stems from the fact that one of the terms involved refers via a contingent associated description. By contrast neither physical nor phenomenal concepts refer via such a contingent reference-fixer. At this point, another analogy emerges: We know how to deal with the a posteriori status self-locating truths. Perhaps phenomenal-physical truths can be modeled on truths like "I am Helen Yetter Chappell" (uttered by Helen Yetter Chappell). But this analogy fails too: Indexicals and demonstratives by their nature shift reference across worlds considered as actual – this is

the very phenomena that explains self-locating truths unique epistemic behavior. Phenomenal concepts do not shift reference in this way. It's natural from here to suggest: Maybe we could avoid this problem by tweaking the so-called "indexical theory" of phenomenal concepts to give phenomenal concepts a reference that's rigid across counteractual worlds. Phenomenal concepts would thus refer directly (without contingent modes of presentation) to *that* (whatever it is), but they would refer rigidly across counteractual worlds. The problem with this theory was that it made phenomenal concepts cognitively insubstantial. We got rid of contingent modes of presentation and replaced them with *no* modes of presentation.

So the lessons learned thus far are: (1) Phenomenal concepts don't shift reference across counterfactual or counteractual worlds. (2) Phenomenal concepts don't refer via contingent modes of presentation. (3) Phenomenal concepts don't refer with no modes of presentation. The natural conclusion that's left is that phenomenal concepts refer via an essential mode of presentation! And what is essential to pain but the hurtiness of pain? Thus the constitutional theory takes the relevant phenomenology itself to provide the cognitive significance and to secure the reference of phenomenal concepts.

Now that I've described the constitutional theory, and given a gloss of its intuitive appeal, we can consider how this theory fares in accounting for our desiderata laid out in chapter 1. I'll argue—drawing heavily on Katlin Balog (Balog 2012a)—that the constitutional theory does an astonishingly good job of capturing what we want out of a theory of phenomenal concepts. But two unaddressed problems remain for the constitutional theory: a challenge for how we can think about phenomenal experiences

when we're not currently having them, and a challenge for individuating phenomenal concepts. In this chapter (section 3.3), I'll tackle what I take to be the easier problem: how we can think about phenomenal experiences that we're not currently having. In the following chapter, I develop a response to the individuation challenge.

3.1 Accounting for the Desiderata

Experience Required: It seems that (at least in normal cases) only subjects who have had or are currently having a phenomenal experience of the appropriate type can have the corresponding phenomenal concept. If phenomenal concepts essentially involve phenomenal experiences as constituents, it is completely transparent why this should be.

Ability to Imagine and Recognize: There seems to be a tight connection between possession of phenomenal concepts and the ability to imaginatively recreate phenomenal experiences and to recognize instances of the experience. The constitutional view offers two possible explanations of this. First, there is plausibly a tight connection between having had a phenomenal experience and being able to imagine and recognize that experience. These abilities may go along with possession of phenomenal concepts simply because possession of these concepts requires agents to have the relevant experiences. Second, there are good reasons—which will be discussed later in this chapter—for allowing that mental imagery can serve as the constituents of phenomenal concepts (e.g. that a red concept could have the form "the experience: *mental image of redness*"). If this view is adopted, there is a direct connection between imaginative abilities and

phenomenal concepts.

The constitutional theory does face a problem in this vein, which will be discussed in the next section: It's obvious that we can employ phenomenal concepts even when we are not actively employing either of these abilities: I can truly think "I am not in pain or imagining being in pain." When I do this, I am neither imagining pain nor am I actively recognizing an instance of pain. But it seems the constitutional theory, which holds that phenomenal concepts essentially involve phenomenal experiences as constituents, cannot allow this. I'll argue later that the theory can be modified to overcome this difficulty.

Epistemic Features: The first distinctive epistemic feature considered in chapter 1 was an asymmetry in how we can know about our own mental states versus those of others. I can only know about other people's phenomenal experiences by inference. But when I am in pain, I can know this directly. I no more have to infer that I'm in pain than I need to infer that I'm thinking P. Since, on the constitutional theory, my awareness of my mental states involves these very states as constituents, we get direct, non-inferential knowledge of them. Because my awareness of other people's mental states doesn't include *their* mental states as constituents, I don't get this special direct knowledge of their mental states. (Balog 2012a)

The second unique epistemic feature was that we at least *seem* to be infallible in our first-personal phenomenal judgments. It doesn't seem that any third-personal beliefs can coherently override your belief that you're currently in pain. The constitutional

theory vindicates this intuition for a restricted class of phenomenal judgments. On the constitutional theory, my judgment that I'm now having a red experience takes the form: "I'm now having the experience: *red experience*." Because I must have the experience in order to think the thought attributing the experience to myself, I cannot be wrong in my judgment of this. There may be other first-personal phenomenal judgments that I can be wrong about—I might misidentify a sensation as the same sensation I experienced yesterday—but this restricted vindication explains why we have this intuition of incorrigibility. (Balog 2012a; Chalmers 2003)

This follows Tyler Burge's account of our privileged access to our own thoughts, developed in "Individualism and Self-Knowledge". Burge argues that when I think the second-order thought "I'm thinking that P", the first-order thought "P" must be thought as a constituent. Because of this, I cannot be wrong about whether I was thinking the first thought: Whatever thought was filling in P, it is both what I will be thinking and what I will be attributing to myself. (Burge)

One might worry that this way of getting us incorrigibility will result in a trivial sort of incorrigibility, of the sort we have in judging "I am here now". But Chalmers argues convincingly that this is not so: My judgment that I am here now contains no real information about how the world is. When one comes to believe it, one doesn't restrict the epistemic possibilities in any way. By contrast, when I come to believe "I'm having a red experience now," I restrict the set of epistemic possibilities to include only those in which I'm currently having *this experience*. (Chalmers 2003)

Acquaintance: We saw in chapter 1 that these unique epistemic features stem from the unique relationship we bear to our phenomenal experiences. We don't stand at a distance from our experiences, but are intimately connected to them. I've followed Balog and Chalmers in calling this relationship *acquaintance*. Since, on the constitutional view, the way we think about phenomenal experiences involves the experiences themselves as a constituent, there is no "gap" between our cognitive understanding of phenomenal experiences (our judgments concerning them) and the experiences themselves. When I judge that warm/soft experiences are pleasant, the warm/softness that I am thinking about is immediately present in my thoughts, giving me noninferential access to it.

Substantialness of Phenomenal Knowledge: I've argued that Mary's new intimate acquaintance with phenomenal redness doesn't merely give her a new mental tag or way of pointing at phenomenal redness, but rather, gives her a new and rich way of thinking about phenomenal redness. On the constitutional model, phenomenal concepts *carry the phenomenology of the experiences they refer to with them*. This phenomenology is what gives our phenomenal concepts their cognitively significance, and is in turn what makes our phenomenal knowledge *substantial*. When Mary sees red for the first time, she learns that seeing red is *like this *red phenomenology**. In learning this, she rules out all sorts of epistemic possibilities (those on which seeing red has the green phenomenology, the orange phenomenology, and so on), and so gains substantial insight into the nature of the world.[14]

[14] We can remain neutral over whether there were genuinely possible worlds consistent with her physical knowledge that come to be ruled out.

Super-Rigidity: I noted that phenomenal concepts are not merely rigid—referring to the same things in all counterfactual worlds—but are *super-rigid*. That is, phenomenal concepts refer to the same thing across all *counteractual* worlds. When Mary thinks "*This experience* (phenomenal redness) is something I've never experienced before," she characterizes *this experience* in terms of what it's like to have it, by latching onto its distinctive nature. In doing so, she rules out the epistemic possibility of '*this experience'* having a different qualitative nature: It is no longer epistemically possible that *this experience* she is latching onto is phenomenal greenness. Hence, any epistemically possible world that Mary could consider as actual is a world at which *this experience* is phenomenal redness. Its referent is stable across counteractual, as well as counterfactual worlds. (Chalmers 2003, Levin 2007)

On the constitutional view what *makes* my red concept a concept of phenomenal redness, is the particular experience that is embedded in it. The referent of the concept is not determined by some description that could be satisfied by multiple sorts of entities, or by pointing (that could be directed at multiple sorts of entities). The reference is fixed *by the embedded experience*. And the concept is individuated—though only partly, as we shall see in chapter 4—*by that very embedded experience*. There is no room for the concept to come apart from the referent either in worlds considered as counterfactual or in worlds considered as counteractual.

Vagueness: Many (but not all) phenomenal concepts seem to be vague. The

vagueness of ‘red’, ‘soft’, ‘salty’, ‘bright’ presumably inherit their vagueness from their corresponding phenomenal concepts. On the other hand, there intuitively cannot be borderline cases of colored experiences or phenomenal experiences. While a complete account of vagueness in phenomenal concepts would require a dissertation of its own, I think the constitutional theory has the resources to give a non-arbitrary account of vagueness in phenomenal concepts. In the next chapter (section 4.3), I note connections between the potential for indeterminacy in imagery and vagueness in phenomenal concepts, and suggest that this may be able to answer the vagueness challenge.

Respecting or Explaining Away Dualist Intuitions: Finally, the main reason phenomenal concepts have generated such interest is for their potential use in explaining dualist epistemological intuitions. There seems to be an unbridgeable epistemic and explanatory gap between the phenomenal the phenomenal truths and the physical truths. I’ll address the potential of phenomenal concepts to provide a physicalistically respectable way of explaining why we find these gaps in chapter 7. For now, I’ll just give a brief preview of the strategy, and how the constitutional theory is supposed to advance it.

First, the epistemic gap: Standard cases of a priori derivations require one of the terms involved to refer via an a priori associated description. Given a complete description of the microphysical facts, I’m able to determine a priori that water is H_2O, because I am able to determine a priori that H_2O plays the role a priori associated with ‘water’. But on the constitutional theory, phenomenal concepts don’t refer via a priori

associated descriptions—rather relying on embedded instances of the appropriate type of experience. Consequently, there is alleged to be no possibility for a priori derivations of phenomenal-physical truths. This generates the epistemic gap.

The explanatory gap is the result of our phenomenal concepts carrying their phenomenology with them. Physical concepts, such as Mary's concept NEURAL STATE 49, don't carry their phenomenology with them. But phenomenal concepts, by their nature, do. Because of this, physical concepts seem to leave out the "what it's likes". This is allegedly what makes an explanation of such truths seems impossible, generating the explanatory gap. (Balog 2012a, 28-30)

While I think that the constitutional theory is at least as successful in explaining why there is an ideal epistemic/explanatory gap as any other theory of phenomenal concepts, I argue at length in chapter 7 that no theory of phenomenal concepts can succeed in this aim. However, in the process, I argue that phenomenal concepts—on the constitutional theory, as well as others—can give an explanation of why *creatures psychologically like us* find such gaps. I suggest that this is really both all that we can ask for from a theory of phenomenal concepts, and is all that physicalists really need.

The theory is also compatible with dualism being true. If dualism is true, then the phenomenal experiences that partially constitute our phenomenal concepts are not physical. The what-it's-likes that we have cognitive access to when we think about our experiences are irreducible nonphysical properties. Our intuitions that the phenomenal is irreducible are not to be explained away, but embraced as accurate.

3.2 Two Challenges

We have seen that the constitutional theory offers an extremely attractive account of phenomenal concepts, which seems equipped to deliver everything we wanted out of such a theory. I turn now to two serious problems facing this account. First, it seems that on the constitutional theory, it will be impossible for a subject to use a phenomenal concept when they are not having the relevant phenomenal experience. But surely I can truly think "I'm not experiencing phenomenal redness." The second problem, first raised by Michael Tye, is a problem for how to individuate phenomenal concepts. Consider the phenomenal concept of redness. On the constitutional theory, this concept has the structure "the experience: *red*," where *red* stands for an instance of phenomenal redness. Tye notes that each phenomenal experience that we have instantiates many different properties. An instance of red will also be an instance of maroon, dark red, having a color, being a phenomenal experience, and so on. We need an explanation of why "the experience: *red*" is a red concept, as opposed to a scarlet concept, a concept of being colored, or a phenomenal-experience concept. (Tye 2003: 94-95, 101-102)

In this section, I'll address the first of these challenges. I'll argue that this problem is relatively straightforward to overcome, and that the residual difficulty answering this challenge is a more specific version of the individuation problem. In the following chapter, I'll turn my attention to the individuation problem.

3.3 Employing Concepts Without the Experience

According to the constitutional theory, phenomenal concepts contain instances of

the phenomenal experiences they refer to. Red phenomenal concepts contain red phenomenology as a constituent; pain concepts contain pain phenomenology as a constituent. While I've touted the virtues of such a view, it has the highly counterintuitive result that I can't employ a phenomenal concept of pain without *being in pain.* But this is absurd! Surely I can think about painful experiences when I'm not having them. Surely I can truly think "I'm not currently in pain."

3.3.1 Phenomenal Concepts and Mental Imagery

The first move we might make in responding to this worry is to expand the range of phenomenology that can serve as constituents of phenomenal concepts, allowing imaginative recreations of phenomenal experiences to serve as constituents of phenomenal concepts. The concept of phenomenal redness would then require either a red experience or an imaginative recreation of a red experience to fill the gap. (This move follows Papineau (2002).) The benefit of this is that mental imagery, unlike perceptual imagery, can be generated at will.

The viability of this move hinges on the idea that imaginative recreations of redness are qualitatively similar to red experiences themselves. I find this highly plausible. When I close my eyes and imagine a firetruck, there's something *reddish* about the phenomenology. The phenomenology of imagining a piece of music is often so similar to hearing the music that one might not notice if the music they were imagining started playing very softly in the background.

But you might worry that this isn't always the case. Does imagining pain really

hurt? I think even here, the answer is 'yes'. If I concentrate very hard on imagining slamming my hand in a car trunk, there's something about it that echoes the experience itself. Perhaps another way to bring this point out is through reference to empathic reactions. When I see someone in an awkward situation in a movie, I feel uncomfortable. It's plausible that this sort of empathetic reaction results from a tendency to imaginatively transport ourselves into the situations of others. I automatically imagine what it would be like to be in their situation, and so feel uncomfortable myself. Three years ago, as I watched videos of protesters in Iran being shot and beaten, I felt waves of pain. A blow to the back of a protester's head sent echoes of pain through *me* if I attend to it in the right way. If empathetic reactions involve a kind of very vivid imaginative recreation, this suggests that there can be something *painful* about imagining pain.[15]

This reliance on imaginative recreations of phenomenal experiences creates a further problem: I've argued that imaginative recreations of phenomenal experiences are qualitatively similar to the original experiences, such that we can rely on them to provide the cognitive content of concepts that refer to the original experiences. So a concept of *phenomenal redness* might have the form "the experience *imaginative recreation of red*". But what makes this a concept of of *phenomenal redness*, as opposed to a concept of *imaginative recreations of phenomenal redness?*

This is a special case of the individuation problem. The individuation problem is a problem for saying what experiences count as sufficiently similar to the phenomenal "filler" that the concept refers to them. The problem here is how to specify what

[15] Since being in pain is a matter of feeling a certain way, if I imagine pain sufficiently vividly that it's qualitatively very similar to what I experience e.g. when I'm stabbed, we might say that my imagining causes me *to be in pain*.

experiences are sufficiently similar to a "filler" that is a mental image, rather than a perceptual image. I'll postpone my discussion of this problem until the next section, and simply note that *if the individuation problem can be solved, then this challenge can be overcome*. This doesn't create a further problem over and above the problem that the individuation challenge already posed.[16]

So we now have a modified theory, on which phenomenal concepts can involve experiences or imaginative recreations as constituents. Alas, this still doesn't solve the original problem. It seems clear that I can truly think "I'm not experiencing or imagining phenomenal redness." In response to this, proponents of the constitutional theory suggest that there are two kinds of phenomenal concepts: The most fundamental kind of phenomenal concepts involve instances of the experience referred to (or imaginative recreations of that experience) as constituents. We might call these *full phenomenal concepts*. But another sort of phenomenal concept does not require that an appropriate experience serve as a constituent. These concepts refers derivatively, by their connection to full phenomenal concepts. We might call these *pointer phenomenal concepts.*

3.3.2 Two Kinds of Phenomenal Concepts

None of the existing constitutional theories of phenomenal concepts offers a fleshed out account of these pointer concepts. Balog suggests that these concepts are individuated by conceptual roles that link them to full phenomenal concepts (Balog

[16] Depending on how finely phenomenal concepts are individuated, it will either be the case that (1) phenomenal concepts involving imaginative recreations will refer to both imaginative recreations and to sufficiently similar perceptual experiences, or (2) that they will refer only to imaginative recreations. One might find this unpalatable, but if you accept the claim that imaginative recreations of experiences are phenomenally similar to the experiences—as you surely do if you find it attractive to allow mental imagery to serve as a constituent of phenomenal concept—I don't think this should seem too odd.

2012a). Chalmers suggests that the content of these concepts is determined by dispositions to imaginatively recreate experiences of the appropriate sort and to recognize instances of that kind (Chalmers 2003).

I think that Angela Mendelovici's Efficient Concept View provides a useful way of thinking about the relationship between full and pointer phenomenal concepts, and can do so in a way that shows the distinction between full and pointer phenomenal concepts not to be ad hoc. We must first draw a distinction between the *content* and the *vehicles of representation* of our concepts. Vehicles of representation are the things that do the representing: in the case of written language, the squiggles on the page; in the case of our thoughts, the brain-states or nonphysical states that mentally represent. Contents are what the thoughts are about.

We're finite beings, who have remarkably complicated thoughts. Mendelovici argues that if this is to be feasible, the vehicles of representation for our concepts must be relatively simple. Further, she argues that the contents of our thoughts cannot be more complex than the vehicles that represent them. So the contents of our thoughts must also be more simple than we sometimes suppose. Consider the following two sentences:

> *(Conditional$_{EASY}$):* John Hawthorne endorses the conditional analysis of phenomenal concepts.
>
> *(Conditional$_{HARD}$)*: John Hawthorne endorses the view about phenomenal concepts on which they have a conditional structure, such that if the world is merely physical then they refer to physical states, but if the world has nonphysical states of the appropriate sort then they refer to these nonphysical states.

It's *harder* to think *(Conditional$_{HARD}$)* than it is to think *(Conditional$_{EASY}$)*. But it's difficult to see how we could explain this comparative difficulty if the two thoughts involve

equally complex vehicles of representation. Suppose I think *(Conditional$_{EASY}$)*, and you then ask me what the conditional analysis is. It might take me a few moments to be able to answer you. But if the content of *(Conditional$_{EASY}$)* is just the same as that of *(Conditional$_{HARD}$)*, the answer should be immediately available to me: After all, I'd have just thought something with that content! This suggests that the content of the two thoughts differs. Answering this question requires me to "cash out" the concept CONDITIONAL ANALYSIS into a concept with more complex content. And in addition to the time constraints involved in answering questions about the conditional analysis, there's an introspectable difference in the phenomenology of thinking *(Conditional$_{EASY}$)* and thinking *(Conditional$_{HARD}$)*. This adds further support to the idea that there's a difference in content between *(Conditional$_{EASY}$)* and *(Conditional$_{HARD}$)*. (Mendelovici, 131-160)

We might think of phenomenal concepts on this model. It's quite hard to imaginatively recreate phenomenal redness (let alone pain). Most of the time, when I have a thought about phenomenal redness, I don't do this, just as most of my thoughts about the conditional analysis don't have content spelling out a full definition of the conditional analysis. But when I "cash out" my concept of phenomenal redness, it unpacks into a full phenomenal concept, which includes an instance of the type of experience referred to as a constituent. I might do this when probed ("What is this 'phenomenal redness' you're thinking about?") or I might do it automatically, when I attend to the content of the concept. (You might imagine text on a computer, such that when you hover the mouse over a word, it expands into a fuller content.) Once we see the

reasons for adopting Mendelovici's model, we find that it's not ad hoc to posit two types of phenomenal concepts: the need stems from general principles arising from limitations on our cognitive capacities.

What is the cognitive content of these pointer phenomenal concepts? As their name suggests, I take these concepts to simply function as pointers: contentless place-holders that get their cognitive significance derivatively through their connection to full phenomenal concepts. These concepts are individuated by reference to the full concepts they unpack into.

There are two reasons for thinking this. First, when we consult the phenomenology of these concepts, there doesn't seem to be anything substantial to them; they don't seem to reveal anything of their referent. (If this isn't immediately obvious, imagine that someone at the DMV asks you what color your car is. Focus on that thought you have when you answer. You can surely *cash out* the thought "It's black" into something substantial. But the content of the thought that you entertained when you responded, was likely no more cognitively substantial than a a blind pointing towards the parking lot.) If pointer phenomenal concepts are themselves contentless, we need to give some account of why it is that they refer to the phenomenal experiences they do. The full concepts they unpack into provide the answer.

More importantly, we don't want to allow that subjects who have not had the relevant phenomenal experiences can have phenomenal concepts (even if they are only pointer phenomenal concepts that refer to phenomenal experiences in a derivative way). In the case of full phenomenal concepts, we have an explanation of why zombies cannot

have these concepts: They involve phenomenal experiences as constituents, so anyone lacking experiences cannot possess them. But we don't have this explanation in the case of pointer concepts. The obvious way to explain why zombies cannot have these concepts is by appeal to their connection to full phenomenal concepts. If pointer phenomenal concepts are contentful only insofar as they cash out into full phenomenal concepts, it is clear why zombies cannot possess them: They lack the full phenomenal concepts that they rely on.

One might worry that someone could have a phenomenal concept, but not be able to "cash out" the the pointer concept at will. Someone might lack imaginative powers, such that they cannot form a concept that includes an imaginative recreation as a component. Nevertheless, they may be able to have the experience when confronted with the right perceptual stimuli, to reliably recognize it when they had the experience, and to form concepts including these experiences as constituents. Papineau suggests that we're in a situation like this regarding smells. Our ability to classify smells far outreaches our ability to recreate them through imagination. (Papineau 2002, 70-71) Do subjects who lack the imaginative capacities required to cash out a concept at will have pointer phenomenal concepts?

I think it would be plausible either to deny that these subjects possess pointer phenomenal concepts or to grant that the ability to cash out the concepts in certain circumstances (namely, when confronted with the right sort of sensory experience) is sufficient to lend content to the pointer concepts.

One might want to deny that this was sufficient for possession of phenomenal

concepts, on grounds that the pointer concepts would not have the properties we take phenomenal concepts to have: If Joe's concept PHENOMENAL REDNESS doesn't automatically unpack into a full phenomenal concept, involving an instance of phenomenal redness, then it's not clear that Joe—when not confronted with the relevant perceptual stimuli—will find an explanatory gap between phenomenal redness and physical properties.[17] And Joe's phenomenal thoughts seem to be missing something. It's plausible that my thought "Apples look red" involves a phenomenal concept only because, when I attend to RED, I'm disposed to flesh the concept out into a concept that employs an instance of the experience. Joe's concept lacks this; it seems to lack the "what it's like" that's so crucial to phenomenal concepts. If this is right, while Joe would lack a phenomenal concept of red when he's not having a red perceptual experience, he would have a concept of phenomenal redness under certain circumstances, namely, those circumstances when he was having a red perceptual experience. I find it completely plausible that Joe doesn't have a phenomenal concept of redness (when he's not experiencing it). I similarly find it plausible that I don't have a phenomenal concept of the smell of my grandmother's coat closet (when I'm not experiencing it), though I'm sure that I can reliably recognize it.

But it could also be argued that Joe does have the disposition to unpack his concept PHENOMENAL REDNESS in this way, it's just that he can only do the unpacking in a restricted set of circumstances. But why should this mean that it's not a phenomenal concept? When I'm tired or distracted, I might not be able to imaginatively

[17] Perhaps he would remember *that there was* a "what it's like" to redness, just as his current experiences have a "what it's like"\", and could infer that this generates an explanatory gap, even if he was not currently aware of the "what it's like".

recreate phenomenal redness, but surely I don't lack the concept in these circumstances.

I'm not sure there's really a problem here. We can all agree that Joe lacks a full phenomenal concept of red, when he is not experiencing phenomenal redness. I'm not sure there's a substantial issue as to whether his pointer concept PHENOMENAL REDNESS is a phenomenal concept or not. It unpacks into a full concept in certain circumstances. But it doesn't (in most circumstances) unpack into a full concept, and so (in most circumstances) doesn't inherit the interesting properties of full phenomenal concepts (generating an explanatory gap, carrying interesting phenomenology, etc.). This might be all there is to say.

3.3.3 Summary

Let me summarize what I've argued. The constitutional theory seems to make it impossible to employ phenomenal concepts when one isn't actively having the relevant type of experience. But we clearly can employ phenomenal concepts in these circumstances. I've first considered the reply that phenomenal concepts can take *imaginative recreations* of the relevant experiences as constituents. This raises a further question of why a concept e.g. incorporating an imaginative recreation of phenomenal redness, should be a red concept, rather than an imaginative recreation of red concept. This is an instance of a broader problem over how to individuate phenomenal concepts, which I'll consider in the next chapter. But even if the individuation problem can be overcome, this does not solve the original problem, as it's clear that we can employ phenomenal concepts even when we're neither having nor imagining the relevant

experience.

I've noted that there's good reason to think that many of our thoughts involve much simpler contents than one might initially think. We can think incredibly complicated thoughts very efficiently, but it takes time to retrieve all the content buried in these thoughts. Because of this, one might be attracted to a view of concepts like Mendelovici's, on which many concepts have simple vehicles of representation and contents, which can be "cashed out" into more complicated contents when attended to. Similar reasoning pushes one to accept that the phenomenal concepts that we often use in everyday circumstances ("My car is black") are simple. Imagining blackness takes cognitive work, much as thinking about what it takes to endorse the conditional analysis of phenomenal concepts. It's more efficient to use place-holding *pointer* concepts that will unpack into full concepts when necessary. These pointer phenomenal concepts are what we use in every day circumstances. But they inherit their cognitive significance and their unique properties (explanatory and epistemic gap, connection to abilities, etc.) in virtue of their essential connection to full phenomenal concepts.

Finally, it's worth noting that even if you disagree with my claim that mental imagery is phenomenologically similar to perceptual experience, this is not reason to abandon the constitutional view. One could hold a version of the constitutional theory on which full phenomenal concepts necessarily involve perceptual experiences of the relevant sort, where pointer concepts do all the work of explaining how we are able to employ concepts when not having the relevant experiences. We would be in the situation that Joe, in the example above, is in: unable to unpack our pointer phenomenal concepts

at will. But if you are happy with either of the options I offered for dealing with Joe, this isn't a problem.

Chapter 4

Concept Individuation on the Constitutional Theory of Phenomenal Concepts

In this chapter, we turn to the challenge of how to individuate phenomenal concepts on the constitutional theory. Recall that according to the constitutional theory, phenomenal concepts have a quotational structure, "the experience __", where the blank is filled in by a token of the relevant type of experience. So the phenomenal concept red will have the structure "the experience *red experience*", where what fills in the '*'s is not a concept, but an instance of the relevant type of phenomenal experience. On the constitutional theory, phenomenal experiences are not only the referents of our phenomenal concepts, but provide the concepts their cognitive significance and fix their referents. As I look at my teacup, I find myself thinking about the *blueness* of the experience I'm having. In doing so, the blueness of the teacup experience becomes a component of my thought: "this experience: *teacup color experience*". This is what makes my knowledge of phenomenal experiences rich and significant. This is why my concept BLUE is about blue experiences.

Michael Tye raises the following challenge for the constitutional theory: Any phenomenal experience I have, which could serve as a component of a phenomenal concept, will exemplify many different properties. The experience I have right now, as I look at my tea cup, is blue, dark blue, navy, $blue_{591}$, a visual experience, and so on. But now, Tye argues, we need an explanation of why my concept "the experience: *teacup color experience*" is a blue concept, rather than a navy concept or a visual experience concept. (Tye, 101)

This problem arises for every phenomenal concept. It seems that *any* experience that could partially constitute a phenomenal concept will instantiate numerous different phenomenal properties, ranging from the highly particular to the completely general sensory modalities. So simply saying that phenomenal concepts are partially constituted by phenomenal experiences isn't enough to individuate the concepts. We need an explanation of why *this* concept "the experience: *my teacup color experience*" is a blue concept, while *that* concept "the experience: *my teacup color experience*" is a navy concept, and *that* one is a color experience concept.

One might be tempted to respond that the relevant property for concept individuation is the one which the agent is *attending* to: My concept is a concept of blueness, rather than something more or less specific, because I'm attending to the *blueness* of the experience. But Tye argues that this will not do: Focusing on a property isn't like focusing your eyes on a region of space. To attend to a property, I must be able to isolate that property, and abstract it from other properties that the object of my experience instantiates. But to do this presupposes possession of a phenomenal concept! This response seems to presuppose the existence of phenomenal concepts in order to explain how phenomenal concepts are individuated, rendering it circular.

This challenge poses a problem not only for the constitutional theory, but for the indexical theory as well. Recall that on the indexical theory, phenomenal concepts are inner demonstratives (this experience). Contrary to the constitutional theory, these concepts are not individuated by reference to the experiences that are pointed to. But the indexical theory nevertheless faces a version of Tye's challenge. When I employ a

phenomenal concept on the indexical view, we might ask: *In virtue of what am I pointing to the blueness of the experience, as opposed to the navy-ness or the visualness?* Again, it seems that we cannot answer this question by appeal to the quality that the agent is attending to, on pain of circularity.

In this chapter, I will propose and analyze three possible ways of avoiding Tye's individuation challenge. In section one, I consider and refine a dispositional response to the challenge. I then raise several objections to the dispositional account of individuation. Section two considers how we might modify the dispositional approach to avoid these objections. The result is a version of the attentional approach Tye criticizes as circular. I defend this view against Tye's criticisms. Section three considers a radical new approach to phenomenal concept individuation, based on the idea that mental imagery can allow a great degree of indeterminacy. I will focus on the constitutional theory, as the most promising theory of phenomenal concepts. And because the constitutional theory has been advocated by both physicalists and dualists, I will be careful to consider whether the various proposals are open to both sides of the mind-body debate. This will be useful in chapter 8 where I put the constitutional theory to work defending epiphenomenalism.

In the fourth section I argue that none of the theories I've considered can by themselves give us everything we want in a method of individuation. The dispositional method fails to offer a satisfactory method of individuating the bulk of our phenomenal concepts: specific concepts like RED, DARK, A-FLAT, SMOOTH TEXTURED, WARM, and SALTY. And thought the attentional and indeterminate mental imagery views can each capture a great deal of what we want in individuating phenomenal

concepts—telling compelling stories about the relatively specific concepts the dispositional account fails with—neither can offer a complete account of phenomenal concept individuation. Both face difficulty individuating more general phenomenal concepts such as those expressing sensory modalities: VISUAL, TEMPERATURE, TACTILE, PHENOMENAL.

Though this might appear to be an admission of defeat—that there is no uniform way of individuating phenomenal concepts to be had on the constitutional theory—I think that it something we should both expect and embrace. I'll argue that different sorts of phenomenal concepts exhibit different sorts of psychological traits, such that we should not expect or even *desire* a single unified way of individuating all phenomenal concepts. I'll argue that the dispositional, attentional, and mental imagery theories of individuation all work together to give a complete theory of phenomenal concept individuation. I conclude with a brief discussion of the parallels between this discussion of phenomenal concept individuation and the debate between Berkeley and Locke about the nature of abstract general ideas.

In the final section, I consider parallels between the constitutional theory and Tye's version of the direct reference theory.

4.1 Dispositional Response

Ned Block is, as far as I am aware, the only person who has attempted to respond to Tye's individuation challenge. Block suggests that what quality a concept refers to might be determined by what one is disposed to accept as instances of the same concept.

If the subject is disposed to treat any other instance of a bluish color as falling under the same concept, but denies that other non-blue phenomenal experiences fall under that concept, then it is a concept of blue. If the subject is disposed to treat only blue experiences in some more narrow range as falling under the concept, then the concept is more narrow (perhaps a navy concept). (Block, 39)

I'll begin by considering an argument Tye gives against the dispositional approach, arguing that it is unsuccessful. Next I'll present two objections to the dispositional approach that I take to be more problematic. Finally, I'll consider whether the dualist can appeal to the dispositional strategy.

4.1.1 Tye Against the Dispositional Approach

Tye argues that the dispositional account is circular. He argues that to be disposed to treat instances of a bluish color as falling under the same concept, an agent must subsume the instances under the concept HAVING THE SAME PHENOMENAL CHARACTER. This in turn requires that the agent have a concept of HAVING A PHENOMENAL CHARACTER. But this is a phenomenal concept. So, Tye argues, the dispositional account of individuating phenomenal concepts must *presuppose* a phenomenal concept. Thus the account is circular. (Tye 2009, 46)

Tye assumes that for an agent to be disposed to treat experiences e_1 and e_2 as falling under concept C_1, there must be a further concept C_2 that the agent must subsume e_1 and e_2 under. I'm not convinced that this is the case.

Plausibly, we have some sort of innate pattern recognition detection faculty that

operates at a nonconceptual level and enables (e.g.) infants to differentiate their mothers' faces and voices from those of strangers. If this is right, then there's no reason to think that recognition of a commonality between e_1 and e_2 presupposes a concept that they're subsumed under. There might simply be subconscious processing that detects a pattern between, say, the neural correlates of the experiences before they are made conscious.

It would seem hard to deny the existence of such a primitive, nonconceptual pattern recognition faculty: Without such a faculty, we would need to have a set of innate concepts sufficient to ground all further detections of similarity. This proposal seems difficult to take seriously.

4.1.2 Problems for the Dispositional Approach

While I don't think that Tye's problem should worry the advocate of the dispositional theory, two more serious problems face the dispositional account.

First, the dispositional proposal seems to get the order of explanation wrong. Block himself notes this: According to the proposal, it is "because one is taking the experience of a specific shade of green as a green-experience rather than as $\text{Green}_{126,731}$ experience that makes it function as a concept of the green experience rather than the concept of a highly specific shade of green. The dispositionalist view seems to get things backwards" (Block, 40). Intuitively, we take our experience of $\text{Green}_{126,731}$ to be a green experience *because* it falls under our green concept. But on the dispositional view, it's the other way around.

Second, the dispositional proposal seems at odds with the motivation for the

constitutional theory. When Mary leaves the black and white room, she has a new experience. She also becomes able to think about red experiences in new ways. Intuitively, when she does this, she *latches on to the distinctive nature of phenomenal redness*. This intuition, that Mary becomes able to think about redness by latching onto the character of her experience, was part of the intuitive motivation for accepting the constitutional theory over a direct reference theory, on which Mary's new concept is simply cashed out in dispositional terms. By appealing to dispositions in order to individuate phenomenal concepts, we seem to cut against this motivation. It's not an awareness of the *redness* of her experience that makes her concept a red concept on this view: It's the fact that she's disposed to use it in certain ways.

It's worth noting here that this disposition can't rely on a conscious process, whereby Mary consciously matches up her experiences to a generic red experience and determines whether or not they're of the red type. This would require Mary to have a generic red experience, which applied to all experiences of phenomenal redness. But, as Tye plausibly assumes, any phenomenal experience will be of a determinate type, not a general all-purpose-red shade that could serve the described role.[18] So the disposition must be the result of some sort of subconscious processing. A subconscious sorting process in the brain will be responsible for the disposition to treat certain experiences as "of the same type," while excluding others. Perhaps our brains are just wired to respond in a certain way to light in a certain range of wavelengths, and not light in other ranges, such that we're disposed to assent to all and only red things as falling under a particular

[18] I'll question this assumption later. But I think that if the assumption is flawed, one doesn't need the dispositional account to respond to Tye's challenge. There's a far more direct way to respond, taking these generic red experiences to be the fillers of our red phenomenal concepts.

concept.

This highlights the extent to which this method of individuation runs counter to the motivation for the constitutional theory. As Balog describes the constitutional theory, part of its beauty is the ability to account for the intuition that "I have a 'substantive' grasp of [feeling giddy's] nature. I grasp it in terms of *what it's like* to have that experience." (Balog 2012a, 29) But while I may have a substantive grasp of the nature of the specific experience that I'm currently having, on this view, I don't have a substantive grasp of the nature of *giddiness* in particular. What makes my concept the concept that it is isn't anything that I have a substantive grasp of. It's a disposition arising from subconscious workings of my brain.

This point can be made in set-theoretic terms. We want our concept red to apply to a vast set of experiences. Many of the members of this set will vary wildly from one another. On the dispositional account, dispositions are used to give the boundaries of the set. But, if the dispositional method of individuating is correct, I will have insight into the nature of only *one* of these experiences. And this insight will not be what is responsible for determining what other experiences fall into the set. But surely we think that it's the nature of the experience I'm having that enables me to think about the set of all red experiences. Surely we want to determine the set's boundaries by something that I consciously grasp when I have a phenomenal red concept.[19]

One could bite the bullet: accept the backwards order of explanation, and abandon some of the intuitive appeal of the constitutional theory. But I think these considerations are enough that we should hope that there is another way of individuating phenomenal

[19] Thanks to Frank Jackson for suggesting this way of putting the problem.

concepts.

4.1.3 Dualism and Dispositions

Before I move on to the attentional approach to individuation, I want to consider whether dualists who endorse the constitutional analysis would be able to take this route to individuating phenomenal concepts. Dualists don't want to accept that zombies have phenomenal concepts. And the constitutional theory grants this: When I think "this experience: *painful experience*", my concept crucially involves an instance of painfulness. But my zombie's corresponding concept doesn't include painfulness as a constituent. She isn't acquainted with painfulness, and accordingly, she doesn't refer to or think about pain. The dispositions aren't all that matter for possession of phenomenal concepts: You must have the right dispositions, *together with an appropriate embedded phenomenal experience*. But you might think that partial zombies, who have a limited range of phenomenal experiences, could pose a challenge to dualists who want to adopt this view.

Pained Zombies are physically indistinguishable from us, but they lack all phenomenal experiences *except* pain experiences (which they have just the same ordinary people). Suppose I think "Not all phenomenal experiences are painful," and in employing the concept PHENOMENAL EXPERIENCE, the embedded experience is some pain that I'm currently feeling. Since my Pained Zombie Twin is physically identical to me, it seems that she will share my disposition to treat things as falling under the concept 'phenomenal experience'. And since she feels pain just as I do, she will have the very

same phenomenal experience as I do embedded in the concept. According to the current proposal, she has the very concept that I have. But when she thinks "Not all phenomenal experiences are painful," she's intuitively thinking something quite different from me. My concept is something that encompasses red experiences, cold experiences, rough experiences, high pitched auditory experiences, and so on, in addition to pain experiences. Her concept doesn't encompass anything besides pain.[20] So my Pained Zombie Twin and I share everything that matters for possessing a concept of phenomenal experience, according to the dispositional analysis. But intuitively, we *don't* share this concept.

I don't think this argument is successful. The dualist who wanted to adopt a dispositional strategy could counter that this argument stems from a misunderstanding of the dualist's conception of dispositions. I wrote that since my Pained Zombie Twin is physically identical to me, it seems that she will share my disposition to treat things as falling under the concept PHENOMENAL EXPERIENCE. But while this may sound plausible to physicalists, the dualist might hold that she needs a more complex account of dispositions. My Pained Zombie Twin may be physically identical to me, but—because of differences in the psychophysical laws in her world—her mental life differs from mine. As a result, while the same *physical* processes occur in our brains, the processes in my brain result in my being disposed to apply my concept to red experiences, cold

[20] You might be inclined to think that she's employing the very same concept as me, on grounds that she's having a *phenomenal experience*, and so is able to latch onto what's distinctive about phenomenal experiences. But while it's compelling to think that we can latch onto what's distinctive about phenomenal experiences, and thereby have a concept of phenomenal experience, this is not what the dispositional view gives us. On the dispositional approach, the only sense in which we latch onto what's distinctive about phenomenal experiences is in that we have certain (subconsciously driven) dispositions.

experiences, and so on. The physical processes in her brain don't have this result. So our dispositions differ.

Consider an analogy: There's a vase on Earth, which is disposed to shatter when struck. This vase has a qualitative duplicate in another world, where the physical laws are different, such that when you strike the vase in that world, it doesn't break. Although the vases are intrinsic duplicates, this difference in physical laws yields a difference in dispositions. Similarly, the difference in psychophysical laws between my world and that of my Pained Zombie Twin's, yields a difference in our dispositions. Since phenomenal concepts are individuated dispositionally, this difference in our dispositions will result in a difference in our phenomenal concepts, as desired.

4.2 Attentional Response

I now want to try to develop a different response to the individuation problem. This response starts by taking seriously the problems with the dispositional approach, and ends up as an attentional theory of individuation, holding that phenomenal concepts are individuated by reference to the particular qualities of our experiences that we're attending to.

One problem with the dispositional approach to individuation is that it seems to get the order of explanation backwards: Intuitively, I'm disposed to use my concept to apply to all and only red things *because it's a red concept*, not the other way around. But my disposition to use the concept in this way isn't random; it doesn't result through a

magical, inexplicable process. I have the disposition because I have information stored in my head that guides me to apply it in certain circumstances and not in others. We might call this guiding information a *subconscious template.* The first move you might make in addressing the problems of the dispositional approach is to insist that phenomenal concepts are individuated by their connection to subconscious templates: It's the information stored in the template that makes it a red concept. And it's the information stored in the template that is responsible for the disposition.

There's a second way we could rely on these templates to individuate phenomenal concepts, which I think is more attractive, in that it can additionally allow us to get around the second problem of the dispositional account—that it went against the motivation for the constitutional theory. Any template that's capable of guiding me so as to apply a concept to all and only red experiences must capture what's distinctively in common between red experiences (or, if you're a dualist, what's distinctively in common between the brain-states that generate these experiences). Perhaps in addition to our brains *storing* a template with this information, the template might affect the way we perceive experiences. Consider a related case. When I see the duck-rabbit as a duck, my background information affects my phenomenology in a way that makes the experience introspectively different from when I see it as a rabbit or as an uninterpreted pattern of lines. Maybe the template can alter the phenomenal experience I have when I look at a red tomato, such that the *redness* of the phenomenal experience (rather than the red_{42}ness, or the coloredness) jumps out at me. There would then be an *introspectable difference* between a red concept with red_{42} as a constituent, and a red_{42} concept with red_{42} as a

constituent. I would be grasping *the what its like of seeing red* when I employed a red concept.

This gets around the second worry I raised for the dispositional approach, that it seems to run counter to the intuitive motivation for the constitutional theory. On this approach, when Mary leaves the room, she comes to grasp the distinctive redness of seeing red. Further, we now have an alternative way of individuating phenomenal experiences: When I employ the concept "the experience: *red_{42}*", what makes it a red concept is the fact that information stored in my subconscious template causes the *redness* of the experience to pop out at me. It's this conscious grasping of the redness of the experience that is responsible for individuating it. (Following this suggestion, a dualist who imagined a creature physically just like us, but where their red template didn't affect their phenomenology, might argue that this creature's concepts would be different from our own.)

Notice that what we have arrived at is a version of the view that Tye criticized, on which my concept "the experience: *red_{42}*" is a red concept (as opposed to a red_{42} concept) just in case I'm attending to the *redness* of the experience. Tye argued that this view was untenable: Attention to redness (or any quality) seems to require a concept of that quality. But this would be a phenomenal concept. So we're presupposing that an agent possesses phenomenal concepts in order to explain how to individuate their phenomenal concepts. But then, Tye argued, the view becomes circular.

I don't think this view is circular. Tye can call these subconscious templates "phenomenal concepts" if he wants. But they are not phenomenal concepts in the

philosophically interesting sense. They are not phenomenal concepts in the sense that proponents of the constitutional theory are after. Among the things we wanted out of a theory of phenomenal concepts was (1) that they crucially involve a direct *acquaintance* with phenomenal properties, (2) that knowledge involving phenomenal concepts is *substantial* and cognitively rich, and (3) that they can respect or explain away the explanatory gap. While these subconscious templates may be "phenomenal concepts" in the weak sense that they pick out phenomenal qualities, they cannot meet these goals.

(1) I don't have to be—or to ever have been—acquainted with phenomenal redness in order to have a red template. I simply need my brain to store information of the appropriate sort. This may typically be caused by my having been acquainted with phenomenal redness, but there's nothing crucially connecting the two. (2) Phenomenal judgments involving these subconscious templates alone cannot give us the sort of rich, substantial knowledge that Mary seems to gain on leaving the black and white room. (Unless a subconscious template is *activated*—used to generate a conscious instance of the relevant quality—the phenomenal knowledge it generates amounts to no more than a disposition to make certain judgments.) (3) If our phenomenal concepts were nothing more than "subconscious templates", this would not give physicalists their desired explanation of the explanatory gap: These templates will guide us to label certain experiences as 'red experiences'. If physicalism is true, these experiences are nothing more than physical states, that could be described in physical terms. But we have no explanation of why it should be so very *mysterious* why this physical state feels the way it does. We find we have all of the problems that the direct reference theory of

phenomenal concepts faced in chapter 2.

So in offering this attentional response to the individuation challenge, we are not relying on the very concepts we're trying to give an analysis of: We're relying on something we might think of as a proto-phenomenal concept, in order to explain the ultimate target. Since the term 'phenomenal concept' is unclear, it might be useful to think of this exchange without reference to "phenomenal concepts". What we're ultimately after is an explanation of how Mary comes to change cognitively upon leaving the black and white room. There are many surprising ways in which Mary differs once she's released—discussed as "desiderata for a theory of phenomenal concepts" in chapter 1. At present, we want to know why, when Mary has a thought about the new experience she's having, it's a thought about the redness of the experience, rather than some more or less particular quality. In order to explain this, I've suggested that we need to rely on her brain storing information that could guide her in picking out red experiences. But this is not the thing we want an explanation of. So we are not guilty of circularity, however much else must be said.

One might wonder how we get these subconscious templates. Don't we need to tell a further story about *them*? A further story will ultimately need to be told. But this story is best left to the neuroscientists and others who study how our brain is able to detect patterns in the world. It's not something that can be answered from the armchair.

Note that to make sense of the attentional response in a noncircular way, we rely on the very low-level pattern detection capability that we relied on to save the dispositional response from Tye's charge of circularity. In order to have dispositions, we

must store information that guides our dispositions. This is the very information that the attentional account makes use of in individuation. With respect of the circularity charge, the dispositional and attentional views stand or fall together.

Summary:

The information that guided our dispositions on the dispositional account is taken to be capable of affecting our phenomenology, in such a way as to highlight (e.g.) the redness of our red experiences, or the loudness of a sound. This can be thought of by analogy to how background knowledge can alter the phenomenology you have when you see duck-rabbit. Phenomenal concepts are individuated by the qualities of the embedded experiences that "pop out" at you. While this presupposes some sort of abstraction mechanism, it does not presuppose phenomenal concepts in the philosophically interesting sense, or in any sense that is not also presupposed by the dispositional analysis. This view has the virtues of getting the intuitively correct order of explanation, and of fitting with the intuitive motivation for the constitutional theory.

One might question whether introspection really reveals that attention to a particular quality of our phenomenal experience can alter the overall phenomenology. Is the phenomenology of attending to the redness of a red_{42} experience different from the phenomenology of attending to the red_{42}ness of a red_{42} experience? I think introspection at the very least doesn't rule this out, and may positively supports such a possibility. Focus your eyes on something blue-green, and attend to the *blueness* of that thing. Now

attend to its *greenness*. When you shift your attention, the phenomenology changes as well.[21] This supports the idea that activation of a template can alter embedded phenomenology, which is required for individuation on this account. Similarly, when you attend to a dusty purplish shade of pink, there seems to be a phenomenological difference between attending to the *pinkishness* of it, and attending to the *particular shade of pink*.

4.2.1 Attention and Dualism

I now want to consider whether dualists can accept this approach to individuation. First, I want to point out that Pained Zombies also don't create a challenge for dualists' adopting this way of individuating phenomenal concepts. When I think about phenomenal experiences by quoting a painful experience ("the experience: *painful experience*") my Pained Zombie Twin quotes the same experience. In thinking this thought, I rely on a subconscious template, which has the effect of making the *phenomenalness* of the embedded painful experience jump out at me. But my Pained Zombie Twin has the very same template, playing the very same functional role. The phenomenalness of her embedded experience should jump out at her too. If this is what makes the concept a concept of phenomenal experience, then we should both be employing a concept of phenomenal experience. The fact that my Pained Zombie Twin has only had one sort of phenomenal experience will not be relevant. And, if she is indeed consciously aware of the *phenomenalness* of her experience in particular, this seems

[21] While I gave these instructions using color terms 'blue' and 'green', there's no reason to think this attentional process must depend on consciously intending to focus on 'blueness' or on possession of a term for blueness. This feature of my writing is simply the result of my need to provide explicit consciously understood instructions to convince you that phenomenology does change with attention.

intuitively the right result.

(Note that the attentional view and the dispositional view differ in whether they grant my Pained Zombie Twin the concept PHENOMENAL EXPERIENCE. This is because, on the dispositional account, my Pained Zombie Twin is not latching onto the *phenomenalness* of her painful experience—and is not disposed to apply her concept to painful experiences—whereas on the attentional account, she is consciously grasping the phenomenalness of her experience.)

But we can construct more puzzling cases than this. Consider my Disjoint Twin. She's physically identical to me, but she lives in a world with very different psychophysical laws. Suppose that for me R_1, R_2, R_3, R_4, and so on, are all brain-states that correspond to different shades of red (red_1, red_2, etc.). For any brain-state R_i, when I'm in R_i, my Disjoint Twin is in R_i. But my Disjoint Twin's phenomenology differs from mine in the following way: Every even numbered *i* corresponds to a red experience. But every odd numbered *i* corresponds to a green experience. When I am shown a color spectrum from pink to maroon, my experiences have uniformity. My Disjoint Twin, on the other hand, has experiences that shift sharply from pink to pastel green to pale red to pale green, and so on.

Keep in mind that my Disjoint Twin and I are *physically* identical. This difference in our phenomenology arises from a difference in the psychophysical laws of our respective worlds. Her world's psychophysical laws are erratic in a way that ours are not. But because we are physically identical, she will have a *subconscious template* that is physically identical to my red template.

Suppose that right now, I'm in brain-state R_2 and am entertaining a red phenomenal experience. I'm attending to this experience and thinking about the redness of my present phenomenology, using a concept "the experience: *red_2*", where the redness of the embedded experience is highlighted by my attention to it. My Disjoint Twin is also in brain-state R_2, having a red phenomenal experience, and thinking about "the experience: *red_2*" (where her subconscious template, like mine, alters the embedded phenomenology). But, because of the psychophysical laws, certain conscious experiences that her template would apply to—that she would be disposed to categorize as "red experiences"—are inverted.

What should we say about my Disjoint Twin's concept? Is she currently employing the same concept as me?

One reason you might think she has a different concept is that you might think that she's disposed to apply her concept to different experiences: all those green experiences that her subconscious template applies to. But this presupposes that the concept she's applying to the green phenomenal experiences is the same concept that she's applying to the red experience she's having now. But this fails to acknowledge that on the attentional approach, phenomenal concepts are individuated by reference to the highlighted qualities.

A better way to make sense of my Disjoint Twin's concepts parallels what we said in the case of the Pained Zombie. On this way of understanding my Disjoint Twin's concepts, she is currently employing the very same concept as me: a concept that applies to all and only red experiences, because it is formed by attention to the redness of her

current red experience. But when we're in brain-state R_3, and are having different experiences, we will also be employing different concepts. I will be employing my red concept (by virtue of having a red experience, and my template highlighting the redness of the experience). And—assuming activation of her subconscious template functions to highlight the greenness of her phenomenal experience—she will be employing a green concept (by virtue of having a green experience, and her template highlighting the greenness of the experience). This follows what we said in the Pained Zombie case: My Pained Zombie Twin was employing the same concept of me insofar as we were both relying on an embedded pain experience, where we were attending to the *phenomenalness* of the experience. But when I thought about experiences by employing other phenomenal experiences (which my Pained Zombie Twin lacked), we did not share our concepts.

If this is right, my Disjoint Twin's mental life differs from mine not only in that we (sometimes) have different experiences, but also in that we (sometimes) employ different concepts. But all this doesn't mean that we're not also sometimes using the same concepts! While this may look strange on the face of it, we're facing a very strange possibility. This is surely the most normal thing that could be said about this case, as it respects both the intuition that there's a strong connection between our phenomenal experiences and our cognitive lives, and also the intuition that individuals who are having the same experiences and processing them in the same way are cognitively alike.

4.3 Indeterminate Imagery Response

I want to consider one final way of responding to the individuation challenge that denies a key assumption of the challenge. To introduce the approach, consider the following apparently unattractive response to the challenge: *Phenomenal concepts refer to the most particular quality instantiated by the embedded phenomenal experience.* Why isn't this the obvious response to the challenge? There's an assumption that every phenomenal experience is determinate. For any red experience that could be embedded in a red concept, there's an assumption that it will be of a determinate shade. As a result, this response seems to entail that we can only have phenomenal concepts of e.g. red_{42}, red_{43}, $tickle_{789}$, etc. But this is absurd. We clearly have general phenomenal concepts, which apply e.g. to all and only red experiences, or to all and only tickles.

What I want to consider in this section is a way to defend this very simple proposal for individuating phenomenal concepts by denying that phenomenal experiences are always determinate in the way that Tye implicitly assumes. It's clear that perceptual experiences tend to be determinate. But I'll argue that mental imagery can exhibit fairly radical indeterminacy. The proposal will then be that (1) general phenomenal concepts require embedded mental imagery that is highly indeterminate, and that (2) they refer to the most particular quality the embedded imagery instantiates.

On this view, a general phenomenal concept red takes the form *the experience *phenomenal experience that's determinately red, but not determinately any particular shade of red**. Since this is quite a radical suggestion, let me start by making an intuitive case for the idea that mental imagery can be indeterminate in the way described. I'll then

put forward and respond to two reasons to question this intuitive defense of indeterminacy.

I'm trying to recall what color Bill's eyes are. I close my eyes and imagine Bill looking at me. He has eyes in my mental image. His eyes seem dark—yes, they're *definitely* dark. But when I try to focus on their color, it eludes me. They're just dark. There isn't any particular dark color they seem to be. It's not just that when I think back, I can't draw to mind what color Bill's eyes are. I've formed a mental image of Bill looking at me, and in this mental image that I'm currently attending to, there's no answer to be found as to the precise shade of the eyes. Similarly, while I may be able to imagine particular shades of red, it also seems to me that I can form a mental image of redness, such that when I compare it to particular red shades in the world, it doesn't seem to pick out any one over the others.

This provides at least a prima facie case for thinking that there can be a degree of indeterminacy in mental imagery. (Though this does not address the question of *how much* indeterminacy our mental imagery can admit—a question to which I'll return.) But there are two reasons one might reject this intuitive case.

First, one might have different intuitions based on introspection than those I've tried to motivate. When I first began thinking about mental imagery, I'd assumed that everyone's introspection of their own mental imagery would reveal the same things that my own introspection revealed. But after presenting the idea of indeterminate mental imagery to other philosophers, I discovered that people give radically different accounts the phenomenology of their own mental imagery. Some report that their imagery works in

the way I've described, allowing for indeterminacy in certain respects. Some report that visual imagery always seems to be determinate in all respects just like their perceptual imagery is. And some report that their mental imagery seems to bare no obvious relationship to perceptual experiences whatsoever. (This parallels the findings of Galton in the late 19th century, who discovered radical variation in people's reports of the distinctness and vividness of their mental imagery (Galton 1880).)

In the hopes of making this approach to phenomenal concept individuation compelling to those who don't share my apparent ability to consciously generate indeterminate mental imagery, I'll offer some reasons to doubt the appearance that phenomenal experiences supplied by imagination must be determinate with respect to their more particular qualities.

When we consider whether a red mental image could be indeterminate with respect to shade, we tend to close our eyes and try to form a mental image of a red of an indeterminate shade. But this is not the sort of process by which the mental images that supply the content to phenomenal concepts are generated. The mental images that ground our phenomenal concepts are generated without conscious effort. We don't sit down with the instructions "form a red mental image that's indeterminate in these respects" and then make a conscious effort to do so. It simply happens.

Why is this important? When we think about what's possible through imagination, we may be tempted to project inward from what's possible in perception. Several of the people I discussed this with made comments to the effect of "I could never see anything that's red but indeterminate with respect to precise shade, so how could I imagine

something that was red but not any particular shade?" If we believe, before even introspecting on our experiences, that it's not possible for experiences to be indeterminate in this way, this may affect the experiences that we're able to intentionally generate. If I can think of no way to have an indeterminately shaded color experience, when I try to direct the process of imagining one, I will be doomed to fail. I will try to imagine such a color using methods that occur to me as making sense: I will try to imagine the paradigm red, or try to imagine a red thing that I recently saw. It is not surprising that we fail to perceive indeterminately shaded red when we do this. It also tells us nothing about the *possibility* of having such experiences.

Further, there are processes we can consciously go through that demonstrate that it is possible to have color experiences with a degree of indeterminacy[22]: Focus your eyes on a point, and then (without shifting your gaze) consciously attend to the color of an object in your peripheral vision. If your color experiences are anything like my own, your experience of the color will be determinate in some respects (perhaps seeming dark red), but there will not be any precise shade that the object seems to be (it doesn't seem to be any particular dark red shade rather than another). There's a degree of indeterminacy in the color experience, which can't be explained away as a fault of memory. Once you've had a perceptual experience of a color that's indeterminate in certain respects, you can try to replicate this experience in imagination, yielding a color experience that's indeterminate regarding its precise color.

This experiment may not yield indeterminacy on such a radical scale as would

[22] Thanks to Mark Johnston and Sarah-Jane Leslie for pointing this phenomenon out to me.

have to be possible to account for all the general phenomenal concepts we possess. So we're left with a further question of *how much* indeterminacy is psychologically possible. But if we can have experiences that are indeterminate within a narrow range, this suggests that there is no barrier *in principle* to with expanding the range of possible indeterminacy to include e.g. red experiences that are indeterminate in all more particular respects, or even visual experiences that are indeterminate in all more particular respects. The further question of *how much* indeterminacy there can be in mental imagery is not so much a philosophical one as an empirical one.

Finally, there are cases where our phenomenal experiences appear to everyone to be indeterminate in certain respects, which might at first-glance seem impossible. When you imagine a striped tiger, it is determinately striped, but there may be no particular number such that the tiger seems to have that number of stripes. In order to deny the possibility of a red experience that seems indeterminate with respect to its shade, we'd need some principled way of distinguishing between these cases.

To summarize, I've claimed that introspection reveals that mental imagery admits indeterminacy of the sort required for general phenomenal concepts on this view, giving a prima facie case in favor of the possibility of individuating concepts in this way. I've put forward reasons for thinking that contrary intuitions do not override this prima facie case: (1) There's reason to think the mental imagery we can consciously direct ourselves in generating will be more limited than the mental imagery we can generate in general. (2) We can even generate indeterminacy in our perceptual experiences, showing indeterminacy in experience to be possible. (3) There are undeniable instances of

indeterminacy in mental imagery. To deny that the phenomenology I've proposed is possible, we would need a principled basis for distinguishing the two instances of indeterminacy.

But there are independent reasons one might deny the possibility that mental imagery admit indeterminacy in the way I've suggested. Any time we have a representation, we can distinguish between the content of the representation and the form of representation. Consider the sentence "There is a mountain range along the river." This sentence has certain *content*—it represents a certain sort of topography by the river. And the lines on the page are the *form* by which the content is represented. Similarly, when I draw a mountain range—evenly spaced triangles—I represent the topography as being a certain way. This is the picture's content. And the triangles are the form that represents this content.[23]

In both the case of the map and the sentence, we can easily make sense of indeterminacy of *content*. Language isn't rich, and so can leave content features open. For maps and other pictures, we have an understanding of what certain features represent, in a way that can leave information unspecified. (This might be conventional, explicitly stated as in a map's legend, or grounded in our psychology.) But in both these cases, the *forms* are never indeterminate. There is a precise fact about the relationships of the lines that make up the sentence "There's a mountain range along the river." There's a precise

[23] The distinction between content and form is similar to the more familiar distinction between the content and vehicle of a representation. The difference is that term *vehicle* seems to apply both to what I would consider the *form* of a representation, and to what I would consider the *realizer* of that form. We might say that the vehicle of representation for a drawing was patterns of graphite on dried wood pulp, or that the vehicle of representation for a human mental representation was a particular pattern of neural firings. Neither of these things is the form of the representation. The form is at a step higher of abstraction: The pattern of lines in the case of the drawing, or the functional state in the case of the mental representation.

fact about how many triangles there are on the map.

But in the case of phenomenal experiences, the form and the content cannot be neatly separated. When I think "this experience *phenomenal red experience*," I'm thinking something that *represents* phenomenal redness. But the phenomenal redness is also what fixes the reference. It's the redness of the experience that *does the representing,* analogously to the triangles in the text and map cases. The phenomenal experience does double-duty as both content and form. Hence to have indeterminacy in the content, we must also have indeterminacy in the form.

But this looks problematic. While we can imagine a map or a sentence that represents an indeterminate number of mountains, we can no more imagine that the form by which it does this is indeterminate than we could imagine that the mountain range itself was indeterminate. So we have a challenge: How can the *form* of mental imagery be indeterminate? Why should the form of my mental imagery be any different from the form of other representations? (Even if you don't like the distinction between form and content in the case of phenomenal experiences, we find this problem. It's clear that indeterminacy in mental imagery isn't generated in the same way as in the map or a body of text. We don't have something determinate that leaves open certain aspects of what it *stands for*. So we have a puzzle: How does indeterminacy in mental imagery arise?)

4.3.1 Indeterminacy in Mental Imagery

Mental images are, to use Colin McGinn's terminology, *attention-dependent:* You have to attend to a mental image in order for the image to exist. If you're sitting in a

lecture, daydreaming about your favorite forest walk, and the speaker makes a point that catches your attention, the image of the forest ceases to exist the moment you shift your attention from it. When the talk becomes dull and your mind again focuses on the forest, you can again conjure up a mental image of it. But during the time when you were attending to the lecture (and your limited attention capabilities precluded you from simultaneously attending to the forest), it seems clear that no forest image was present but not attended to. Your mind is the source of your mental imagery, and your attention is required for its sustenance.

This applies on a local scale as well as a global one. There can be no aspect of my mental image that exists, but that I am not attending to. But our attentional capabilities are quite limited. Even when I stand in the middle of a forest, I cannot attend to everything in my visual field simultaneously. When I close my eyes and imagine the forest, I could not replicate the detail of the world even if I tried. (And in forming mental images, I may not always try to replicate the detail of the external world.) Because our attention is selective, mental images lack what McGinn calls *saturation*. In the saturated perceptual image, "every point of the visual field is such that some quality is manifested there, whereas this is not true of the [mental] image. Percepts represent the world as dense, filled, continuous; but the [mental] image is gappy, coarse, discrete."[24] (McGinn 2004, 25)

These observations are hard to deny, and I think they can help explain just the sort of indeterminacy we're looking for: We don't merely find an indeterminacy of the

[24] While McGinn may be overly optimistic about the degree to which out percepts are saturated – peripheral vision is a prime counterexample – he is surely right about mental imagery lacking saturation and lacking it to a far greater degree than our perceptual imagery.

content of what is represented, in virtue of a determinate form being taken to stand for many different particulars. We find an indeterminacy of *form*. The lines and color of the mental images themselves leave gaps because *we* don't put enough in to fill them.

Consider the following examples. I close my eyes. Visually, nothing is being represented to me. If I try to comment on my current visual experience, I might say that I see blackness. But this blackness is clearly not a mental image. Now I imagine a red triangle. My mind is responsible for supplying the content of the image. It puts in the shape and color of the triangle, but that's all that it supplies. Intuitively, the *content* of my mental image is a red triangle (not a red triangle suspended in blackness, or a red triangle with no surroundings). But furthermore, the *form* of the mental image seems to be given by the lines and coloring of the triangle (not by the triangle, it's coloring, and the blackness). The form of the mental image is what lines and color and so on *my mind* contributes to the mental image. The form does not include the absence of everything that I fail to contribute. Because of this, the mental image "leaves open" or indeterminate the triangle's surroundings.[25]

[25] This seems intuitively the right thing to say about mental imagery, but it relies on a very different way of understanding mental imagery and perceptual imagery, about which more needs to be said. Suppose I get information through vision, telling me that there's a red triangle, but not saying anything about its surroundings. Contrary to the case of mental imagery, the perceptual experience that I have as a result would represent *a red triangle suspended in blackness*, and the form of this experience would include blackness as a component. What accounts for this difference?

I'm inclined to think that our perceptual experiences present themselves as complete in a way that our mental imagery does not. We process perceptual experiences as informing us "here's what's in your field of experience, *and that's all*." Because of this, when P is not included in a perceptual experience, we take reality to determinately lack P. Our mental imagery doesn't seem to be processed in this way. It tells us "here's what's in the situation," but makes no claim as to the depiction's completeness. The source of our mental imagery is in us, and so is not complete. Furthermore, *the way we process the mental imagery reflects this fact*. It doesn't even *seem to us* that the law of excluded middle should hold of our mental imagery.

Because of this, something more is required for a mental image to represent an absence, beyond simply failing to include the thing. We must depict not only what *is* a part of the image, we must also depict what is determinately *not* a part of the image. So when a color in my perceptual experience includes redness, but leaves out all other hues, this color is perceived as a particular shade:

Now that we've seen that indeterminacy of form as well as representational content is possible in the case of mental imagery, we can see how to apply this idea to the case of color indeterminacy. Suppose that particular color-shades aren't fundamental, rather are built up out of hue, saturation, and brightness building-blocks. *Hue* includes components like redness, greenness, blueness, and yellowness. *Saturation* describes how rich or intense in hue a color is. *Lightness* describes how light or dim the shade is.

We might tell something like the following story about how particular color shades are built up out of these building-blocks. (Note: I am not committed to this being the correct story about how colors are composed. I simply provide it as a toy example for purposes of making sense of how there might be indeterminate color imagery. The real point that I want to make is that *if* color experiences are built up out of more fundamental components, then we can use this fact to account for indeterminacy in mental imagery. I provide the following example of how specific color shades might be build up to illustrate the proposal.)

When I have a perceptual experience of a pomegranate, the external world supplies my mind with a complete profile of the color of the pomegranate: The exact proportions and ratios of red, green, blue, and yellow, the degree of blackishness. The ratios of the hue components gives us the precise hue of the pomegranate. The absolute portions of these components gives us the saturation. (Two shades might each have a ratio of 90% red, 10 % blue, 0% green and yellow. But one might achieve this by including 18 units of redness and 2 units of blueness. The other by including 45 units of

"supersaturated red." While, when a color in my imagination includes redness, but leaves out all other hues, this is perceived as a red of an indeterminate shade. (This simple sketch cannot be quite right, since we need to account for the perceptual experience of color indeterminacy in our peripheral vision. But I suspect something along these lines needs to be at work.)

redness and 5 units of blueness. The latter, though the same in hue, is more intense, more saturated.) Finally, the degree of blackishness determines how light or dim the shade is.

How can this accommodate indeterminacy in color imagery? At the beginning of this section, I claimed that I could form a mental image of an eye, where the eye was determinately dark, but not determinately any precise dark color. We can now make sense of this in the same manner as the triangle's leaving open its surroundings. My mind might put into the image the blackness of the eye, without including other color shade components. As a result, the eye would be determinately dark without determinately being any particular dark shade.

In a similar way, we might make sense of mental imagery that's determinately red, but not determinately any particular shade of red. My mind puts redness into the image, but fails to add any amounts of blue, green, yellow, whiteness, or blackness.

Indeterminacy in saturation is slightly more tricky, but I think can be achieved in a similar way. Recall that we can form mental images of tigers that are determinately striped, but don't have any determinate number of stripes. This suggests that mental imagery can leave magnitudes open. We can then make sense of imagery that's determinately purplish-red, but indeterminate with respect of its intensity. My mind might supply that it is 75% red and 25% blue, but fail to specify any particular magnitude of either hue.

4.3.2 Indeterminacy Individuating Phenomenal Concepts

So far, I've argued that we can make sense of the notion of a phenomenal

experience that is indeterminate in various respects. I recognize that this is a very strange proposal, and that even if the arguments I have given in defense of it are accepted, your phenomenal introspection may not reveal the sorts of phenomenal indeterminacy I have postulated. I hope that I have at least succeeded in arguing that we can *make sense* of this proposal. How does this help us to individuate phenomenal concepts?

The idea would be that (e.g.) a general phenomenal concept of redness is partially constituted by mental imagery that's determinately red, but not determinately any particular shade of red. Phenomenal concepts refer to the most particular qualities exhibited by their fillers. So a more specific concept would involve an experience that was determinate with respect to qualities this experience leaves indeterminate; a more general concept would involve an experience that was indeterminate in more respects. The concept's reference is fixed not only by the qualities that the experience has, but also by the qualities it lacks.

It should be obvious that this method of individuation is as open to the dualist as it is to the physicalist: Phenomenal concepts are individuated not by physical states or processes, but by the phenomenology itself. Thus there is no potential, on the dualist's understanding, for the concepts to come apart from their referents.

4.3.3 Psychological Restrictions on Indeterminacy

This method of individuation faces a difficult challenge: Do *all* phenomenal concepts have phenomenal experiences that partially constitute them, and that are indeterminate with respect to all qualities more particular than those that the concepts

pick out? Could we have a phenomenal experience that was left open what *sort* of experience it was (e.g. whether it was a color experience, kinesthetic experience, sweet experience)? Could we have a colored experience that was indeterminate with respect to its color? What about blue-or-red experience that left the redness and blueness indeterminate? Or an intense experience that was indeterminate in all other respects?

The answer to each of these questions *seems* to be 'no'. But what explains why it's possible for experiences supplied by the imagination to be indeterminate in certain respects, but not in others? What principled basis could there be for drawing such a distinction?

There seem to me to be (at least) two broad classes of phenomenal qualities that we cannot make sense of experiencing in absence of more particular qualities. The first class contains gerrymandered experiences: red-or-blue experiences, dark-or-yellow experiences, and experiences (such as intense-experiences) that look like they form unified types, but are really just as gerrymandered as red-or-blue experiences. The second class contains general concepts that express different modalities of experience like phenomenal-experience, color-experience, kinesthetic-experience, and temperature-experience.

Red-or-blue experiences seem to be ruled out by our psychology. This seems difficult to deny. When I have a speckled experience, it may not seem determinate whether it was an experience of 42 speckles or 52 speckles. But it might seem determinately not the case that there it was an experience of 1,319 speckles. It seems that every quality that it's indeterminate whether an experience has must possess a certain

degree of similarity, where the degree threshold is determined by our psychology. Likewise, I think we can make good sense of having a mental image of a color, where it's not determinate whether the experience was of brick-red or of scarlet, but where it is determinately not a blue experience. But my brain doesn't seem to recognize red and blue experiences as sufficiently similar to imagine a color that was red-or-blue, but not determinately either. This seems to be a psychological fact.

Why can't we have an intense-experience that isn't any particular type of experience (e.g. a red, pain, or loud experience)? A pain, a red, a tickle can all be intense. INTENSE doesn't seem to pick out a unified collection of experiences, rather it seems to modify concepts, picking out a different quality in each of these cases. There is no single quality they share, which could exist in isolation from other qualities. Similarly, a bright experience doesn't seem to be any single thing.[26] BRIGHT seems to modify different color concepts, picking out a different quality in each case. While these concepts may appear to pick out individual qualities, they are really analogous to the red-or-blue experience. It again seems plausible that there is something about our psychology that rules these experiences out. (At first glance, it might seem strange that I have claimed that dark experiences can be indeterminate with respect to color, but that bright experiences cannot be indeterminate in this way. This can be explained by the fact that there is something in common that all darkly colored experiences share—a blackishness—but that there is no analogous common thread through the brightly colored experiences. In the one case, the relevant color dimension is the binary of light to dim; in the other, it

[26] I use 'bright' here as in the common usage to mean something like *vivid*. A bright yellow is a bold and vivid yellow. A bright purple is a bold and vivid purple. Bright in this sense might be something like 'highly saturated'.

is that of saturation.)

What are we to say about concepts like PHENOMENAL EXPERIENCE, COLOR EXPERIENCE, KINESTHETIC EXPERIENCE, and TEMPERATURE EXPERIENCE? How do the experiences that constitute them differ from experiences that intuitively can leave more particular qualities open, such as red experiences, cold experiences, high pitched, and dark experiences?

One might try to give a response to this that is similar to the response I gave to the possibility of a red-or-blue experience. Perhaps, due to my psychology, red, blue, green, yellow, etc. aren't similar enough for an experience to be registered as determinately *one of them*, but indeterminate regarding which one it was. Likewise, it might be a brute fact of my psychology that cold-experiences and hot-experiences aren't similar enough to allow for a temperature-experience that was indeterminate regarding whether it was a cold-experience or a hot-experience. This is of course extremely speculative, and further might be found unsatisfying, as it doesn't give any *reason* for thinking that there is an important distinction to be made between the two clusters of qualities.

Here's one reason to think that there's an important distinction between these two clusters of qualities. Some phenomenal concepts are vague – RED, COLD, LOUD – admitting borderline cases. But not all phenomenal concepts are vague. PHENOMENAL EXPERIENCE, VISUAL EXPERIENCE, and TEMPERATURE EXPERIENCE all have sharp boundaries: for any potential referent, the concepts either determinately apply or

determinately don't apply.[27] For any mental image we can form that admits indeterminacy, we find that the corresponding concept is vague, admitting borderline cases. Conversely, all of our vague phenomenal concepts (RED, COLD, LOUD and so on) are such that we can form mental imagery that determinately exhibits the relevant quality, but is indeterminate in other respects. The distinction between cases where we can and can't form indeterminate mental imagery parallels the already familiar distinction between vague and precise concepts. This reinforces my claim that there's a genuine difference between those phenomenal qualities that we can and cannot form indeterminate mental imagery of.[28]

4.4 A Unified Account?

If we want to give a unified account of all phenomenal concepts – RED and LOUD, as well as COLORED and AUDITORY – then it's obvious that this mental imagery suggestion won't cut it. Even if it's only a psychological fact that creatures like us can't form mental imagery that's determinately colored, but not determinately any particular color, we'd like to have an account of how phenomenal concepts work *for*

[27] This might seem wrong: Surely it can be indeterminate when you stop having a sensation of warmth, and hence (you might think) when you stop having a temperature experience. But while it can clearly be indeterminate when an experience of warmth becomes an experience of lukewarmth or of coolness, this doesn't show what it can be indeterminate whether you're having a temperature experience at all. A warm temperature experience can fade into what we would perceive as a "neutral" temperature experience, but this is not the same thing as blurring into an absence of temperature experience.

[28] This also suggests some curious things about vagueness: I've suggested that it's a matter of our psychology how similar two experiences must be for our minds to form a mental image that determinately captures one or the other, but is indeterminate beyond that. If this is right and there's a tight connection between indeterminacy in imagery and vagueness of concepts, then creatures psychologically different from us may find VISUAL EXPERIENCE and even PHENOMENAL EXPERIENCE to be vague.

creatures like us. And this proposal cannot be used to individuate our concepts of phenomenal modalities.

But I'm not sure that this is a point against this method of individuating. While parsimony is ceteris paribus a virtue, there's reason to think that we *shouldn't* want a unified theory that treats concepts like RED and PHENOMENAL in the same way. When I motivated the constitutional theory, I noted that it was compelling to think that what changes cognitively in Mary upon seeing red for the first time is that she becomes able to directly latch onto her new experience and pull it into her thoughts. One of the virtues of the constitutional theory was purported to be its ability to account for the substantialness of phenomenal knowledge. Defenders of the theory describe phenomenal concepts as giving us "unmediated insight into the essences" of our phenomenal experiences, making the redness of red, the tickliness of tickles, and so on consciously present to us. I argued that we, likewise, plausibly have direct insight into not only the particular experiences we 'quote', but also the particular *qualities* of these experiences that we conceptualize. While all these things look very compelling in the case of redness and tickles, it's not obvious that these are things we want out of a theory in the case of phenomenalness or coloredness.

Neurosurgeons give a blind man the ability to see for the first time. Far from intuitively desirable, it seems downright counterintuitive for a theory to hold that when the blind man first has a visual experience, he consciously latches on to the *visualness* of his experience, in particular. I certainly don't seem to be able to directly latch onto the *visualness* of my experiences and thereby pull it into my thoughts. And it's not obvious

that I have substantial knowledge of what it is to have a visual experience over and above my knowledge of the particular sorts of experiences that seems to count as a visual experience.

So perhaps we should expect to give a different treatment to concepts of phenomenal modalities, and concepts of more particular phenomenal qualities. One could insist that the constitutional theory offers a good account only of certain phenomenal concepts (those that present phenomenal qualities to us in a way that gives us substantial grasp of their nature), while perhaps a direct reference account is better served to capture concepts of phenomenal modalities. But a more unified way to go, that would preserve the constitutional theory as a good theory for all sorts of phenomenal concepts, would simply rely on different ways of individuating concepts of phenomenal modalities from the way we individuate more particular phenomenal concepts. The point I raised against the dispositional method of individuating, that it seems to run contrary to the motivation for the constitutional theory, seems to lose its force for concepts like PHENOMENAL EXPERIENCE. So we should not be afraid of appealing to it in these cases.

Perhaps then the dispositional method of individuating is how we should individuate concepts of phenomenal modalities, but the attentional or indeterminate imagery approaches do the work of individuating more fine-grained phenomenal concepts. The need to individuate in different ways will be the result of certain psychological factors: We seem to be able to consciously grasp the redness of red experiences, in a way that gives us substantial knowledge. Because of this, we want a way of individuating phenomenal red phenomenal concepts that reflects this. But – due to

the way we're psychologically constructed – we are incapable of consciously latching onto the *phenomenalness* of our conscious experiences. Hence, we don't want a way of individuating these phenomenal concepts that either presupposes or has the result that we are capable of this. So, at least for creatures psychologically like us – able to consciously latch onto the redness of our experiences, but not their phenomenalness – we need different explanations of how to individuate these two sorts of phenomenal concepts. A different sort of creature might require only one sort of individuation mechanism, but we require two.

Even the attentional theory, which looked highly attractive in relation to the dispositional theory, doesn't seem to be capable of doing all of the required work. While it seems at least somewhat plausible that the *greenness* of a particular blue-green shade could pop out at me, it seems obvious that such general qualities as the *phenomenalness* or the *visualness* cannot be highlighted in my phenomenology in the same way. So as with the mental imagery view, we find the need to appeal to another way of explaining concepts of phenomenal modalities. None of the three approaches we've considered can – or should be expected to – pull all the work of phenomenal concept individuation on its own.

4.5 Summary

Let's briefly consider where the three methods of individuation I've discussed stand in light of this discussion.

The dispositional method of individuation looks unappealing as a way of individuating the sorts of phenomenal concepts that have pervaded the phenomenal concepts literature, concepts like: RED, BLUE, GIDDY, COLD, LOUD. On this view, while I may be consciously aware of the *particular* pink experience that's partially constituting my pink concept, I'm not *aware of the pinkness*. What makes my concept a concept of pinkness is simply subconsciously driven dispositions. But while this looked like a bad result for concepts like RED, BLUE, GIDDY, COLD, and LOUD, it is a welcomed result for concepts of phenomenal modalities, such as: COLORED EXPERIENCE, VISUAL EXPERIENCE, TEMPERATURE EXPERIENCE, and PHENOMENAL EXPERIENCE.

By contrast, the attentional and indeterminate mental imagery ways of individuating seem to deal well with the concepts that are often discussed in the literature, but a poor job of dealing with concepts of phenomenal modalities. It's hard to see how a unified treatment of individuation is possible, and because of the psychological differences in how we grasp these different sorts of phenomenal qualities, this doesn't look like a bad thing.

How do the attentional and the indeterminate mental imagery views fare in comparison to one another? The strongest objection to each seems to lie on introspective grounds: One might question whether phenomenal qualities can really "pop out" in the way I suggested on the attentional view. Or one might reject the idea that mental imagery can be indeterminate in the ways I've suggested, despite my attempts to rebut these contrary intuitions.

The attentional and mental imagery views also have different strengths and weaknesses. The attentional method of individuating seems to do well in cases where the phenomenal concept is filled by a perceptual experience. I look at my tea cup, and attend to the experience I'm having, thinking: *this experience *tea cup color, where the blueness pops out at me**. But it doesn't seem so attractive in the case of phenomenal concepts that employ mental imagery. Suppose there's no blue object in my environment. On the attentional view, to form a blue concept, I'd need to generate blue mental imagery. This mental imagery would presumably be a *particular shade* of blue, as the advantage of the attentional theory was that it avoided positing (possibly suspect) indeterminate imagery. And then the *blueness* of this particular experience would need to be highlighted in the mental image. This seem to presuppose some pretty impressive attentional powers: I find that when I try to form a mental image of a particular shade of blue, and then mentally "narrow in" on the *blueness*, I can't simultaneously attend to the blueness, and continue devoting the limited attentional resources to the particular shade to maintain it. The shade simply collapses into a general blue, the moment my attention begins to dwell on the *blueness*. So while the attentional theory offers a very attractive looking account of particular phenomenal concepts when they involve perceptual experiences, it doesn't look so compelling for those concepts based on mental imagery.

Conversely, while the indeterminate mental imagery view seems to offer a plausible account of concept individuation for those phenomenal concepts involving mental imagery, it has nothing at all to say about the concept I generate by "quoting" the color of my teacup. For the indeterminate mental imagery view to offer a general account

of individuation, even for relatively particular phenomenal concepts like red, it would need to posit that *every* phenomenal concept involves embedded mental imagery. So even as Mary sees a red apple for the first time, when she thinks that the apple is red, she is using mental imagery as well as the perceptual imagery that she's having in virtue of seeing the apple. This might not be a terrible bullet to bite. Perhaps perceptual experiences can only give us concepts of determinate phenomenal qualities (RED_{42}, $BLUE_{591}$, $GIDDY_{75}$, $COLD_{91}$, and $LOUD_{248}$), and to have a more general phenomenal concept, which involves substantial insight into the nature of the experience, we have to have an indeterminate experience. But if you are not satisfied with this result, it seems we must admit even *more* possible ways for individuation to work.

One might take the attentional and mental imagery views to work together, much as they might work with the dispositional theory. I don't think this is as ad hoc as it appears. Indeterminacy in mental imagery might simply be a special case of attention to a particular quality. When we attend to a quality in a perceptual experience, the quality we attend to pops out at us, but – because our mind is not the source of all the other qualities of the experience – the experience has other more particular qualities besides the one we're attending to. By contrast, when we attend to only a particular quality in mental imagery, since we are the source of each aspect of our mental imagery, and our attentional capacities are limited, the quality we attend to winds up being the most particular quality of the experience.

4.6 Echoes of Locke

In many ways, this discussion parallels the debate between Locke and Berkeley about *abstract general ideas*. In particular, my discussion of the attentional and indeterminate mental imagery views might be seen as offering a limited defense of Locke's theory of abstract general ideas.

Locke wanted to understand how it is that our words refer to things in the world. He held that our words refer by signifying *ideas* in the minds of their users. In our terminology, we might roughly think of *ideas* as *concepts*. This leads to the question: How is it that ideas are able to refer to things in the world? This is especially puzzling when we think about ideas that apply to many different particulars – what Locke calls *abstract general ideas*. ANIMAL refers to all animals, EXTENDED refers to everything with extension, RED refers to all red things, and so on. How is it that general ideas are able to refer to many different particulars that all bear some commonalities, but also have important differences?

There is some dispute about Locke's understanding of the nature of ideas. Berkeley, a famous critic of Locke, took ideas to be mental images that refer to things in the world in virtue of bearing sufficient resemblance to them (Pitcher 67, 70-71). He took Locke to be committed to a similar understanding of ideas. I think that if we join Berkeley in this imagistic interpretation of Locke, we find some interesting direct parallels between (1) Berkeley's dispute with Locke, and (2) my discussion of how to individuate phenomenal concepts on the constitutional theory. My discussion may even be seen as offering a limited defense of an imagistic reading of Locke.

So let's join Berkeley in understanding *ideas* to be mental images that refer by virtue of bearing sufficient resemblance to things in the world. We can already see a similarity between abstract general ideas and the constitutional theory of phenomenal concepts: Both involve concepts for which mental imagery plays a reference-fixing role.

How do words refer to things in the world? Words signify ideas in the minds of their users. These ideas – we are joining Berkeley in supposing – refer to things in the world by bearing sufficient resemblance to them. In the case of ideas that refer to many different particulars, Locke argues that we frame these general ideas through a process of abstraction: Having encountered many particulars that are alike in some respects, while differing in others, we come to form an idea that *includes what is shared by all the particulars and excludes what is particular only to some*. I encounter people of all colors, shapes, and sizes. Through a process of abstraction, I am then able to form an idea that leaves out the differences of color, shape, and size, including only what is common to all people. Locke writes that in doing so, we "make nothing new; but only leave out of the complex idea they had of Peter and James, Mary and Jane, that which is peculiar to each, and retain only what is common to them all" (Locke 1690, III, 7). To form an idea of motion, I omit what is particular to each instance of motion: the particular body, speed of the motion, direction, acceleration, an so on. I form an idea that includes only what is left. To form an idea of red, I omit what is particular to each instance of red: the particular size and shape of the redness, the particular shade of red. I form an idea that includes only what is common to all instances of red. "[I]deas become general by separating from them the circumstances of time and place, and any other ideas that may determine them to this

or that particular existence." (Locke 1690, III, 6)

C.C.W. Taylor argues that there are two ways that we could understand the process of Lockean abstraction. He calls these the *paradigm instance* and the *schematic representation* views of abstraction. On the *paradigm instance* view, to form the general idea HUMAN BEING, I consider a particular human being (say Peter). And I consider only those sensory features of Peter that are relevant to being a human being. Thus one and the same idea (a representation of Peter) can at some times function as the idea PETER, and at others function as the idea HUMAN BEING, depending on what features of the idea we're considering (Taylor, 99). What I have called the "attentional view" of phenomenal concept individuation is akin to this. On the attentional view, a single type of phenomenal experience (say a red_{42} experience) can serve either as a RED_{42} concept or as a RED concept, depending on what features of the experience are being attended to. This can be thought of as a fleshing out of what it takes to "consider a feature of" an idea.

On the *schematic representation* view, abstraction involves the creation of a representation that literally leaves out those features that are not common to all. Rather than forming an idea of a particular man and only "considering" those features of him that are shared by all men, the schematic view has it that we actually form a representation that leaves out all of these non-shared features. Keeping in mind that we are following Berkeley in understanding ideas as mental images, my indeterminate mental imagery view is clearly a version of schematic representation. I've argued that the concept of redness is constituted by a mental image, where this mental image *leaves out* what is peculiar to each different shade of red, and retains only what is common to them

all.

Talyor takes Locke's writing to be indeterminate between these two understandings of abstraction. I take the schematic view to be a more natural reading of Locke: he describes abstraction as requiring "separating from" ideas all those features that are particular to only some referents, and describes abstraction as involving the "framing" of an idea that leaves certain features out (Locke 1690, III, 6-7). But other passages do suggest the paradigm view, such as his statement that abstraction makes "nothing new" (Locke 1690, III, 7). In any case, Berkeley clearly takes Locke to hold the schematic view. And it's this view that he attacks in the introduction of *A Treatise Concerning the Principles of Human Knowledge.*

Berkeley takes this Lockean view of abstraction to be incoherent and straight-forwardly absurd. To form an idea of a person, I must form an idea that omits what is particular to only some people, retaining everything that is shared by all. But this means forming an idea that is not any particular color (since there is no one color that all people share), and yet is colored (as all people are some color or other). This means that one must form an idea that is not any particular shape (as there is no single shape that all humans share), and yet has a shape (as all humans have some shape or other). It must have a sex, but no particular sex. It must neither have nor lack a beard. And so on. (Berkeley 1710, Intro 9) Berkeley writes "I cannot by any effort of thought conceive the abstract idea above described [of a person]. And it is equally impossible for me to form the abstract idea of motion distinct from the body moving, and which is neither swift nor slow, curvilinear nor rectangular." (Berkeley 1710, Intro 10)

Given his commitment to an imagistic understanding of ideas, it is not difficult to see why he thinks abstraction to be absurd. But my defense of indeterminate mental imagery offers a limited rebuttal to Berkeley. I have argued that – at least in the case of simple phenomenology such as the experience of red that's not any particular shade – such abstraction is possible and something we find when we introspect. This shows that, even on an imagistic, schematic reading of Locke, abstraction is possible – at least to some limited extent.[29]

4.7 Pointer Concepts and Tye's Direct Reference Theory

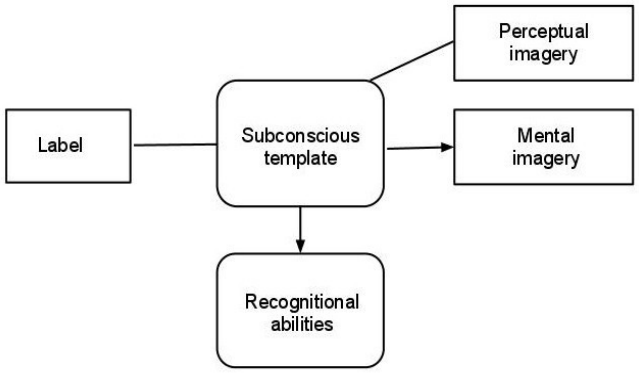
I want to conclude by considering the relationship between the fully fleshed out constitutional theory and Tye's "direct reference" theory from chapter 2. As we'll see, Tye's direct reference view of phenomenal concepts is closer to the constitutional view than it might appear at first glance. Recall that Tye held a theory of phenomenal concepts on which concept C is a concept of phenomenal redness, just in case C is reliably triggered by (and because of) phenomenal redness – essentially acting as a label for phenomenal redness. But this isn't sufficient to characterize what it takes for C to be a *phenomenal* concept. To be a phenomenal concept, C must be laid down in memory as the result of a phenomenal experience and must tend to trigger mental images of phenomenal redness in response to certain mental activities.

This view seems quite far apart from the constitutional view: phenomenal

[29] In other work, 'On the Possibility of Abstract Ideas', I offer a more full-fledged defense of Berkeley's imagistic reading of Locke. I argue that it's metaphysically possible that some creatures could form mental images of triangles with no determinate relationships between the sides, images of motion with no determinate speed, and so on.

concepts simply function as labels, and the phenomenology doesn't play a reference-fixing role. But given a fully fleshed out version of the constitutional theory and of Tye's view, I think we'll find the two views to have very little distance between them. For simplicity, I'll assume that physicalism is true.

On Tye's account, phenomenal concepts are like labels that are individuated by the subject's dispositions to apply them. These dispositions don't just arise at random, or by some divine miracle. Rather, the dispositions are doubtless guided by information stored in the agent's head. We might – this will sound familiar – call this guiding information a *subconscious template*. This template will be responsible for the subject's recognitional abilities, and will individuate their phenomenal concepts. But this only gives us a picture of what it is for a concept C to be a concept *of phenomenal redness*. It doesn't tell us what it is for C to be a *phenomenal concept* of phenomenal redness. To capture this part of the story, C must tend to trigger mental imagery of phenomenal redness. This mental imagery also helps to explain our feeling that phenomenal knowledge is rich and cognitively significant. This mental imagery is doubtless generated by the subconscious template, as that is what stores the relevant information. (It's hard to see how the label could generate imagery on its own.) Finally, we have perceptual imagery as well, that is connected to the subconscious template insofar as the subconscious template facilitates our recognition of the perceptual imagery. Here is an illustration of the view we've described:



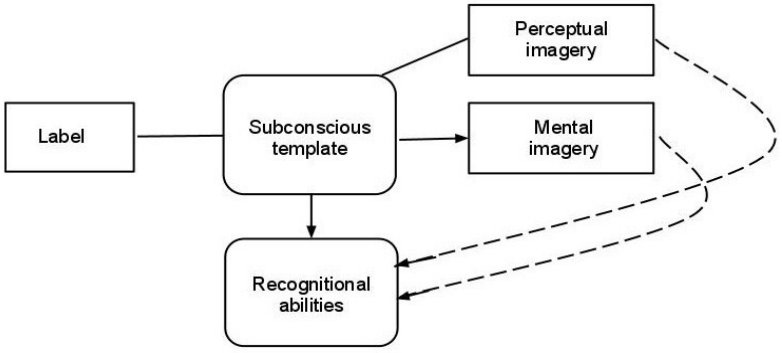
The arrows from subconscious template to imagery and to recognitional abilities indicate that the subconscious template is responsible for producing these things. The line between label and subconscious template indicates an essential relationship between the two, but not a direction of production. All of these components are essential for a concept C to be e.g. a phenomenal concept of phenomenal redness.

Now consider the constitutional theory, as we have fleshed it out. On this view, phenomenal concepts essentially contain imagery – either mental imagery or perceptual imagery – which gives them their cognitive significance. The mental imagery is presumably something that we are able to generate because our brains store the appropriate information about e.g. red experiences, pain experiences – because out brains store subconscious templates.

We have seen that there's a derivative sort of phenomenal concept (pointer phenomenal concept), which labels certain experiences without employing mental imagery. This sort of concept gets its meaning derivatively, by its connection to the "full", imagery-employing phenomenal concepts. It can't be that my full, imagery-employing

concepts generate these pointer concepts, since I can employ the pointer concepts even in the absence of the relevant imagery. Presumably the pointer concepts are connected to the full concepts by virtue of their both being connected to the same subconscious templates.

Finally, my possession of phenomenal concepts goes hand in hand with my ability to recognize phenomenal experiences. I am disposed to apply my phenomenal concepts (both pointer and full) to the appropriate types of experiences: red concepts to red experiences, blue concepts to blue experiences, and so on. (This is true whether or not we assign such recognitional abilities a role in concept individuation. It's something that any plausible theory of phenomenal concepts must accept.) Again, these dispositions are not generated at random. In the case of our pointer phenomenal concepts, the natural explanation is that our subconscious templates store information about experiences, which guides us in applying these labels. In the case of full phenomenal concepts, the imagery that partially constitutes the concepts will play a role in our recognitional powers. But if physicalism is true, the subconscious template and the imagery will not be able to come apart. The subconscious template stores information about what's going on in the subject's brain when she has e.g. a red experience. Since physicalism does not allow for the possibility of inverted spectra, anytime the subject has an experience that matches the red subconscious template, the corresponding imagery will be red imagery. So when the imagery plays a role in our recognitional powers, this will not have the potential to come apart from what would be the case if our subconscious template played this role. This give us the following picture.



Notice that this picture looks just the same as the illustration of Tye's view, with the exception of the added arrows from mental and perceptual imagery to recognitional abilities. These arrows indicate that for *full* phenomenal concepts, perceptual and mental imagery plays a role in our recognitional capacities. This is not the case for pointer phenomenal concepts, as these concepts do not need to be expanded into full phenomenal concepts in order to facilitate our recognitional capacities.

So Tye's model of phenomenal concepts, when fully fleshed out, looks just like the model of *pointer* phenomenal concepts on the constitutional theory. The constitutional theory differs in its account of full phenomenal concepts, for which imagery can play a role in our recognitional abilities. The constitutional theory is more attractive insofar as it's independently plausible that (1) mental and perceptual imagery can – in some cases – facilitate recognition, or that (2) our thoughts about experiences can make use of experiences in a way that renders them more significant than a mere label or empty pointer. Further, the constitutional theory, unlike Tye's direct reference theory, can give an account of the explanatory gap. Tye might try to borrow the approach taken by the constitutional theory, saying something like: Because phenomenal concepts tend to

trigger appropriate phenomenology, while physical concepts don't, we're fooled into believing that physical concepts don't *refer* to phenomenology. But I don't think this is a satisfying move given that on Tye's view, the phenomenology doesn't play a reference-fixing role. Just because employment of one concept *tends to trigger* something, while employment of another concept doesn't tend to trigger that thing, why think that this is reason to believe the concepts can't corefer? By contrast, on the constitutional theory, the very thing that fixes the reference for our phenomenal concepts is apparently missing from our physical concepts.

Chapter 5

Phenomenal Concepts and the Mind-Body Problem

In the first half of this dissertation, we assessed what we want out of a theory of phenomenal concepts and how to account for these requirements. I argued that neither the indexical nor the simple direct reference theory can give us what we want from a theory of phenomenal concepts. I then offered a defense of the constitutional theory of phenomenal concepts devised by David Papineau and Katalin Balog. The primary challenges to such a view are that it seems to make thoughts about phenomenology dependent on instances of the relevant phenomenology, and the problem of which of the properties exhibited by the phenomenal experiences that constitute these properties are relevant for their individuation. The past two chapters have sought to answer these challenges.

Having assessed what we want out of a theory of phenomenal concepts and how to account for these desiderata, we can turn to the philosophical work that these concepts can do in answering the mind-body problem.

We saw that there's a strong intuition that phenomenology is not reducible to the physical. A satisfactory account of the relationship between phenomenology and physical reality cannot simply ignore this intuition: It must either conform to the intuition (as the dualist does) or explain it away. Physicalists have recently sought to explain the intuition away by telling a physicalistically respectable story according to which phenomenal-physical truths, while necessary, are only knowable a posteriori. Even an ideally rational agent with complete physical information would find it a priori conceivable that there be

a zombie world; even an ideally rational agent with total general physical information would not be able to work out a priori that red looks like *this.*

In chapter 1, we touched on the most popular way of defending this view: the co-called *phenomenal concept strategy* (PCS). According to PCS, phenomenal concepts are so radically different from physical concepts that even an ideally rational agent with complete physical information could not work out the phenomenal-physical entailments a priori. Suppose that a physicalistically acceptable account of phenomenal concepts can be given that vindicates this move. While this would not *close* the epistemic and explanatory gaps, it does the next best thing: It gives a physicalistically respectable story of why the gaps arise. This argument schema is filled in with a particular theory of phenomenal concepts, such as the indexical, direct reference, or constitutional theories of chapters 2-4. The ideas is that if phenomenal concepts refer like indexicals (or directly or in some other completely different way), it might be impossible for an agent to realize from the inside that they referred to the same things as physical concepts. Thus phenomenal-physical truths could only be known a posterori, accounting for the epistemic gap and the sense that phenomenology is distinct from physical reality.

Another way of using phenomenal concepts to give a physicalistic account of the epistemic gap is the *conditional analysis of phenomenal concepts*, proposed by David Braddon-Mitchell (2003) and John Hawthorne (2002). Proponents of the conditional analysis hold that for all we can know a priori, the world could be merely physical or it could have nonphysical states. (It would, after all, be quite presumptive to think that we could know from our arm chair that the nature of reality is entirely physical!) They argue

that phenomenal concepts have a conditional structure, such that what they refer to depends on the actual nature of the world: If the actual world contains nonphysical states, then our phenomenal concepts refer rigidly to these states. But if the actual world is merely physical, then our phenomenal concepts refer rigidly to whatever physical states play the right functional role in the actual world.

If the conditional analysis is right, then we cannot rule out the possibility of zombies a priori: It's a priori conceivable that we live in a world with nonphysical stuff. But if such a world were actual, then zombies would be possible (since phenomenal concepts would refer rigidly to nonphysical states). On the other hand, should the actual world turn out to be purely physical (as the physicalist maintains), then zombies are *not possible* (since phenomenal concepts refer rigidly to physical states). The same will hold for each particular phenomenal experience's relationship to the brain. For example, it will be conceivable that being in pain is not just a matter of being in brain-state 28470823.

In the following two chapters, I argue that neither of these appeals to phenomenal concepts is successful: The structure of phenomenal concepts cannot give the physicalist an ideal epistemic gap. In chapter 6, I argue against the conditional analysis of phenomenal concepts. I argue that as defended, the conditional analysis offers a radically incomplete semantic analysis. But when fully fleshed out, the analysis becomes viciously circular. I also consider just what the "zombie intuition" truly is, and defend the claim that a zombie world is conceivable-as-actual.

Chapter 7 moves on to the phenomenal concept strategy. I argue that no version of the phenomenal concept strategy can succeed. According to PCS even an ideal agent,

with complete physical information, would not be able to work out the connections between their physical way of thinking about the worlds and their phenomenal way of thinking about the world, because these two ways of thinking about phenomenology are so different. I show that for any theory of phenomenal concepts that can be proposed to support PCS, we can construct an agent to serve as a counterexample to the claim that an ideal agent would find an epistemic gap. I start by developing this line of argument against what I take to be the strongest version of PCS—that based on the constitutional theory presented in chapters 3 and 4—and then show how the idea can be extended to cause trouble for versions of PCS based on alternative theories of phenomenal concepts.

To do this, I design an agent who (1) has full general physical knowledge, (2) has phenomenal concepts that work just as the constitutional theory claims, and yet (3) is wired such that she would be in a position to immediately work out the phenomenal-physical truths. I argue that—assuming the truth of physicalism—this creature's immediate phenomenal-physical judgments amount to a priori knowledge. It follows from this that there is not an *ideal* epistemic gap between the phenomenal truths an the physical truths. Thus, the special nature of phenomenal concepts does not entail an ideal epistemic gap. The phenomenal concept strategy fails.

But from my argument that the special nature of phenomenal concepts cannot support the existence of an *ideal* epistemic gap, emerges an alternative and highly compelling way to be a type-A physicalist. We clearly have a very different psychology from the imaginary agent I'll describe. I suggest that it is this psychological difference of ours—we might say "this rational defect" of ours—that explains why *we* find an

epistemic and explanatory gap. I argue that this gets the proponent of PCS what they really were after all along: It gives us a way to be a physicalist that respects the intuitions that *Mary* would be fooled if her captors presented her with a blue banana, that *we* would find zombies conceivable no matter how much physical information we possessed, that *we* will never be able to bridge the explanatory gap. Accepting this version of physicalism requires abandoning the idea that these gaps are *ideal*, but plausibly our intuitions about zombie cases and the like were never about what creatures with radically different psychologies from our own would be able to conceive of; rather they were about creatures psychologically like us.

This should be attractive to physicalists of all varieties: Type-A physicalists can continue to maintain a tight connection between ideal conceivability and possibility, while respecting our most firmly held intuitions (something they've had difficulty doing). And though my argument shows that the letter of type-B physicalism is untenable—physicalism is not compatible with an ideal epistemic gap—we have a new way to be a physicalist that captures what I take to be the *spirit* of type-B physicalism. We have a way to be a physicalist while respecting our dualistic intuitions, a way that has the virtue of avoiding messy debates about the relationship between conceivability and possibility.

Does this new version of type-A physicalism pose a special challenge to dualism, and (particularly) for epiphenomenalism? This question is taken up in chapter 8. If there is a physicalistically acceptable explanation for why Mary and other creatures like us find an epistemic gap, shouldn't epiphenomenalists accept the same story about why we have our dualist intuitions? And, given that arguments for dualism are based on these

intuitions, doesn't this debunk the epiphenomenalist's arguments?

I argue that this does not debunk epiphenomenalism. Dualism – and epiphenomenalism in particular – is a defensible position. And just as phenomenal concepts can help the physicalist to develop a bullet-proof version of her theory, so phenomenal concepts can be of use to the dualist in developing an attack-proof version of dualism. The way we think about consciousness can help to reveal compelling and coherent versions of both physicalism and dualism.

The result is a stand-off. Neither the physicalist nor the dualist will be able to get any traction against the (reasonable) proponent of the other side. There is no independent way to adjudicate between the two positions, as defenses of each must start by assuming their own position.

This raises two questions, one within philosophy of mind and the other methodological. First, (how) can we make progress on the mind-body problem? If neither side can get traction against the other, does that mean that we must simply give up? Second, what is (and should be) our aim in debates like the mind-body problem? Is our aim to convince proponents of the other side? To convince neutral third-parties? What counts as a neutral third-party? Do human agents – with the weird intuitions that follow from our contingent psychologies – take priority over hypothetical agents with different wiring and so different intuitions?

We begin with the first way of appealing to phenomenal concepts to defend type-B physicalism: the conditional analysis of phenomenal concepts.

Chapter 6

Circularity in the Conditional Analysis of Phenomenal Concepts[30]

The conceivability argument started from the idea that a zombie world – a minimal physical duplicate of our world, in which there is no qualia – is ideally conceivable. Ideal conceivability is taken to be a good guide to possibility.[31] Hence, it's concluded that a zombie world is possible. But if that's so, then being physically like our world isn't sufficient for being like our world simpliciter: Physicalism is false.

We have seen that the standard responses to the zombie argument either deny that zombie worlds are ideally conceivable (a la type-A physicalism), or deny that conceivability is a good guide to possibility (a la standard type-B physicalism). But it seems hard to deny the intuition that zombies are conceivable, and unsatisfying to dismiss the intuition without explaining its grip on us. And many are committed to the idea that conceivability does tell us something about possibility.

David Braddon-Mitchell's and John Hawthorne's conditional analysis of phenomenal concepts aims to give a way of rejecting the zombie argument, while granting both (1) the zombie intuition, and (2) that conceivability is a generally good guide to possibility. Thus the conditional analysis aims to give a friendlier, more intuitive version of physicalism that can account for our dualist epistemic intuitions.

After laying out the conditional analysis and its motivation, I'll consider a standard objection to the analysis, due to David Chalmers and Torin Alter. They object

[30] A modified version of this chapter is forthcoming in *Philosophical Studies* as "Circularity in the Conditional Analysis of Phenomenal Concepts".

[31] This, of course, includes the earlier caveats that (1) only *ideal* conceivability is taken to tell us anything about possibility, and (2) the fact that it's conceivable e.g. that water is XYZ shows only that it's possible that there be watery stuff thats chemical makeup is XYZ.

that the analysis doesn't respect the *true* zombie intuition (that it's conceivable that a zombie world is actual), but rather gives us a weaker imitation zombie intuition (that it's conceivable that a zombie world is possible). I'll then present several counterarguments that aim to show that we *don't* really have the intuition that it's conceivable that a zombie world is actual. I'll argue that these arguments are unsuccessful.

In the final two sections of the paper, I present what I take to be a stronger case against the conditional analysis, arguing that when fully fleshed out, the analysis either becomes viciously circular or collapses into a rigidified version of analytic functionalism.

6.1 The Conditional Analysis of Phenomenal Concepts

According to the conditional analysis, phenomenal concepts have a conditional structure. On this analysis, the following conditionals are knowable a priori and place a constraint on the meaning of our phenomenal concepts:

1. If the actual world contains nonphysical states, then our phenomenal concepts refer rigidly to these states.[32]
2. If the actual world is merely physical, then our phenomenal concepts refer rigidly to whatever physical states play the right functional role in the actual world.[33]

[32] This clause might be too simple: Presumably, not just any nonphysical state is a phenomenal state, just as not just any physical state is a brain state. For this reason, one might want to rephrase the antecedent to be about the actual world containing nonphysical states *of the right sort*. Along these lines, Hawthorne includes the restriction that the nonphysical state "typically [be] caused in the way that folk think pain [or whatever other phenomenal experience] is caused." (Hawthorne, 2002, 25) Braddon-Mitchell does not include such a restriction. My argument against the conditional analysis, given in section three, hinges on the feasibility of spelling out this of-the-right-sort clause.

[33] Braddon-Mitchell offers an alternative formulation on which, if the actual world is merely physical, then in the actual world our phenomenal concepts refer to the physical states that actually play the right functional roles, and in each counterfactual world, our phenomenal concepts refer to whatever physical states play the right functional role in that world. I'll ignore this for simplicity. It won't affect my argument against the conditional analysis.

If the conditional analysis is right, then we cannot rule out the possibility of zombies a priori: It's conceivable that we live in a world with nonphysical stuff (we surely can't rule this out a priori). But if such a world were actual, then zombies would be possible. This is supposed to explain the intuition behind the first premise of the zombie argument. And though intuitions of conceivability might generally be a reliable guide to possibility, we can't infer the possibility of zombies from this intuition: If the actual world contains nonphysical stuff, then zombies are possible. But if (as the physicalist maintains) the actual world turns out to be purely physical, then zombies are *not possible.* So we allegedly have a way of accepting the zombie intuition, while rejecting the argument's conclusion.

Why think that phenomenal concepts have a conditional structure? Advocates of the conditional analysis rely on two-dimensional semantics to argue that the way we would behave, given different revelations about the actual world, reveals something about the meaning of our phenomenal terms and the nature of our phenomenal concepts. According to two-dimensional semantics, what our concepts (not different concepts by the same name) refer to in different worlds considered as actual reveals the meanings of these concepts.[34] For example, in analyzing the concept WATER, we consider what WATER will refer to if the actual watery stuff turns out to be XYZ, if the actual watery stuff is H_2O, if there is actually no watery stuff, if the actual watery stuff is a mass hallucination, etc. This reveals that WATER, in any world considered as actual, picks out the clear, odorless, tasteless … liquid. According to two-dimensional semantics, this

[34] I've avoided using any technical terminology. But for those unfamiliar with two-dimensional semantics, it may be helpful to refer to Chalmers' introduction to the subject: Chalmers, "Two-Dimensional Semantics", as the basic ideas are essential to understanding the conditional analysis and my objections to it.

reveals the meaning—or an important aspect of the meaning—of WATER.

When we analyze a concept in this way, we don't restrict ourselves to considering as actual worlds that conform with what we know to be the case a posteriori. (If we restricted ourselves to worlds that conform with what we know a posteriori, we would find that WATER *means* H_2O: something clearly false!) Rather, to analyze the meaning of a concept, we consider each *a priori conceivable* world, and assess what the concept refers to in that world. As Braddon-Mitchell writes, "[t]he conditional analysis I have proposed is not supposed to be a posteriori. We do not look at the nature of actual qualia to find that the conditional analysis is true. It is, after all, an analysis" (Braddon-Mitchell 2003, 131).

Proponents of the conditional analysis take two-dimensional semantics to reveal that phenomenal concepts have a conditional structure. Braddon-Mitchell argues that we *hold fixed* that we have phenomenal experiences, even given shifting assumptions about the metaphysical nature of these experiences: "[I]f I become persuaded of dualism I think [phenomenal experiences] are dualistic, if I lose faith I decide they are physical, if I regain the faith I think they are dualistic, and so on" (Braddon-Mitchell 2003, 130). The idea is that, when we abandon dualism in favor of physicalism (or visa versa), we don't come to *mean* something different by PAIN.

Hawthorne makes this point vivid through a thought experiment in which an oracle tells you about the nature of the world. It seems conceivable a priori that the oracle tell you that the world is merely physical, or that she tell you that the world contains nonphysical states. If she tells you that the world is merely physical, you will not

conclude that there are no phenomenal experiences in the world. Rather, you will conclude that phenomenal experiences are the physical states that play the relevant functional roles. Similarly, if the oracle tells you that there are nonphysical states of a certain kind, you will conclude that these states (and not physical-functional states) are the phenomenal experiences. (Hawthorne, 26)

Whatever sort of state our phenomenal concepts refer to, they seem to refer rigidly: If we learn that there are nonphysical states of the right kind, then when we consider counterfactual worlds lacking these states, we will take them to be zombie worlds. Similarly, if the actual world is purely physical, then even in counterfactual worlds with nonphysical stuff, the physical states will be what our phenomenal concepts refer to; the nonphysical stuff will simply be a weird extra. Thus, our dispositions to use phenomenal concepts given different situations considered as actual reveal that what phenomenal concepts refer to depends on the actual nature of the world.[35]

6.2 Does the Conditional Analysis Yield the Right Intuition?

Chalmers and Alter have argued that the conditional analysis doesn't explain the right intuition. In this section, I discuss the Chalmers/Alter objection and offer some arguments to bolster it. In the following sections, I set this objection aside to present

[35] It's not clear that these intuitions are by themselves sufficient to push one towards a conditional semantic analysis for phenomenal concepts. For example, Papineau and Balog's "constitutional theory" of phenomenal concepts, satisfies the motivation for the conditional analysis, but doesn't seem to offer a conditional *semantic* analysis. On this theory, phenomenal concepts are partially constituted by the phenomenal experiences they refer to. Nothing in this analysis of phenomenal concepts specifies whether the experiences that partially constitute phenomenal concepts are physical or nonphysical. So it is not *analytic* that phenomenal experiences refer to either physical or nonphysical states. Thus, when we come to believe in or abandon dualism, this does not require our phenomenal concepts to change. It just requires our beliefs about the nature of the states composing the concepts to change. While the constitutional theory may be conditional at a *metasemantic* level, insofar as the analysis *of our analysis* of phenomenal concepts depends on the nature of reality, it is not conditional at a semantic level.

another—and I think stronger—objection to the conditional analysis.

Chalmers and Alter argue as follows: The conditional analysis tells us that it's conceivable that it's possible that there are zombies. (It's conceivable that the actual world is such that some counterfactual world is a zombie world.) But the intuition we want to explain is the intuition that it's conceivable that the actual world is a zombie world. We might call this an intuition of "direct" conceivability, to contrast it with the doubly modal claim the conditional analysis yields.

Braddon-Mitchell, in defending the conditional analysis, simply denies that the zombie world is directly conceivable (Braddon-Mitchell 2003, 129). The defender of the conditional analysis might argue that we confuse the intuition that it's conceivable that a zombie world is possible for the intuition of direct conceivability, and so mistakenly think that zombies are directly conceivable. Plausibly, our modal intuitions aren't sufficiently fine-grained for us to reliably distinguish between the intuition that zombie worlds are *conceivable as actual* and the intuition that they're *possibly conceivable*. So explaining why we have an intuition that's in the ball-park might be good enough to explain the intuitive pull of the first premise of the conceivability argument.

Alter counters that, while we sometimes misinterpret our modal intuitions, this doesn't seem to be a case of misinterpretation. When we misinterpret a modal intuition, we find that the intuition changes once we've clarified the various possible intuitions we could be trying to latch onto. (E.g. once we've distinguished *the claim that Hesperus is not identical to Phosphorus* from *the claim that Hesperus is not the evening star*, we no longer have the intuition that the former is possible.) But Alter insists that even after

distinguishing the conditional analysis's doubly modal claim from the claim of direct conceivability, it still seems that the zombie world is directly conceivable. The claim of conceivable possibility doesn't push us to abandon the idea of direct conceivability. (Alter, 14)

In addition to arguing that our modal intuitions are insufficiently fine-grained to tell against the conditional analysis, defenders of the analysis have argued that we have positive reasons to deny having the intuition that Chalmers and Alter describe as the true zombie intuition. If this is right, the intuition underlying the zombie argument must be something else, like the conditional analysis's doubly modal claim. While I don't think it's immediately obvious that our intuitions reliably push us either way on such subtle modal issues, I think that these further arguments against the intuition of direct conceivability are unsuccessful. In assessing why these arguments fail, I hope to (1) strengthen the idea that we have the intuition of direct conceivability, and (2) reveal some problems for the conditional analysis that will be exploited in the argument against the conditional analysis given in section three, and (3) reveal some interesting facts about the relationship between considering worlds as actual and as counterfactual.

Jussi Haukioja argues that it's not possible to directly conceive of a zombie world, as "this would entail the direct conceivability of *myself* as lacking consciousness" (Haukioja, 10). But surely it couldn't turn out that I actually lack consciousness. If this is right, the zombie intuition must be something other than an intuition of direct conceivability, such as the doubly modal claim achieved by the conditional analysis. This would provide a point in favor of the conditional analysis.

It's not entirely clear what Haukioja has in mind. One thing he might be thinking is this: I know that I have phenomenal experiences. I know this with a kind of certainty that I cannot know other things (like the chemical makeup of the watery stuff). While I can imagine a situation in which it turned out that I was wrong about the nature of water, I cannot imagine a situation in which it could turn out that I was wrong that I'd had phenomenal experiences. But to conceive of a world as actual just is to imagine discovering that the world is actually that way. So I can't conceive of a world in which I lack phenomenal experiences as actual.

But as we have seen, when analyzing a concept, we cannot restrict ourselves to considering as actual worlds that conform with what we know to be the case a posteriori. (No one thinks that what's conceivable *compatible with our a posteriori knowledge* reveals something about the meanings of our concepts.) When I consider the meaning of WATER, I must consider what WATER would refer to if the actual watery stuff was XYZ, if the actual watery stuff were a mass hallucination, and so on. Although my knowledge that I have phenomenal experiences has a kind of certainty that my knowledge that water is H_2O doesn't have, it is like my knowledge that water=H_2O insofar as I can only know that I have phenomenal experiences *through experience*. In analyzing our phenomenal concepts, we can no more restrict ourselves to considering the worlds with qualia than, in analyzing WATER, we can restrict ourselves to worlds in which the watery stuff is H_2O. But it seems *a priori* conceivable that I, along with everyone else in the world, lack qualia. Unless we're given some reason to deny this, the analysis must accommodate a third possibility: that the world is merely physical, and

there are no phenomenal experiences. And this would run contrary to the motivations of the conditional analysis to accept that physicalism is true, but that conceivability bears a tight connection to possibility. Alter makes a similar point in response to Braddon-Mitchell and Hawthorne. (Alter, 6-10)

But since Haukioja is offering a response to Alter, perhaps this is not what he has in mind. Perhaps he is intending to offer a reason why the zombie world is not *a priori* directly conceivable. The idea might be something like the following: To consider a world as actual, I must have a perspective into the world: a viewpoint through which to view what the world is like. The world's center is the location of this viewpoint. If there's nothing that it's like to be the center of the world, I lose my grasp on what the world is like. The viewpoint goes dark. So at the center of any world we consider as actual, there must be consciousness. Hence we can't conceive of a zombie world as actual.

But just because I must determine what the world is like, this doesn't require there to be a what-it's-like situated in the world I'm considering as actual. This reading of Haukioja seems similar to Berkeley's argument that there couldn't be an unperceived tree. Though I must form a representation of the tree in order to conceive of it, that doesn't mean that the content of my representation includes that the tree is represented (is conceived of). Similarly, though I must imagine what worlds are like from a viewpoint inside the worlds, that doesn't mean that what I imagine must include that the viewpoint has a "what it's like". Speaking metaphorically, we might say that the world's center serves as a window. I can stand outside of the world and look through the window to imagine what the world is like, with out there being any "what it's like" represented

within the world.

One might object to this on grounds that what I've described—standing outside of a world, looking in—amounts to considering a world as counterfactual. When you consider possible worlds as actual, the argument goes, you situate yourself *inside* the world you're taking to be actual. Any world that you consider from an outside world is merely a world you're considering as counterfactual.[36]

If this were right, it would rule out far more than the direct conceivability of zombie worlds. If we're to assess a world from *inside* that world, there must not only be a what-it's-like situated in the world, but also conscious thoughts occurring in the world. Thus, the above suggestion would have it that we cannot conceive-of-as-actual worlds without conscious, rational thought (rock worlds, worlds populated by animals with no rational thoughts). But clearly such worlds are directly *a priori* conceivable. So we should hope that this suggestion is misguided.

To respond to this objection, I think that we need to differentiate between the *abilities* we use to evaluate worlds and the things we build in as *features of the worlds* we evaluate. In evaluating what a world is like, I have to employ certain abilities: I have to be conscious, I have to be capable of forming coherent thoughts, and so on. When I evaluate a zombie world, I employ these abilities. When I evaluate a world consisting of a rock in a void, I employ these abilities. When I consider a world populated by rational, conscious beings, I employ these abilities. Whether I consider these worlds as counterfactual or counteractual—and whether the world I'm considering has conscious, rational beings or not—I must employ these abilities *in this world.* (After all, this is the

[36] Thanks to Jack Woods for this objection.

only world I am in!) But just because I must employ these abilities to consider a world, doesn't mean that I need to build the abilities in as features of the world I'm considering. To think that this is required is to make the mistake Berkeley made with his unconceived tree.

The difference between considering worlds as actual and as counterfactual lies not in what *abilities* we use, or where these abilities are used, but in what *features of our world* we hold fixed. In considering a world as counterfactual we mentally "hold fixed" certain features of our world while we assess a world whose features differ. In considering worlds as actual, we don't mentally "hold fixed" features of our world, but allow our thoughts to become submersed in whatever features we are imagining. So this objection—that zombie worlds aren't conceivable, because I cannot conceive of myself actually being a zombie—does not work.

Once we appreciate why these arguments against the intuition of direct conceivability fail, I think it becomes more plausible that we really do have the intuition of direct conceivability. The intuition of direct conceivability is a surprisingly minimal claim. The claim is *not* that, for all we know, we actually do live in a zombie world. We can be certain that we have phenomenal experiences; hence we can be certain that the actual world is not a zombie world. The claim is simply that, *for all we could know a priori*, we actually live in a zombie world.

Consider the claim that a world with no coherent thoughts is directly conceivable—that is, that it's a priori conceivable that the actual world lacks coherent thought. (This could be a world with nothing but rocks, or a world filled with extreme schizophrenics

who never have a moment of lucidity.) I know that the actual world is not such a world, and I could not even conceive of such a world were it actual—as conceiving of such a world requires the formation of a coherent thought—nevertheless for all I know a priori, such a world could be actual. This is all that the intuition of direct conceivability requires.

In the next section, I present what I take to be an even more devastating objection to the conditional analysis. I argue that, as stated, it offers a radically incomplete semantic analysis. Once we flesh out the analysis, it winds up becoming viciously circular.

6.3 Fleshing Out the Analysis and Finding Circularity

The strategy used by proponents of the conditional analysis was to consider various a priori conceivable worlds *as the actual world*, and assess what our phenomenal concepts refer to in these worlds. This is supposed to reveal something important about the nature of our phenomenal concepts: that they refer to physical states in some worlds, nonphysical states in other worlds, and an analysis of them must be conditional in structure—revealing that what sort of thing phenomenal concepts refer to depends on what the actual world is like. The oracle thought experiment was designed to show (1) that it's a priori conceivable that the world be merely physical or that the world have nonphysical states, (2) that what our phenomenal concepts refer to depends on the metaphysical structure of the actual world in the way captured by the conditional analysis.

In this section, I'm going to use the very strategy employed by the conditional analysis to show that, as stated, the analysis is radically incomplete. I'll modify

Hawthorne's oracle thought experiment to show that the conditional analysis neglects to consider certain a priori conceivable cases. If the conditional analysis is to be offer a complete semantic analysis, we must add more clauses to accommodate these additional conceptual possibilities. I'll argue that we cannot flesh out these clauses without rendering the conditional analysis viciously circular.

When the oracle says that the world contains nonphysical states, the conditional analysis tells us that our phenomenal concepts necessarily refer to these nonphysical states. This seems to miss a possibility. Imagine that the oracle tells you that the world contains nonphysical states, and additionally tells you that *even if you'd lacked these nonphysical states*, you still would have had what-it's-likes. This seems conceivable.[37] If the oracle tells us this, we will surely not conclude that our phenomenal concepts refer to nonphysical states. We will either conclude that phenomenal concepts refer to physical states (and the nonphysical states are just some weird extra) or we would conclude that phenomenal experiences are overdetermined, and phenomenal concepts refer to both physical and nonphysical states. Zombies are not be possible relative to this world.

But proponents of the conditional analysis don't consider this possibility. According to the conditional analysis, if there is nonphysical stuff in the world, then that nonphysical stuff is qualia; zombies are possible. This suggests that proponents of the conditional analysis are building the truth of (non-overdetermined) dualism into worlds with nonphysical stuff. This fits with the motivations they give for the analysis: They

[37] One might deny that this was conceivable, on grounds that we should be talking not merely about "nonphysical states", but about "nonphysical states *of the right sort*". If you favor adding an "of the right sort" clause, you can read this as a call to flesh out this clause in a meaningful way. As will become clear, I think fleshing this clause out will either require reliance on phenomenal concepts, making the resulting analysis circular, or will collapse into a form of analytic functionalism. Further, such a response seems to rule out as inconceivable the possibility of overdetermination.

argue that we shouldn't have zero credence that dualism is true. And if dualism is true, then zombies are possible.[38]

It's natural to wonder why we can't fixate on one of these counterfactual zombie worlds and consider it as actual, *holding fixed the truth of dualism.* The proponents of the conditional analysis explicitly disavow such a possibility: When the oracle tells us the world is merely physical, the conditional analysis says that our phenomenal concepts refer to physical states. The justification for this is the presupposition that phenomenal concepts always refer—a presupposition that in any world considered as actual, there is qualia. Because of this presupposition, any purely physical world considered as actual is a world in which physicalism is true.

Hence, on the conditional analysis, when I consider as actual a world in which there are nonphysical states, I build the truth of dualism into the world. Then when I fixate on a zombie world (possible relative to that world), and consider it as actual, what I imagine changes: I shift to building the truth of physicalism into the world. This tacit shift in theory-of-consciousness is what's doing all the work.[39]

But we're neither warranted in building the truth of dualism into all worlds with nonphysical states, nor are we warranted in building the truth of physicalism into all

---

[38] As Braddon-Mitchell implies, while a physicalist, he is not an arrogant one, and hence "[does] not give zero credence to physicalism being false" (Braddon-Mitchell 2003, 127).

[39] This implicit shifting in theory-of-consciousness when we consider worlds in different ways is quite strange. When we ask whether a world is a zombie world or a world with consciousness, we're asking whether there's a *what-it's-like* in that world. Intuitively, whether or not there's a *what-it's-like* in a world doesn't depend on whether that world is being thought of as actual or counterfactual. It depends on *what there is in that world*. Words may be able to shift their referents depending on the attitude we take to the world they're being used in. But surely *whether there's something that things feel like* doesn't depend on the attitude you take towards a world! On the conditional analysis whether there's a what-it's-like in a world or not depends not just on intrinsic features of that world, but on whether *we* are taking an attitude towards the world whereby we "hold fixed" features of our world (considering the world as counterfactual) or whether we mentally submerse ourselves in the world (considering the world as actual).

merely physical worlds. First, consider what I have described as the tacit assumption of physicalism in all purely physical worlds. This is grounded in the presupposition that phenomenal concepts refer in every world considered as actual. (No matter what the oracle tells us, we'll conclude that our phenomenal concepts refer.) As we've seen, this presupposition is justified on a posteriori grounds (the fact that we've had experiences), which cannot place restrictions on the meaning of our phenomenal concepts.[40] So there are two a priori conceivable possibilities consistent with the oracle telling us that the world is merely physical:

1. There are phenomenal experiences, and our phenomenal concepts refer to physical states.
2. There are no phenomenal experiences, and our phenomenal concepts don't refer,

And, if what I have said above is correct, there are three possibilities in the case in which the oracle tells you that the world has nonphysical states:

3. There are nonphysical states, and phenomenal concepts refer to these states.
4. There are nonphysical states, but phenomenal concepts refer to physical-functional states (or refer to both nonphysical and physical-functional states).
5. There are nonphysical states, but phenomenal concepts don't refer.[41]

[40] One might object that I can't conceive of a world in which my phenomenal concepts don't refer to anything: Surely "zombies" would refer to something with their uses of 'pain'. But this is beside the point, much as it would be beside the point that inhabitants of some world might use 'pain' to refer to puppies. What's relevant is what *our* term would refer to in such a world, and it seems clear that it wouldn't refer.

[41] This would be a case where the physical states aren't phenomenal states, and though there are nonphysical states, these states are intuitively not *phenomenal* states either. (Surely we can imagine that there are nonphysical states that are just some weird extra, and have nothing to do with our mental lives!) So in such a world, phenomenal experiences, although there are both physical and nonphysical states.

The conditional analysis only recognizes (1) and (3). If the situations captured by (1)-(5) are a priori conceivable, then the conditional analysis is in trouble. It is intended to give a semantic analysis of our phenomenal concepts, but the analysis is radically incomplete. The proponent of the conditional analysis might argue that we simply need to add more clauses to account for these additional conceivable scenarios.[42] To account for (1)-(5), the analysis would need to be modified to something like this (using PAIN as an example):

1. If the actual world is merely physical *and we have pains*, then PAIN refers rigidly to the physical states that actually plays the right functional roles.
2. If the actual world is merely physical *and we don't have pains*, then in the actual world PAIN doesn't refer, and in no merely physical counterfactual world does PAIN refer.
3. If the actual world contains nonphysical states *and we wouldn't have pains were there not nonphysical states*, then PAIN refers rigidly to these nonphysical states.
4. If the actual world contains nonphysical states *and we would have pains even if the world were merely physical*, then PAIN refers rigidly to the physical state that plays the right functional role in the actual world.[43]
5. If the actual world contains nonphysical states *and we don't have pains*, then PAIN doesn't refer in this world or in any counterfactual world or in any counterfactual world that only contains this sort of nonphysical states.

This is supposed to give us an analysis of the concept PAIN. But in order to accommodate the additional conceptual possibilities, the analysis has been forced to

[42] This is a strategy Haukioja takes in responding to Alter's objection that an oracle could tell us that the world is purely physical and nothing plays the role of qualia.

[43] This is something of a simplification, as we have not included a clause to accommodate the possibility of overdetermination. For simplicity—and because it will not pose any problems to the argument I go on to give—I will omit this clause. Note that, if anything, the need to accommodate such a clause poses even *more* difficulty for the conditional analysis.

employ the very phenomenal concepts it is trying to explain. The analysis has become circular.

(Note that even if you reject the idea that a zombie world is directly a priori conceivable, the analysis is still circular, in virtue of the need to spell out what nonphysical states are phenomenal states: (3) - (5). It is presumably undeniable that we can directly conceive of nonphysical states that are not phenomenal states! If you find (2) dubious, you can read the remainder of this paper as a call to flesh out (3) - (5).)

One might respond by trying to spelling out these additional clauses without employing phenomenal concepts. But it's difficult to see how such a move could be successful. In the next section, I'll consider three types of strategies one might use to spell out these clauses in a non-circular way. I'll argue that the only hope of saving the conditional analysis requires adopting a rigidified version of analytic functionalism. But first, I want to flesh out just what is problematic about the above analysis of phenomenal concepts and to consider whether the circularity the conditional analysis faces is vicious.

6.3.1 Vicious Circularity

The circularity I've argued that the conditional analysis faces certainly seems to be vicious. If someone who had no prior notion of PAIN were given the analysis, it would offer them little help in understanding the concept. Given this analysis, to know whether there are pains in a world—and if so, which things are the painy things—we must already know whether there are pains in the world!

Perhaps a better test of vicious circularity is to imagine creatures who do have a

concept that I would verbalize as ‘pain’, but where they verbalize it in a different way. Suppose Sue has a concept that she calls ‘phi’ that is in fact the same concept we express as ‘pain’. Would the conditional analysis—together with the relevant facts about the nature of each a priori conceivable world—enable Sue to work out the synonymy? Sue’s concept, by hypothesis, has the following structure:

1. If the actual world is merely physical *and we have phi-s*, then PHI refers rigidly to the physical states that actually plays the right functional roles.
2. If the actual world is merely physical *and we don’t have phi-s*, then in the actual world ‘X’ doesn’t refer, and in no merely physical counterfactual world does PHI refer.
3. If the actual world contains nonphysical states *and we wouldn’t have phi-s were there not nonphysical states*, then PHI refers rigidly to these nonphysical states.
4. If the actual world contains nonphysical states *and we would have phi-s even if the world were merely physical*, then PHI refers rigidly to the physical state that plays the right functional role in the actual world.
5. If the actual world contains nonphysical states *and we don’t have phi-s*, then PHI doesn’t refer in this world or in any counterfactual world or in any counterfactual world that only contains this sort of nonphysical states.

Our concept ‘pain’ has the same structure. But this similarity in structure isn’t enough for Sue to determine that the concepts are the same. To determine this, Sue would need to know whether phi-s (in the antecedent) and pains (in the antecedent of the analysis of our concept) were the same thing. And this is precisely what Sue doesn’t know!

The problem is roughly this: The meaning of PAIN is given by what PAIN refers to in different worlds considered as actual. To know this—to have a complete analysis of PAIN—we must be able to specify all of the relevant features of each world

considered as actual. But one relevant feature (as we can see by comparing clauses (1) and (2), or clauses (3), (4), and (5)) is whether the world *has pains in it*. And the analysis doesn't have a way of telling us this. If we plug in what the conditional analysis does tell us, in place of 'pain' in the antecedents of the conditionals, we get an uninformative infinite regress.

Even if Sue knew all about my uses of 'pain' in purely physical worlds and in worlds with nonphysical states, she wouldn't know everything to know about my uses of 'pain' unless she also knew *whether these worlds had pains in them.* We're lacking a neutral way for Sue to know about the worlds, and hence to know whether I use 'pain' in the same way she uses 'phi'.

You might think that *any* analysis of phenomenal concepts will be problematically circular in this way. For example, consider the constitutional analysis of phenomenal concepts, from chapters 3 and 4. According to this theory, phenomenal concepts are partially constituted by the experiences they refer to. The concept RED is taken to have the structure "the experience *red experience*," where what fills the '*'s is an instance of the relevant type of experience. When we employ a phenomenal concept, the what-it's-like of the experience embedded in the concept is revealed to us, but the metaphysical status of this experience (as physical states or nonphysical states) is not. The constitutional theory doesn't face straightforward circularity worries: Although we must use phenomenal concepts in order to *describe* the constitutional analysis, it is not the case that the *analysis* employs phenomenal concepts. The constitutional analysis employs phenomenal *properties* as constituents, not phenomenal concepts.

But when we use two-dimensional semantics to think about the constitutional theory, it seems to face the very circularity problem faced by the conditional analysis. Papineau and Balog suppose that it's *epistemically* possible that the conscious experiences be physical states, or that they be nonphysical states. They also allow that zombie worlds are epistemically possible, and would doubtless grant that it's epistemically possible that there are worlds with nonphysical states that are not phenomenal states. To give a two-dimensional analysis of these phenomenal concepts, we must have a way of specifying not only whether the world in question has nonphysical states, but of specifying what states (if any) are the *phenomenal* states.[44] But, since not all nonphysical states are phenomenal states, they'll still have to specify what states – if any – are the phenomenal states.

And defenders of the constitutional theory are committed to the idea that we cannot specify what states are phenomenal states in terms other than phenomenal terms: They think that phenomenal concepts are conceptually isolated from physical, functional, and all other nonphenomenal terms. Thus to specify all the relevant features of each world considered as actual (so as to give a two dimensional analysis), we must use phenomenal concepts. This looks like the very circularity faced by the conditional analysis.

But this doesn't pose the same sort of challenge to the constitutional theory that it does to the conditional analysis. The constitutional theory takes what we might call an *experience first* approach: It starts from the assumption that there are these things,

[44] Dualists who endorse the constitutional theory, will deny that it's epistemically possible that conscious states by physical states.

phenomenal experiences, and uses these experiences to build up an account of how we refer to them. It makes no pretenses of giving a being able to pick out these experiences in third-personal terms. The fact that phenomenal experiences cannot be analyzed in nonphenomenal terms is (according to physicalist proponents of the view) one of theory's main virtues: It's because phenomenal concepts are conceptually isolated from other concepts, that zombies are conceivable and that Mary seems to learn something *new*.

Contrast this with the conditional analysis. The conditional analysis takes what we might call a *concepts first* approach: We start by describing worlds, and then looking for what our concepts latch onto. This can be seen in the process by which defenders motivate the conditional analysis: Suppose the world turns out to be like this, what do our phenomenal concepts refer to? (Contrast this with the constitutional theory: Here's a phenomenal experience; how do we form a concept of it?) It can also be seen in the fact that according to the conditional analysis, whether a world is a zombie world depends *not just on what is/isn't in the world* but on the attitude that we take towards the world. It's not the intrinsic nature of the world that makes it a zombie world, but whether *our phenomenal concepts* refer to anything in the world.[45]

Whereas the constitutional theory is simply concerned to tell us how to construct phenomenal concepts given the prior assumption of phenomenal experiences, the conditional analysis is after something more ambitious. The analysis of our concepts is supposed to reveal to us whether there are phenomenal experiences in the world. It cannot do this if the only way to specify a world fully enough to tell whether it has

[45] Thanks to Richard Yetter Chappell for his helpful discussion of the aims of the constitutional theory and the conditional analysis, and for the descriptors 'experience first' and 'concept first'.

phenomenal experiences is to employ phenomenal concepts in describing the world.

Why can't the conditional analysis follow the constitutional theory in simply assuming that there are phenomenal experiences, and tackling the more minimal problem of how we refer to these experiences? The conditional analysis could then accept that to fully specify a world, such that we could know what (if anything) our phenomenal concepts refer to in that world, we must specify whether the world has phenomenal experiences. And the only way to do this is to use phenomenal concepts.

One problem is that the conditional analysis just doesn't seem to have anything to say on the matter of *how* we refer to phenomenal experiences. That's simply not what the theory is designed to do. Further, if we accept that the only way to know whether our phenomenal concepts apply in a world is to have the world described in phenomenal terms, we lose all motivation for the conditional analysis. The conditional analysis was put forward as a novel physicalist-friendly way of explaining the zombie intuition. If the only way we can know whether our phenomenal concepts refer is to have a world presented to us in phenomenal terms, the phenomenal terms must be conceptually isolated: It must be the case that an a priori derivation of the phenomenal from the nonphenomenal is impossible. But if our phenomenal concepts are conceptually isolated from all other concepts, we already have a straightforward explanation of the zombie intuition: that provided by the phenomenal concept strategy. If our phenomenal concepts are conceptually isolated, the physical description of the world won't be enough to tell us whether there are phenomenal states, hence zombies will be directly conceivable. This is a more satisfactory explanation of the zombie intuition, as it gives us what Chalmers and

Alter argue is the true zombie intuition. And it means there's no work to be done by the conditional analysis's alternative explanation.

If what is distinctive about the conditional analysis is to be retained, it cannot hold that to know whether our phenomenal concepts refer, we must use phenomenal concepts to specify the world.

6.4 Can Circularity be Avoided?

Proponents of the conditional analysis argued that how we would behave given different revelations about the actual world reveals something important about the nature of our concepts: they have a conditional structure, such that what they refer to depends on the metaphysical nature of the actual world. The resulting analysis of our phenomenal concepts is supposed to give a novel way of accommodating the zombie intuition. I've argued that proponents of the conditional analysis fail to consider certain conceptual possibilities, and so give an incomplete analysis of our phenomenal concepts. But to flesh the analysis out, we seem to have to use the very concepts that we're trying to give an analysis of. This renders the analysis viciously circular. We've seen that the conditional analysis can't accept that phenomenal concepts are conceptually isolated in a way that renders a non-circular analysis impossible, if it is to have anything distinctive to say.

In what follows, I consider three types of strategies one might use to spell out five clauses of the revised conditional analysis in a non-circular way: an appeal to "the folk", an appeal to theories of reference, and an analytic functionalist analysis. I'll argue that the first two approaches do not help avoid circularity. While I take the analytic functionalist

model to be internally coherent, I'll argue that it is unmotivated and highly unattractive.

6.4.1 Appeal to "the Folk"

We might try to appeal to something like "what the folk use 'pain' to refer to", but then we've simply passed the buck. We're left without any analysis of what the folk use 'pain' to mean. What's worse, phenomenal concepts plausibly *are* folk concepts, so we're left without an analysis of what we've been after all along.

Hawthorne implicitly does something along these lines. He suggests building into the conditional analysis a restriction on what sort of nonphysical states are eligible to be the referents of phenomenal concepts: those that are "typically caused in the way that folk think pain [or whatever other phenomenal experience] is caused" (Hawthorne, 2002, 25). It's far from obvious that the folk have any robust beliefs about pain other than that it's "hurty". But even if there is common folk agreement about how nonphysical pain-states are caused, we surely don't want to bind the reference of our phenomenal concepts to these beliefs. Surely it's conceivable that hurtiness is a matter of being in a nonphysical state, and that the folk are wrong about its causal history. And in this case, we (and all the folk!) would surely agree that 'pain' referred to this hurty feeling, despite its causal history. What is essential—to the folk as well as philosophers—is the feeling of the state.

6.4.2 Theories of Reference

The following "toy" response might seem to suggest a way of fleshing out these causes without employing phenomenal concepts: Suppose that the reference of PAIN is

given by whatever state causes a tokening of the concept PAIN. We could then replace all *uses* of the concept PAIN with "state that causes tokenings of the concept 'PAIN'"—where this merely *mentions* the concept as opposed to using it. Thus, it might be held, we don't need to employ the concept we're analyzing in the antecedent. While this simplistic view of reference is not promising, we can build on the format of this response. Take some more complex view of reference. If what it takes to be the referent of a concept can be spelled out without making use of that concept, then we can similarly replace all uses of the concept PAIN with "state that stands in so-and-so reference relation to the concept of 'PAIN'." Again, we don't need to employ the concept we're analyzing in the antecedent.[46]

But I don't think that this strategy works. Start by considering the "toy" example, where we replace uses of PAIN with "state that causes tokenings of the concept 'PAIN'." To fully flesh out this clause, we need to specify what counts as a token of the concept 'PAIN', and to do so without again employing the concept.

As a model, suppose we take something to be a token of the concept PAIN iff the agent expresses it using the verbalization 'pain'. It's clear that this will not do. Following this suggestion, we'd have to replace PAIN with "state that causes tokenings of a concept that is verbalized 'pain'." But this is not an acceptable replacement of PAIN (the concept *we* use, which we are attempting to give an analysis of), as in some other world, thoughts about puppies might be verbalized using the word 'pain'.

This is indicative of a broader challenge: All sorts of concepts stand in reference relations to all sorts of objects. TABLE refers to tables, CAR refers to cars, THREE refers

[46] This strategy was suggested to me in conversation by John Hawthorne.

to the number three. We don't want it to be the case that PAIN—*our* concept, not some other concept by the same name—could refer to any of these things in some other world. (This was what went wrong in our toy example: The relation of "causing tokenings of a concept that is verbalized 'p-a-i-n'" is something that can relate concepts to all sorts of different things that aren't pains.) If we are to be able to substitute "state that stands in so-and-so reference relation to the concept of 'PAIN'" for PAIN in the problematic clauses, we need to make sure that this clause will always refer to *pains*. There seem to be two ways that this could be achieved. Either we could try to build something into the reference relation that would restrict the class of things that are suitable to be the referent of "the concept of 'PAIN'," or (more plausibly) we could give some way of individuating "the concept of 'PAIN'" that will ensure that "the concept of 'PAIN'" picks out *pains*. But whichever way we try to solve this problem, we must not make *use* of the concept pain. For, recall, the goal here was to find some way of referring to pains without employing the concept we're trying to analyze.

But now we have found ourselves back where we started. How can we spell out why "the concept 'PAIN'" must pick out pains, if not by reference to the *hurtiness* or *paininess* of what they pick out?[47]

[47] Another thought one might have for avoiding the problems that I've suggested is to replace uses of PAIN with "state that *actually* stands in so-and-so reference relation to the concept of PAIN." There are two ways to read this suggestion, neither of which is helpful. Let's look at how this phrase would be embedded in one of the conditional analysis's clauses. (1) becomes (1').

(1') If the actual world is merely physical *and we have states that in the actual world stand in so-and-so reference relation to the concept 'PAIN'*, then 'PAIN' refers rigidly to the physical states that actually play the right functional roles.

(1') can be read in one of two ways. On the first, both 'actual world's refer to the same world, so that the second reference to the 'actual world' is redundant. On the second way of reading it, they refer to different worlds, giving us something like this: If the world we're considering as actual is merely physical and, *in this world, we have states that are the very same states which in our world—the "really" actual world—stand in so-and-so reference relation to the concept 'PAIN'*, then 'PAIN' refers rigidly to the physical states that, in the world we're considering as actual, plays the right functional

6.4.3 Functionalist Analysis

One way you might answer this question is by appealing to analytic functionalism. (This might be thought of as a specific version of the "theories of reference" approach discussed above, which has the potential to avoid the arbitrariness referent worry.) Proponents of the conditional analysis may not be happy with this move. Braddon-Mitchell, for example, explicitly builds into his motivation for endorsing the conditional analysis that "[t]he claim that 'qualia are whatever plays certain functional roles' is *not* directly analytic." (Braddon-Mitchell, 2003, 113; emphasis added) But, faced with vicious circularity, I want to see whether the conditional analysis has any path out of this bind.

If analytic functionalism is true, then we will not need to add clauses (2), (4), and (5) to the conditional analysis, as the situations I used to motivate the need for these clauses will not be a priori conceivable. For suppose it's a priori that PAIN refers to "state (physical or nonphysical) that plays so-and-so functional roles." Barring overdetermination, it will then *not* be ideally conceivable that the oracle tells us "the world has nonphysical states (of the relevant sort), and even if you lacked these states, you still would have had what-it's-likes." It will not be ideally conceivable that the oracle tells you "there are nonphysical states (of the right sort), but there are no what-it's-likes." Nor will it be conceivable that there be a world physically just like our world, but with no

roles. On this reading, the second 'actual' isn't trivial. But the resulting analysis is of no help to the conditional analysis. For according to the conditional analysis, pain might have a very different nature in some other world-considered-as-actual from the nature it has in our world. What we emphatically do not want is to build into the antecedent that pain in some other world we're considering as actual is the state that in our world is the referent of 'PAIN'.

what-it's-likes. So we will not need the additional clauses I've suggested we need.

But once we have given an analytic functionalist style analysis of phenomenal concepts, it's difficult to see what further work is done by this conditional structure. (Haven't we already accepted an analysis of phenomenal concepts by accepting a functional analysis?) Perhaps this again is acceptable. Maybe the take-home message from the conditional analysis should be to highlight something that was already lurking in our analysis of phenomenal concepts.

A standard, non-rigidified analytic functionalism is antithetical to the conditional analysis. On standard analytic functionalism, the referent of PAIN in counterfactual worlds does not depend on its referent in the actual world. Only intrinsic features of the world—namely what plays the appropriate functional roles in that world—are relevant to which states are the conscious states. So the sort of analysis the conditional analysis gives of the conceivability of zombies is not open to the standard, non-rigidified analytic functionalist.

But suppose that you take phenomenal concepts to have the following a priori analysis: *The state (perhaps physical, perhaps nonphysical) that* actually *plays so-and-so functional role.*[48] It follows from this that if the actual world is merely physical and has physical states that play the right functional roles, then phenomenal concepts refer rigidly to physical states. And if the actual world has non-physical states that play the right functional roles, then phenomenal concepts refer rigidly to these states. It won't be

[48] This could be thought of as a condensed way of putting the following conditional:
1. If the actual world contains nonphysical states *that play the relevant functional roles*, then phenomenal concepts refer rigidly to these nonphysical states.
2. If the actual world is merely physical *and has physical states that play the right functional roles*, then phenomenal concepts refer rigidly to these physical states.

conceivable that the actual world is a zombie world, but it will be conceivable that a zombie world is possible. This is just the result that the conditional analysis was aiming at. And, as we have seen, this view is able to avoid my charge of circularity. Perhaps the moral of the conditional analysis should be that accepting a rigidified version of analytic functionalism gives you a way of accepting something like the zombie intuition, while avoiding dualism.

Unlike other ways of precisifying the conditional analysis, I take this view to be internally coherent. Nevertheless, I think such a view is unattractive and unmotivated. This view still faces the Chalmers/Alter objection that it fails to account for the intuition that zombie worlds are directly conceivable. And it's not clear that proponents of the conditional analysis will find such a view attractive. Braddon-Mitchell explicitly states as part of his motivation for proposing the conditional analysis that functionalism is a substantial truth and "not directly analytic." He disparages what he calls "procrustean analytical functionalism" (a non-rigidified functionalism), and similarly rejects a rigidified version on which "qualia are whatever *neural state types* are in fact actually causally responsible for our discriminative powers" (Braddon-Mitchell 2003, 113). Part of the reason for this is the "implausibility" of such a priori analyses. And part of the reason for rejecting these views is that they arbitrarily privilege some intuitions (namely the intuition that qualia play certain causal roles) over other intuitions (such as the zombie intuition).

The view I've suggested on behalf of the conditional analysis inherits all of the implausibility of analytic functionalism. But those who take the conditional analysis to

respect our zombie intuitions may not find it guilty of the same sort of privileging of intuitions, as rigidified analytic functionalism has the same story to tell about the conceivability of zombies. (Though I'll argue below that there are other intuitions that this analysis flies in the face of.)

More worrying than whether particular adherents of the conditional analysis would find this view palatable is the worry that *rigidified* analytic functionalism is a highly unattractive view. Suppose that state P plays the pain role in the actual world. Consider a counterfactual world in which state Q plays the pain role. According to rigidified analytic functionalism, in this counterfactual world, state P—not state Q—will be pain. This runs contrary to all functionalist intuitions. Readers may have noticed that the proponent of the conditional analysis isn't actually committed to such a view. As I noted in footnote 33, Braddon-Mitchell suggests an alternative version of the conditional analysis on which:

1. If the actual world contains nonphysical states (that play the relevant functional roles), then our phenomenal concepts refer rigidly to these states.
2. If the actual world is merely physical (and has physical states that play the right functional roles), then in the actual world our phenomenal concepts refer to the physical states that actually play the right functional roles, *and in each counterfactual world, our phenomenal concepts refer to whatever physical states play the right functional role in that world.*

On this proposal, phenomenal concepts refer rigidly only if the actual world has nonphysical states. But we still face the worry raised above in worlds with nonphysical states. Consider the following scenario: The world has nonphysical states. Nonphysical state P plays the pain role in the actual world. Now consider a counterfactual world in

which nonphysical state Q plays the pain role. In this counterfactual world, nonphysical state P—not nonphysical state Q—will be pain. Isn't this scenario just as bizarre as it would be if the states in question were physical states? After all, even in nonphysical worlds considered as actual, this view would have us pick out pains by reference to the functional roles they play, *not* by reference to their intrinsic natures. (This was, again, because we couldn't specify in the antecedent what it took to have the relevant intrinsic nature, without employing the phenomenal concepts we aimed to explain.)

To overcome this objection, perhaps we should understand phenomenal concepts as follows. It's analytic that:

1. If the actual world contains nonphysical states (that play the relevant functional roles), then in the actual world our phenomenal concepts refer to the nonphysical states that actually play the right functional roles, *and in each counterfactual world, our phenomenal concepts refer to whatever nonphysical states play the right functional role in that world.*
2. If the actual world is merely physical (and has physical states that play the right functional roles), then in the actual world our phenomenal concepts refer to the physical states that actually play the right functional roles, *and in each counterfactual world, our phenomenal concepts refer to whatever physical states play the right functional role in that world.*

But why stop there? If what's relevant to a thing's being pain is its playing the relevant functional role, then why separate the cases where the functional role is actually played by something physical from the cases where it's played by something nonphysical (and visa versa)? Suppose that physical state P plays the pain role in the actual world. Consider a counterfactual world in which nonphysical state Q plays the pain role. In this counterfactual world, physical state P—not nonphysical state Q—will be pain. But this

again runs contrary to functionalist intuitions. And the proponent of the conditional analysis can't collapse the two clauses above into a single clause, for that just yields analytic functionalism which we have seen is antithetical to their view.

This resulting view is ad hoc and radically at odds with the intuitions that typically motivate functionalism. It seems the only motivation there could be for fleshing out the analysis in this way is a last gasp for the coherence of the conditional analysis.

Finally, accepting this sort of analytic functionalism seems contrary even to the motivations for the conditional analysis, insofar as part of this motivation for the conditional analysis was the idea that we can't rule out the possibility of dualism a priori. Our dualist intuitions don't stem from the idea that a nonphysical state could be what's playing some relevant functional role. Our dualist intuitions are grounded in the idea that there could be nonphysical states that—*by their intrinsic natures*—feel certain ways. This is why the zombie argument doesn't seem compelling against dualism: If dualism held phenomenal states to be nonphysical states that were simply important insofar as they played certain functional roles, we would be able to conceive of a world where we had these states playing the relevant roles, but we lacked the what-it's-likes. (At least, we'd find we had just as strong an intuition that we could conceive of such a world as we have that we can conceive of the traditional zombie world.) But we don't seem to be able to do this. So, contrary to the motivations for the conditional analysis, this analytic functionalism version of the conditional analysis doesn't seem to respect our dualist intuitions.[49] This is especially problematic for Braddon-Mitchell, who proposed the

[49] Thanks to Richard Yetter Chappell for this insight.

conditional analysis in part out of a desire not to arbitrarily privilege certain intuitions over others.

6.5 Conclusion

The conditional analysis of phenomenal concepts purports to give physicalists a way of understanding phenomenal concepts that will allow them to (1) accept the zombie intuition, (2) accept that conceivability is generally a good guide to possibility, and yet (3) reject the conclusion that zombies are metaphysically possible. It does this by positing that whether phenomenal concepts refer to physical or nonphysical states depends on what the *actual* world is like.

I've argued that the only coherent way to hold such a view is by adopting a species of analytic functionalism, on which it's a priori that phenomenal concepts refer to the state (perhaps physical, perhaps nonphysical) that actually plays so-and-so functional role. Any other way of precisifying the conditional analysis is plagued by vicious circularity. While the rigidified analytic functionalism is coherent, it's highly unpalatable. It faces the Chalmers/Alter objection that it fails to capture the true zombie intuition (direct conceivability), giving us only a mock-zombie intuition (conceivable possibility). I've offered arguments to bolster this objection, showing just how minimal the intuition Chalmers and Alter have argued really is. It—as the original conditional analysis—makes it the case that whether there's consciousness in a world or not doesn't depend on features intrinsic to that world. It runs contrary to the intuitions that commonly motivate functionalism, and when fully spelled out seems unmotivated—failing even at its task of

taking dualism seriously as an epistemic possibility. Finally, this view inherits all the "implausibility" of analytic functionalism—something that Braddon-Mitchell explicitly rejected in his motivation for proposing the conditional analysis. So it's not clear whether even defenders of the analysis would be happy with this view. But happy or not, this is the only coherent precisification of the conditional analysis.

Chapter 7

A Psychological Gap: Against the Phenomenal Concept Strategy

We now turn to the second way of using phenomenal concepts to support type-B physicalism: the phenomenal concept strategy (PCS). According to PCS, phenomenal concepts are so radically different from physical concepts that even an ideally rational agent with complete physical information could not work out the phenomenal-physical entailments a priori. This strategy aims not to close the epistemic gap, but to give a physicalistically respectable story of why the gap arises.

In this chapter, I argue that no version of the phenomenal concept strategy can succeed. For any theory of phenomenal concepts that can be proposed to support PCS, we can construct an agent to serve as a counterexample to the claim that an ideal agent would find an epistemic gap.

I begin in section 7.1 with a precisification of the phenomenal concept strategy, due originally to Daniel Stoljar (2005) and Martine Nida-Rumelin (2004). In section 7.2, I introduce my argument against the phenomenal concept strategy, beginning by focusing on the version of PCS that I take to be the strongest – that based on the constitutional theory. I show that we can design an agent who (1) has full general physical knowledge, (2) has phenomenal concepts that work just as the constitutional theory claims, and yet (3) is wired such that she would be in a position to immediately work out the phenomenal-physical truths. I argue that—assuming the truth of physicalism—this creature's immediate phenomenal-physical judgments amount to a priori knowledge. It follows from this that there is not an *ideal* epistemic gap between the phenomenal truths

an the physical truths. Thus, the special nature of phenomenal concepts does not entail an ideal epistemic gap. The phenomenal concept strategy fails. Section 7.3 defends my argument against four potential objections. In section 7.4, I extend my argument, showing that the basic structure of my argument also creates trouble for version of PCS based on indexical and direct reference theories.

Finally, in section 7.5, I wrap up the chapter and assess its implications. While the special nature of phenomenal concepts cannot support the existence of an *ideal* epistemic gap, my arguments point us towards a highly compelling way to be a type-A physicalist. We clearly have a very different psychology from the imaginary agent I'll describe. I suggest that it is this psychological difference of ours—we might say "this rational defect" of ours—that explains why *we* find an epistemic and explanatory gap. I argue that this gets the proponent of PCS what they really were after all along: It gives us a way to be a physicalist that respects the intuitions that *Mary* would be fooled if her captors presented her with a blue banana, that *we* would find zombies conceivable no matter how much physical information we possessed, that *we* will never be able to bridge the explanatory gap. Accepting this version of physicalism requires abandoning the idea that these gaps are *ideal*, but plausibly our intuitions about zombie cases and the like were never about what creatures with radically different psychologies from our own would be able to conceive of; rather they were about creatures psychologically like us. Understanding the way we think about consciousness reveals not an *ideal* gap, but a *psychological gap*. And this, I suggest, is just what the physicalist should want.

7.1 A Clarification of the Phenomenal Concept Strategy

Before I consider whether the phenomenal concept strategy succeeds, I want to clarify the form of the strategy. It's tempting to think that the reason Mary learns something new when she leaves her black and white room is simply that she acquires a new concept: Having never had a phenomenally red experience before, captive Mary *lacked the concept* of phenomenal redness. It's for this reason that she couldn't work out the phenomenal-physical truths prior to her release: She didn't have the concepts necessary for even formulating the phenomenal-physical truths. While this is a superficially attractive view, it does not answer the dualist's arguments, and it should not be mistaken for the phenomenal concept strategy. (The following argument is due to Daniel Stoljar, drawing on Martine Nida-Rumelin.[50])

To see that this response is not sufficient to answer the knowledge argument, imagine that before her release, Mary is shown a red piece of paper, but is not told what color it is. Based on this, Mary would be able to form a concept of phenomenal redness. Mary now has all the relevant concepts. But intuitively, Mary would not be able to work out a priori that the experience she was having was one of phenomenal redness. No amount of physical background information seems capable of revealing to Mary what sensation it is that she's currently having. Further, it should be obvious that this line of thought does nothing to answer the conceivability argument or arguments from an explanatory gap: Despite our possession of the relevant phenomenal concepts, we find zombie worlds conceivable and phenomenal-physical truths arbitrary, and it doesn't seem

[50] Note that Stoljar coined the phrase 'phenomenal concept strategy', and used it to refer to this sort of view. But those who've since adopted the name to apply to their views don't have this in mind.

that this is simply due to a lack of physical information. (Stoljar, 17-18)

For these reasons, this is not the strategy that advocates of PCS endorse. They rather argue that *because of the special nature of phenomenal concepts, even someone with all the relevant physical information and all the relevant phenomenal concepts would not be able to work out the phenomenal-physical truths a priori.* This basic position is fleshed out with a theory of the nature of phenomenal concepts, which is supposed to support the claim that an a priori derivation would be impossible. This is an attractive and sophisticated position, which deserves careful attention.

I'll now turn my focus to this position, arguing that it too fails to support type-B physicalism. I'll begin with the version of the phenomenal concept strategy based on the constitutional theory, as this theory is the most plausible. But I think that even if the indexical and direct reference theories (argued against in chapter 2) can be defended, they are not capable of supporting an ideal epistemic gap. In section 7.4 I'll show how my argument can be extended to versions of PCS based on these theories.

7.2 Constitutional Theories and the Phenomenal Concept Strategy

According to the constitutional theory, defended in chapters 3 and 4, phenomenal concepts are partially constituted by the experiences they refer to[51]. Phenomenal experiences are not merely the referents of our phenomenal concept, but *constituents* of them, which provide the concepts their cognitive significance. According to the most popular version of this theory, phenomenal concepts have a quotational structure "the

[51] This is at least true of the most fundamental sorts of phenomenal concepts – what I called "full phenomenal concepts" in chapter 3.

experience __" where the blank is filled in by a token of the relevant type of experience. Because the phenomenal experience is a constituent of the concept, the phenomenal concept carries with it the phenomenology of the experience. Knowledge involving these experiences is supposed to give us unmediated insight into the "essence" of the phenomenal experience. (Balog 2012a, 16).

Since phenomenal concepts don't refer via a priori associated descriptions – but rather via instances of the appropriate type of experience – there is alleged to be no possibility for a priori derivations of phenomenal-physical truths. This generates the epistemic gap. And because phenomenal concepts carry their phenomenology with them (whereas physical concepts don't) physical concepts seem to leave out the "what it's likes". This is allegedly what makes an *explanation* of such truths seems impossible, generating the explanatory gap. While I take the constitutional theory to offer a highly compelling account of phenomenal concepts, which can give us much of what we want out of such a theory, I will argue that it cannot vindicate type-B physicalism.

7.2.1 A Priori Derivation Without Associated Descriptions

I want to put pressure on the idea that a priori derivations require that one of the terms involved employs an a priori associated description. I'll offer an alternative picture that plausibly facilitates an a priori derivation, even on the constitutional model. The basic idea is this: Just because phenomenal concepts don't refer via associated descriptions doesn't mean that there's nothing to use in an a priori derivation. The concepts involve instances of the phenomenal experiences themselves. If these

experiences simply are physical processes, then it should be possible for the physical system they're processes of to be constructed in such a way that the processes can be reliably matched up to information stored *describing* the processes. And that would yield an a priori derivation of the phenomenal from the physical.

According to the phenomenal concept strategy, there's an ideal epistemic gap between the physical truths (P) and the phenomenal truths (Q). That is, it's *ideally conceivable* that there be a world physically just like our own, but phenomenally completely different: It's ideally conceivable that P & ~Q. What does it take for some proposition Z to be ideally conceivable? One way of getting at ideal conceivability is by way of what an *ideally rational* agent could conceive of. Since this yields the further tricky question of what it takes to be an ideally rational agent, I'll stick with a more minimal characterization: Z is *ideally conceivable* iff it's not knowable a priori that ~Z.[52]

How should we understand the notion of a priority employed here? Chalmers characterizes a priority as follows: "If I cannot know that P independent of experience, but another less limited being could do so, then it is a priori that P." (Chalmers 2002b) What I take to be relevant for Z being a priori is that *some creature* could come to know, independently of experience that Z. It follows that Z is ideally conceivable iff it's not the case that some creature could know that ~Z independently of experience.

Proponents of the phenomenal concept strategy claim that it's ideally conceivable that P & ~Q: that it's not the case that some creature could know that ~(P & ~Q) independently of experience. I will show that contrary to their claims *it is the case that* some creature could know that ~(P & ~Q) independently of experience. Some possible

[52] This is roughly Chalmers's definition of ideal negative conceivability. (Chalmers 2002b)

agent could work out that P = Q a priori.

The structure of this section is as follows: In order to assess PCS on grounds that its defenders will accept, I will assume the truth of physicalism. I will remain neutral about whether phenomenal experiences should be identified with functional states or the particular neural states that realize them. However, in the interest of simplicity, I will initially present my analogy in functionalist terms—e.g. suggesting that captive Mary's concept of phenomenal redness describes a *functional state* she could enter.[53]

I will design, in low-level, sub-personal terms, an agent who (1) has complete general physical knowledge, (2) has concepts that meet the constitutional theory's requirements for being phenomenal concepts, and yet (3) is wired such that she could immediately determine the phenomenal-physical truths. I'll argue that—assuming the truth of physicalism—this creature's immediate phenomenal-physical judgments amount to a priori knowledge. I'll use a computer to get clear on the sort of internal architecture that an agent would require in order to be able to directly match up the phenomenal truths with the physical truths. We can then imagine a conscious agent who has the same sort of internal architecture as the computer.

If I can construct such an agent, it will show that the constitutional theory of phenomenal concepts does not entail an *ideal* epistemic gap as purported by the phenomenal concept strategy. A less limited agent with the requisite concepts would not find a gap. It follows that *we* find these gaps in part because of our own psychology.

[53] In Objection 1, I show that the argument creates just as much trouble for nonfunctionalist physicalists as for functionalists.

7.2.2 An Alternative Psychology

Let's begin by imagining a computer analogue of a person with full physical knowledge.[54] We might imagine concepts as files, storing information about the referents. Since the computer has complete general physical knowledge, it has files describing every functional state the computer could be in.[55] Call these files *descriptive files*. These files are analogous to e.g. captive Mary's concept of phenomenal redness, which—by hypothesis—picks out phenomenal redness in terms of the functional state of agents experiencing it.

The computer also has analogues of phenomenal concepts, which I'll call *constituted files*. Just as my phenomenal concepts contain tokens of the relevant types of phenomenal experiences (which, by hypothesis, just are certain physical-functional states), the computer's constituted files will contain tokens of the relevant type of computer-states: functional states realized by some particular silicon-based computational states. For now, I'll simply describe these as files containing instances of functional states the computer might be in (though, of course, these functional states will be realized in some particular way).

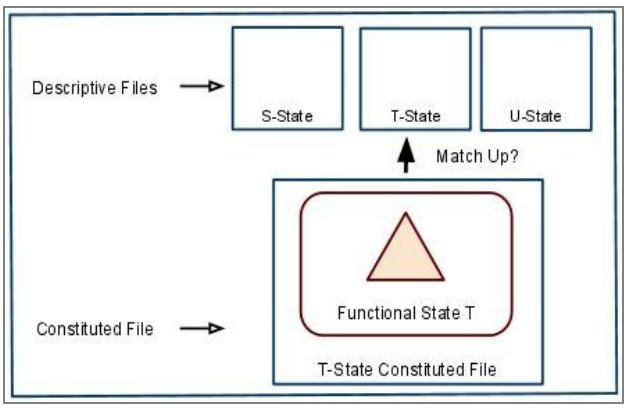
To have these "concepts", the computer must have an analogue of the "the experience" operator, which enables its own functional states to operate as the contents of

---

[54] This structure was inspired by Perry's proposal for the psychology of indexical knowledge described in *Knowledge, Possibility, and Consciousness*. I discuss Perry's version of the phenomenal concept strategy and argue that the same problems apply to it in Section VI.

[55] This isn't quite right. If the computer had files describing *every* functional state it could be in, we'd get explosion. It would need to have files describing not only the state-A, but also the state of describing the state-A, and the state of describing the state of describing the state-A, and so on. But all we really need for this argument to work is that the computer contains descriptions of the first-order states that are equivalent to phenomenal states in conscious creatures. It needn't store descriptions of what it's doing when it recognizes a conscious state or recognizes that it's recognizing a conscious state, and so on. Thanks to Corey Maley for pointing out this complication.

files. Imagine that the computer is in functional state-T, and that the T-state functions as the contents of a file, creating a constituted file of the T-state. As I've designed it, there is also a descriptive file of state-T, which stores a complete description of the T-state.



We now have a computer analogue of the mind of captive Mary when she is presented with a red piece of paper for the first time. Both Mary and the computer have complete general physical knowledge.[56] Mary has a physical concept of phenomenal redness (a concept describing the functional state that observers are in when having red experiences). Our computer has a descriptive file of state-T (a file describing the functional state-T). Mary has a phenomenal concept of redness (a concept that is constituted by her experience of phenomenal redness—which, by hypothesis, is a functional state). The computer has a constituted file of state-T (a file that is constituted by an instance of the T-state).

If the structure of phenomenal concepts is sufficient to generate an epistemic gap between the phenomenal and the physical, then the computer will—like Mary—not be able to match up the constituted file of the T-state with the descriptive file of the T-state.

[56] More carefully, the computer stores complete descriptions of all the physical truths. I don't mean to imply that the computer has intentionality.

Must this be the case? I will argue that it is not: The computer could be constructed such that there's a reliable low-level mechanism by which the constituted file is matched to the corresponding descriptive file. In what follows, I'll flesh out what this means and how it is possible.

If the description of the T-state stored in the descriptive file were an instance of the state itself, then surely a computer could recognize that the contents of the descriptive file and the constituted file were the same. But the descriptive file is not supposed to store an instance of the T-state, but a *description* of it. On this suggestion, the computer is simply matching up two instances of the same type of constituted file. The match-up is trivial.

One might think *of course* a computer could be set up to match the constituted file to the descriptive file: It could simply be programmed to recognize and merge the two files when they were present. But this will not help us. This is analogous to insisting that we might have an innate disposition to match up our phenomenal experiences to certain functional states: an innate disposition to recognize our phenomenal experiences as instances of a certain functional type. While we could have such an innate disposition, if there is no principled connection between the concepts we are matching up—simply a brute tendency (programming) that causes us to do so—the resulting beliefs will not amounted to *a priori knowledge*. We could just as well be "programmed" to mis-match phenomenal and functional concepts. The problem is that the match-up process is not *reliable*.

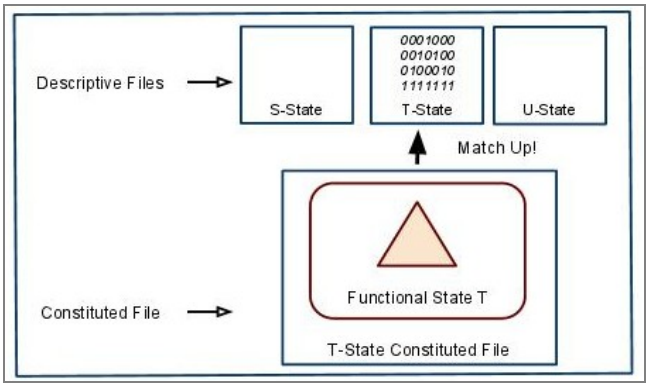
We need to show is that our brains could be wired to match up the phenomenal

concepts and the physical *because* they corefer. I think that this is possible. If I draw a triangle, there are several ways you might describe what I've drawn. You might describe it by drawing another triangle. A computer could doubtless match these drawings up. However what we have here is two instances of a triangle, not an instance and a description. You could describe the triangle by writing the word 'triangle'. The computer could also match the drawing and the word. But because the connection between the two is completely arbitrary, the human analogue doesn't qualify as a priori knowledge. However, you could also describe the triangle by using propositions to describe a two-dimensional array, where each pixel was white or black. A computer could doubtless be programmed to recognize this as a triangle. The relationship between the triangle and this description is not arbitrary, as in the 'triangle' case, nor is it trivial, as in the image case. This example shows that there could be a principled relationship between a description of a state and the state itself, based on which recognition could occur.

Return to our computer analogy. Suppose that there is a principled relationship like this between the information *describing* the T-state (stored in the descriptive file), and the T-state itself. The computer might be set up such that it could recognize the constituted file and the descriptive T-state file as coreferential, as it was programmed to detect that they described the same state. The fact that certain programming is required for this match-up is not problematic. Programming is required to compute that 1+1=2, or any other thing the computer might process. Similarly, any conscious inferences I make are the result of low-level computations done by my brain. The crucial point here is that the programming is not matching the files up arbitrarily, but because of a detected

commonality between them.

The figure below gives an extremely simple, but concrete illustration of the computer architecture I'm imagining. An aspect of the functional state that the computer is in constitutes the content of a computer file (T-State Constituted File).[57] The computer also has descriptive files that store descriptions of various states the computer could be in. One of these files describes the T-state, and because of the principled relationship between its content and the content of the T-State Constituted File, the computer recognizes the two as coreferring.



I've suggested that we can construct a computer analogue of Mary, who would be able to match up constituted files with their corresponding descriptive files, in such a way that a conscious creature with this psychological structure would be able to work out the phenomenal-physical truths a priori. To do this, I've sketched an example of a creature who can match up constituted files with their corresponding descriptive files in a non-

[57] One might object that the way a computer represents a triangle is through representation of a two dimensional array. But note that I'm imagining the functional state *to be* a triangle—not merely to be a state that represents a triangle. You might imagine that somehow, the computer circuits are literally activated in a triangle pattern. If you don't like this, you could imagine that the functional state is a representation of a two dimensional array, and the descriptive file stores instructions for drawing a triangle. The goal is simply to have the constituted file containing some functional state, and the descriptive file describing that state in a *different way*, but a way that bears a natural connection to the functional state itself.

trivial, non-arbitrary way. Before moving on, I want to highlight the features of this match-up process that seem crucial for my further claim that a conscious creature with this psychology could work out the phenomenal-physical truths a priori.

Important features of the alternative psychology: (1) The match-up process I've described is *reliable*. If either the constituted file or the descriptive file had been structurally different, the match-up would not have occurred. (2) The match-up process is *robustly reliable across counterfactual worlds*. Suppose we were to make a small change to the constituted file. The constituted file will be matched to the descriptive file if and only if there's a corresponding change in the descriptive file, such that the match-up is reliable. (3) There's an *analog relationship*[58] between the constituted and descriptive files. While this may not be crucial to the match-up yielding a priori knowledge, it does help to make vivid how the match-up process could be robustly reliable, and so to make the example I've sketched compelling.

If functionalists are right, to have a phenomenal experience is to be in a certain sort of functional state. If I had complete general physical knowledge, and my brain worked in an analogous way to the computer above, this would be sufficient for direct recognition that the experience I'm currently having is the experience described by my descriptive concept. In the following subsection, I will argue that this direct recognition yields a priori knowledge. For now, let's take this for granted and see what follows.

---

[58] The notion of analog *representation* is a familiar one. When I represent my age using 26 candles, this is an analog representation: If my age were to change (say I have a birthday), the number of candles must correspondingly change in a linear fashion. This relationship is symmetrical: If my age were a representation of the number of candles on a birthday cake, when we change the number of candles on the cake, my age would need to correspondingly change. This sort of symmetrical analog relationship holds between the constituted T-State file and the descriptive T-State file. I have simply called it an analog *relationship* because it is not important in my example that either file be taken as a *representation* of the other one.

We are clearly not set up in this way. Even a human with full general physical knowledge would not be able to determine the truth of phenomenal-physical identities without investigating the world. These identities would always seem arbitrary. But if what I have said is right, it is a *contingent feature of our psychology*—a quirk, perhaps a rational defect—that we can't have recognition directly like this.

If this is right, we still have an explanation of why *we* are unable to derive the phenomenal from the physical. But the explanation has to do with *us*, not with the a posteriori status of phenomenal-physical entailments.

For a claim to be knowable a priori it has to be the case that some agent could come to know the truth of the claim independently of experience. And if what I have said is right, we have a sketch of just such a creature. The fact that some creature (us) could possess all of the relevant concepts and yet not be in a position to determine the claim's truth is unimportant. If there could be a conscious creature whose brain functioned in the way I've described, then some conscious creature could immediately derive the phenomenal-physical identities (P=Q), much as we can immediately determine that 1+1=2. It follows that there is no *ideal* epistemic gap between the physical (P) and the phenomenal (Q): There's a possible agent that could work out a priori that ~(P & ~Q).

But recall that the goal of the phenomenal concept strategy was to accept the ideal conceivability of P & ~Q, while blocking the alleged metaphysical implications. If what I've argued is correct, the constitutional theory of phenomenal concepts doesn't provide a way to do this; rather, it provides an explanation of why *we believe* there to be such a gap, though there really is not one. So this may offer a plausible physicalist reply to anti-

physicalist arguments, it does not work as a version of PCS. (Essentially, what we have is a psychological story that supports type-A physicalism.)

7.2.3 A Priori Knowledge

Thus far, I have argued that there could be a creature whose psychology enabled her to immediately and directly "match up" any experience she'd had with the functional description of that experience. I've claimed that the conscious belief that such an agent immediately forms as the result of the described low-level processing yields a priori knowledge, and explored what the implications are if this is true. Now I want to step back and defend my claim that these immediate phenomenal-physical judgments qualify as a priori knowledge. (Recall that proponents of the phenomenal concept strategy can happily accept that an agent with a radically different psychology could work out the phenomenal-physical truths a posteriori—even *we* can work out these truths a posteriori! What they deny, and what I must prove in order to refute their argument, is that a creature could work out the phenomenal-physical truths *a priori*.) I'll begin by arguing that the resulting belief is knowledge. Next I'll argue that this knowledge is a priori, rather than introspective a posteriori knowledge.

Is it Knowledge?

To simplify the discussion, let's imagine a particular case of a creature with the psychology I've described. Imagine an agent who—like Mary—knows all the general physical truths, but has never had an experience of phenomenal redness. This agent

differs from Mary simply in that she has a psychological structure analogous to the computer I have described. Call her Mary*. When Mary* sees red for the first time, she not only forms a new concept (of what this new experience is like), she also forms an immediate judgment that what this phenomenal concept picks out is the same kind of thing her functional concept picked out. The question we're now faced with is: Does this judgment—that thus-and-so experience is of thus-and-so functional kind—amount to knowledge?

Let's start by dismissing one reason you might think it wasn't knowledge. I have argued that a conscious creature, with phenomenal concepts like those described by the constitutional view, could have a physical constitution that enabled it to recognize immediately—without investigation of the world—that phenomenal experiences that it was having were identical to certain physical processes. But this recognition stemmed from low-level subconscious matching of the current states of the creature with information describing physical states. I have not argued that any creature could construct a conscious inference of the phenomenal facts from the physical facts.[59] (And I don't think this is possible.)

One might object that knowledge can only be come to through inferences that are, at least in principle, consciously accessible. If this were the case, what I have described would not be a case where a creature *knows* that the phenomenal experience they're

[59] Inferences made at the very low level I've described may not even be consciously perceived as inferences by the creature. For example, imagine a skilled rower. An oar in water that would look bent to a layperson *looks straight* to the rower. But the processing that causes the rower to see the oar as straight is not consciously accessible to the rower. All that the rower is conscious of is the result of the low-level inference: the oar looks straight. Similarly, we might expect creatures with the psychology I've described to simply be aware of the result of the inference: a direct awareness that the experience they're now having is described by a certain physical concept.

having is a certain physical-functional state, but simply one where they have an immediate intuition to that effect.

But it's not clear to me that all things we take ourselves to know are such that we can form a conscious inference that they are true. I seem to *know* that pain is bad. I don't seem to be able to give a conscious justification for the claim. If this is to count as knowledge, then there doesn't seem to be any problem with the possibility of knowing things that we can't consciously justify.

Why should the physicalist think that such a creature would *know* the phenomenal-physical truths? First, the physicalist is committed to holding that this creature's phenomenal-physical beliefs are *true*. The creature matches up descriptions of functional states with phenomenal experiences that—according to physicalism—are those very functional states.

Furthermore, this creature doesn't just randomly happen to form true phenomenal-physical beliefs. She forms these beliefs *because they are true*. The low-level matching process doesn't arbitrarily match up the descriptive file with the constituted file—it matches them up because the functional state that fills the constituted file is the very type of state described by the descriptive file. It might seem suspicious that I haven't argued that any creature could derive such a match *based on the "phenomenal feel" of being in the functional state*. (Surely the phenomenology should play a role in coming to know that P=Q.) But I don't think the physicalist can justify this move. According to the functionalist, pain essentially feels a certain way. Pain is (also) essentially a certain functional state. While my creature is not working out entailments

based on the phenomenology of the experiences, she is nonetheless working it out based on an *essential* feature of these experiences. There seems no basis for privileging one sort of essential feature and insisting that derivations must be based on that feature to count as knowledge. So my agent's phenomenal-physical beliefs seem to be epistemically justified. It's hard to see how the physicalist could deny that these creatures have phenomenal-physical *knowledge* surpassing our own.

Is the Knowledge A Priori?

More problematic is the question of whether this knowledge is a priori. Type-B physicalists have no difficulty accepting that some creatures can come to know phenomenal-physical truths. (They take us to be such creatures!) What they cannot accept is that such truths can be known *a priori*. And it seems plausible that, contrary to my claims, what I've described is really a kind of introspective a posteriori knowledge.[60]

Consider Descartes's cogito. "I think, therefore I am" can be known through introspection alone, but nevertheless is plausibly only knowable a posteriori.[61] Similarly, my knowledge that *I'm currently thinking about a priori knowledge* is something I know through introspection alone, but also seems a posteriori. How can I hold that Mary*'s knowledge of the phenomenal-physical truths is a priori if I accept that these other cases of introspective knowledge aren't? What is the relevant difference between these two cases?

In cases of introspective knowledge that we intuitively feel to be a posteriori, the

[60] Thanks to Richard Yetter Chappell for suggesting this objection and for helpful discussion about responding to it.

[61] Those who take the cogito to be knowable a priori will doubtless happily grant that Mary*'s knowledge is also a priori.

knowledge is based on contingent truths that are known through introspection. Descartes's cogito takes *as a premise* the contingent truth "I think"; it's his observance (through introspection) of this contingent fact that grounds his knowledge. When I think that *I'm currently thinking about a priori knowledge*, my knowledge is grounded in the contingent fact that I'm currently thinking about a priori knowledge. The contingent fact that I'm currently thinking this *justifies* my belief. But Mary*'s subconscious inference of the phenomenal-physical truths is very different. Though it's true that in order to make such judgments, Mary* must have had certain phenomenal experiences (which she's only contingently had), these experiences play a very different role in her beliefs. Mary* must have had a red experience in order to have the relevant concepts to think about the phenomenal-physical truths. But Mary*'s judgment isn't grounded in the contingent *fact that she's had a phenomenal red experience*. Her judgment is rather grounded in the *necessary* (assuming physicalism is true) fact that such-and-such phenomenal state is the functional state that it is. Mary*'s having had the relevant experience plays an enabling role for her to be able to work out the phenomenal-physical truths, but—unlike clear cases of a posteriori introspective knowledge—it does not play a justificatory role.

There's a second related argument that might be made against the a priori status of Mary*'s knowledge. One might argue that the "match up" process that I've described involves a kind of *inspection* of the agent's concepts, and that knowledge come to through inspection (whether inspection of something external to the agent or something internal) always yields a posteriori knowledge.[62] On this line of thought, there's an important difference between a priori conceptual analysis (which involves *employing*

[62] Thanks to David Chalmers for suggesting this line of argument.

concepts) and a posteriori *inspection* of concepts.

It's hard to get a grasp on the difference between knowledge come to through inspecting concepts and knowledge come to through employing them. (When I conclude e.g. that a bachelor is a marriage-eligible man, does this involve *inspection* of my concept or *employment* of it? How can we determine the difference?) However, there is one clear difference between inspecting a concept and employing it: When I inspect my concept of redness, I can thereby come to know that *I have a red concept*. By contrast, I can't conclude this simply by employing my red concept. It is clear that one can only know that they have a red concept a posteriori. This might seem to be reason to think that there is a principled difference between inspection and employment of concepts, and that inspection only yields a posteriori knowledge.

While I agree that one can only know whether they have a red concept a posteriori, I think that this can be explained in the very same way as the cases of introspective a posteriori knowledge discussed above. My knowledge that I have a red concept is a posteriori because it is grounded in the *contingent* fact that I have a red concept. But once again, the phenomenal-physical knowledge that I've described doesn't fit this model. Mary*'s knowledge of phenomenal-physical truths *presupposes* that she has the concepts necessary for thinking the relevant thoughts. (It's a precondition far employing a concept that you possess it.) But her knowledge isn't grounded in the presupposition that she has the said concept. Rather, it's grounded in the necessary truth that the relevant phenomenal and physical concepts corefer.

Since it's so difficult to get a grasp on the difference between employing and

inspecting concepts, we might characterize the difference by relying on the clear-cut cases where there is a difference between the two processes. If we follow this suggestion, it's natural to think that the marker of judgments made by *inspecting* concepts is that the contingent fact of the concept's existence plays a role in the judgments. But if this is how we analyze the inspection/employment distinction, Mary*'s judgment would not involve inspection.

So we haven't seen a reason for thinking that Mary*'s knowledge is a posteriori. It's possible that one could (1) give a radically different account of the difference between employing and inspecting concepts, on which inspecting concepts always yielded a posteriori knowledge, and (2) show that the knowledge I've described involves concept inspection on this account. But it's difficult to see how such an argument could go, and there is some intuitive reason to think that such an account could not succeed.

Suppose I'm faced with a red block, an orange block, and a blue block. Rather than attending to the blocks, I decide to reflect on the *phenomenal experiences* that I'm having as I look at them. I think to myself: "That experience (red) is more similar to that experience (orange) than to that experience (blue)." If Mary*'s "match up" of the phenomenal and the physical counts as *inspection*, then it's hard to see how this scenario could fail to count as inspection: Surely what I'm doing when I make this judgment is detecting a commonality between my red and orange experiences, much as Mary* detected a commonality between the functional and phenomenal concepts. Further, intuitively what I'm doing is *inspecting* my experiences to determine how similar they are. But intuitively, once we hold fixed that I'm having red, blue, and orange experiences,

my further knowledge that red experiences are more like orange ones than like blue ones is *a priori*. So if this involves inspection of experiences, then not all things we know through inspection are a posteriori.

7.3 Objections

I'll now move on to consider four objections to the argument I've sketched. (1) I consider an objection that my argument creates trouble only for functionalists who wish to endorse the phenomenal concept strategy. I show that my argument creates trouble for nonfunctionalists as well as functionalists. (2) I consider an argument that the agent I've described is employing different concepts from ours, and hence says nothing about what's conceivable using our concepts. (3) I considers Balog's defense of PCS against Chalmers's "master argument", and whether a similar response might cause trouble for my argument. (4) I consider whether the agent I've described would find an *explanatory* gap.

7.3.1 Only a Problem for Functionalists[63]

As stated, my argument is directed at functionalists running PCS. I've taken captive Mary's concept of phenomenal redness to be a concept describing a certain functional state that humans can be in, and the analogous files in my computer to be files describing functional states of computers. I've argued that a computer could be designed such that there was a principled relationship between certain functional states it could be in and the files describing these functional states. This would enable the computer to

[63] Thanks to Jack Woods for pointing this out to me and pressing me on the following objection.

directly "match up" files containing instances of these states with files describing them. But what could I say to the physicalist who argues that being in a certain mental state isn't a matter of being in a certain *functional* state, but rather is a matter of being in a particular *neural state*?

The critic might argue as follows. Experiencing phenomenal redness isn't a matter of being in Functional State R, rather it's a matter of being in the particular neural state that actually fulfills this functional role in humans: having your R-fibers firing. While it's plausible that an organism could be designed to detect patterns between the functional state that it's in and natural descriptions of this functional state, it's not plausible that an organism could be set up to—at a low level—make a principled match between a particular brain activity (R-fibers) firing and a concept describing this particular activity. So your argument only shows that functionalists cannot rely on the phenomenal concept strategy.

In responding, it will be helpful to clarify what I have described as the "matching" of files. The strategy I've used involves building an alternative psychology from the ground up. The "matching up" between descriptive files and constituted files is not a *conscious process*, and as such doesn't involve any sort of conceptualization of either the files or the information resulting from the match-up. Rather, this "match up" should be understood as a low-level detection of similarity: a recognition of a commonality between the state embedded in the constituted file and the state described by the descriptive file. The conscious result of this low-level detection of similarity is an immediate awareness that the two files/concepts are of the *same type*. Note also that the

states that are embedded in constituted files (or in humans' phenomenal concepts) will be members of several types: a functional type, as well a silicon-computational (or neural) type that realizes the functional type.

Now return to Mary*, our alternative psychology version of Mary. I've argued that when Mary* is shown a red piece of paper for the first time, she could determine a priori that she was having an experience of Functional Type R. Further, Mary* will, in virtue of having full general physical knowledge, know that Functional State R is realized in humans by R-fibers firing. Given her complete general physical knowledge, Mary* is in a position to work out—a priori—not only that she is in Functional State R, but also that she is in Neural State R (the state of R-fibers firing).

So regardless of whether phenomenal states are identified with functional states or neural states, if my original argument is correct, a creature with full general physical information and the appropriate psychology could work out all the phenomenal facts a priori. One might wonder whether there wasn't still something that Mary* would be missing. Even if she could work out a priori that she is in Functional State R and Neural State R, could she work out a priori that her current experience was *identical to* Functional State R, or that it was *identical to* Neural State R? It's plausible to think that there's no substantial question here: once Mary* knows that her phenomenal experience is of the particular functional and neural types that it is, she knows all there is to know. But suppose the defender of PCS holds that there is a substantial question here. For this to pose any challenge to my argument, this question would have to be not only substantial but also be something that could only be settled a posteriori. But surely if there's a deep

question as to whether the neural state or the functional state is identical to the phenomenal state, it's a question that can't be solved through empirical investigation. (How could empirical investigation possibly reveal this to us?)

7.3.2 Constitutional Files Aren't Analogous to Phenomenal Concepts[64]

Return to Mary and Mary*. When Mary is shown a red piece of paper for the first time, and forms a new concept of what it's like to see this color, she can't work out a priori that the experience she's now having is an experience that her physical red concept applies to. When Mary* is shown the red piece of paper, she immediately recognizes the experience she's having as a red experience. Call the phenomenal concept Mary forms *red*, and call the phenomenal concept Mary* forms *red**. One might now object that—though *red* and *red** refer to the same experience and have the same cognitive significance insofar as both contain an instance of phenomenal redness—*red* and *red** are different concepts. *Red** facilitates different a priori derivations from *red*, showing that something about their cognitive significance differs. While it can be determined a priori that *red** and *physical red* co-refer, it can't be determined a priori that *red* and *physical red* co-refer.

For the sake of argument, let's grant that *red* and *red** are different concepts. Would this help the phenomenal concept strategy? *Red** is clearly a phenomenal concept: It has the very structure proponents of the constitutional view put forward as the marker of phenomenal concepts. It not only refers to phenomenal experiences, but also

[64] Thanks to Angela Mendelovici for pressing me on this objection, and to Richard Yetter Chappell for his helpful discussion of responses to it.

characterizes them in terms of what it's like to have them. It carries with it the very phenomenology that it refers to. In fact—assuming that the functional state it employs is a phenomenal state (which it could just as well be as in our phenomenal concepts)—these concepts seem to give us everything we could want from a theory of phenomenal concepts.[65] If entailments between physical concepts of phenomenal redness and *red** can be known a priori, then it is not the case that there's something special about phenomenal concepts that prevents the phenomenal-physical entailments from being known a priori. Rather there is something special about a particular class of phenomenal concepts. What class? The only way to describe it is to say: "There is a class of phenomenal concepts that is unique insofar as these concepts can't—simply because of the psychology of the organism they're possessed by—be a priori determined to co-refer with physical concepts."

While I have said that Mary can't determine the phenomenal-physical entailments a priori because a quirk (irrationality) in human psychology prevents us from matching up the phenomenal concepts with the physical, the objector will say: Mary can't determine the phenomenal-physical entailments a priori because her phenomenal concept *red* is a special kind of phenomenal concept—a phenomenal concept embedded within a psychology that can't match it up with physical concepts. But these simply seem like two

[65] Balog, one of the leading defenders of PCS, offers a list of desiderata that any adequate account of phenomenal concepts will satisfy. Apart from the desiderata specifying that phenomenal concepts should generate an epistemic/explanatory gap, red* meets every criterion. Relying on Balog's desiderata, the only way one could deny that red* is a phenomenal concept is by insisting that phenomenal concepts *just by definition* generate an epistemic/explanatory gap—an extremely unappealing route to take. While the defender of PCS could add additional criterion, Balog's list already looks quite thorough. It's difficult to see what independently motivated criterion one could add.

For Balog's desiderata for a theory of phenomenal concepts, see: "Acquaintance and the Mind-Body Problem." (forthcoming-a)

ways of describing the same thing.

What of the conceivability argument? PCS was a strategy for defending type-B physicalism, which aims to accept the first premise of the argument while denying its conclusion: holding that zombies,while not possible, are *ideally conceivable*. I argued that the constitutional version of PCS cannot achieve this aim: If our ability to conceive of zombies results from a quirk of our psychology, and a creature who lacked this psychological quirk would be able to a priori determine these entailments, then zombies are not *ideally* conceivable. But one might argue that on the present line of argument, creatures with this alternative psychology are simply using different concepts. Using these alternative phenomenal concepts, zombies cannot be conceived of. Using human phenomenal concepts, zombies can be (ideally) conceived of.

This objection highlights the question of what it takes to be an ideal agent. Consider two different methods of locating the ideal agent: the first takes a human being, *holds fixed certain basic features of their psychology*, and improves them. Call this agent the *Human Ideal.* The second way of idealizing starts from scratch and locates the absolutely most rational structure for an agent's psychology. Call this agent the *Absolute Ideal*. The Human Ideal will be able to conceive of zombies; the Absolute Ideal will not be able to conceive of zombies.

Is *ideal conceivability* a matter of conceivability to the Human Ideal or to the Absolute Ideal? You can *say* whatever you like here, but if you care about ideal conceivability because of its potential to tell you something about *the world*, it's clear that it's the Absolute Ideal that matters. When the Human Ideal and the Absolute Ideal

diverge in what they find conceivable, the Human Ideal doesn't tell us anything substantial about *the world* beyond what the Absolute Ideal tells us. Rather, it tells us something about the peculiar structure of the Human mind. The whole reason philosophers care about what's ideally conceivable is because it at least seems to tell us something about what's possible. But no one would think that the Human Ideal (with its constraints of basic Human psychology) would have any implications for metaphysical possibility. Hence Absolute Ideal conceivability is clearly the notion that dualists have in mind in the conceivability argument. So acceptance of its first premise requires accepting that an Absolute Ideal agent could conceive of zombies.[66]

Another way of making this point is simply to look at our definition of ideal conceivability. Z is ideally conceivable iff it's not the case that some agent could know that ~Z independently of experience. This makes no reference to the conceptual resources the agent uses. If any agent could work out (using whatever conceptual resources) that ~Z, independently of experience, then Z is not ideally conceivable.

Regardless of whether you choose to describe the concepts I've sketched as "the same" as our phenomenal concepts or not, the upshot is the same: There is nothing special about phenomenal concepts that prevents phenomenal-physical entailments from being known a priori. If physicalism is true, the explanation of this failure must lie in our own psychology, not in anything distinctive about concepts that carry the phenomenology of their referent with them. In the most powerful sense of "ideal conceivability"—the sense that dualists use, and the only sense of philosophical importance—zombies are not

[66] It's also obvious that the Absolute Ideal conceivability of zombies is what type-B physicalists intend to accept. If they merely wanted to grant the claim that a Human Ideal agent could conceive of zombies, there would be a simple route to denying metaphysical conclusions: Human Ideal conceivability, grounded in our fickle human psychology, clearly tells us nothing about metaphysical possibility.

ideally conceivable.

7.3.3 Quasi-Phenomenal v. Phenomenal Concepts

Another line of objection might emerge from an argument Balog gives in response to Chalmers's "master argument" against PCS. Chalmers begins by asking whether it's conceivable that a creature be physically just like us (P), but lack phenomenal concepts (C). That is: Is P & ~C conceivable? He then posses the following dilemma:

> (1) If P&~C is conceivable, then C is not physically explicable.
> (2) If P&~C is not conceivable, then C cannot explain our epistemic situation.
> _____
> (3) Either C is not physically explicable, or C cannot explain our epistemic situation.
> [(4) So PCS fails.] (Chalmers 2007: 174)

His defense of premises (1) and (2) are not relevant for our purposes here. What is relevant is, first, Chalmers's suggestion for avoiding the first horn of his dilemma, and second, the structure of Balog's response—which the physicalist might hope could be appropriated to give a response to my argument.

One might think that by accepting the conceivability of zombies, the type-B physicalist rules out the possibility of accepting that P & ~C is not conceivable, forcing the first horn of the dilemma on which C is not physically explicable: Zombies lack phenomenal experiences; if phenomenal experiences are necessary for C, then zombies necessarily lack C. Chalmers takes it to be obvious that the physicalist wants to avoid the first horn of the dilemma. To do this, he suggests they cast C in topic-neutral terms: giving the conditions for possessing "quasi-phenomenal" concepts—concepts structurally identical to phenomenal concepts, but without the presupposition that they involve or

refer to phenomenal experiences. Zombies doubtless do have quasi-phenomenal concepts, so it may not be conceivable that P & ~C.

In her response to Chalmers, Balog rejects the appeal to the appeal to quasi-phenomenal concepts as antithetical to PCS:

> The physicalist explanation of the substantial grasp [of a phenomenal property via a phenomenal concept] crucially involves the fact that there is something it is like to have an instance of [the phenomenal property]. This means that the constitutional account couldn't be cast in physical or quasi-phenomenal terms and still explain our epistemic situation. … Neither a neurophysiological, nor a mere "architectural" description of phenomenal concepts – e.g. that they are constituted by instances of the referent – can explain the substantial manner in which we refer the phenomenal properties. (Balog, 2012b: 14)

But isn't what I have done simply to give a "mere 'architectural' description of phenomenal concepts"? And if so, won't my argument face the same worries as Chalmers's master argument?

I think that considering Balog's response to Chalmers's dilemma will help to clarify her objection to the use of quasi-phenomenal concepts in such a way that makes clear why her argument seems to provide a successful rebuttal to Chalmers, but leaves my argument untouched.

Chalmers's dilemma is posed as a reductio on PCS. As such, it must begin by assuming PCS. According to PCS, there are two different ways of conceptualizing C: using phenomenal terms, or using physical terms. Balog argues that because of this, the type-B physicalist can accept *both* horns of the dilemma: If we conceptualize C using phenomenal terms, P & ~C is conceivable; C is not physically explicable; but C can explain E. If we conceptualize C using physical terms, P & ~C is not conceivable; C cannot explain our epistemic situation; but C is physically explicable. But, as a

physicalist, Balog takes these two ways of conceptualizing C to simply be different ways of describing the same thing. So C is physically explicable (under one conceptualization) and can explain our epistemic situation (under another). The proponent of PCS is in no worse a position than she started in.

Now consider the structure of the dialectic: Chalmers asks "Is P & ~C conceivable?" Balog responds with a argument about what *we do* when *we* try to conceive of P & ~C. Because *we* have two ways of conceptualizing C, we find two ways of answering the question. The appeal to quasi-phenomenal concepts must be rejected, as it blocks off one of these conceptualizations. This response is based on the idea that what we're doing is trying to construct "from the outside-in" a creature where P & ~C: a creature *we, in our outside perspective,* would recognize as satisfying P & ~C.

But what I have tried to do in this paper is to build up an agent from the inside-out, constructing their psychology "from the ground-up". I have not asked "Do we find a gap?" but rather "Would an agent that functioned in this way at the low-level find a gap?" *Our* conceptualization of the creature is irrelevant. While I have described this agent in physical terms we could now stipulate that the agent has phenomenal experiences (identical to the physical properties filling the constitutional file) and still get the same result. Let the computer be a conscious super-computer. The same results still hold.

7.3.4 An Explanatory Gap Remains

One might think that the creature I have designed would still find an *explanatory* gap. I haven't said anything to decrease the mystery that fills phenomenal-physical

identities. Wouldn't this creature still wonder *why* the phenomenal was identical to the physical? Wouldn't they still feel that these identities were somehow *arbitrary*?

I find it at least plausible that such a creature wouldn't find an explanatory gap. Consider another example of something I can know but can't give a conscious rationalization for: Pain is bad. This truth comes to me immediately, through introspection. Yet I can't explain *why* it is true. Despite the fact that I cannot explain why pain is bad, I don't find myself up at night, wracking my brains for an answer to the question "Why is pain bad?" It doesn't seem at all arbitrary to me.[67] I suspect that anything we could determine to be true directly, through introspection, would be similar to this. It would be bizarre indeed if a creature could immediately intuit that A=B, but was plagued by the feeling that it was arbitrary that A=B.

7.4 Extending the Argument

If my argument is successful against the version of the phenomenal concept strategy based on the constitutional theory, I think it can be extended to pose trouble for versions of PCS based on other leading theories of phenomenal concepts. As I argued in chapter 2, I think these theories fail to offer what we want out of a theory of phenomenal concepts, and hence cannot support PCS for independent reasons. My aim will be to show that *even if* these theories could give us what we want out of a theory of phenomenal concepts, they cannot support the phenomenal concept strategy.

The basic idea is that any successful theory of phenomenal concepts will require

[67] Recall that the presence of an explanatory gap isn't simply a matter of not knowing the answer to the question "Why is A identical to B?" We're in that state with all sorts of identities. (Why is Cicero identical to Tully?) The explanatory gap arises when we feel we *need to explain* why they hold *and we can't*.

agents to store information in their heads (e.g. that guides their ability to recognize the relevant sorts of experience, that guides their ability to imagine the experiences, that they can mentally point to and refer to) If the agent is just a physical system, and this information is stored in some physical state, then—by the same lights as the original argument in section 7.3—it should be possible for the physical system of the agent to match this state up with concepts describing the state. [68]

7.4.1 Indexical Phenomenal Concept Strategy

Recall that Indexical theories take phenomenal concepts to function as internal demonstratives. According to these theories, when I have an experience of phenomenal redness, to bring the experience under a phenomenal concept is to form an internal demonstrative pointing to *that experience (whatever it is)*. Reference to the experience is rigid, picking out the same type of experience in each counterfactual world (just as reference to "that (pointing at a cup)" is rigid). But, unlike on the constitutional theory, on the indexical theory, phenomenal concepts are not individuated by the experiences they refer to. Had I *actually* been demonstrating a different experience, the concept would have referred to a different experience (just as "that" would refer to a bowl had I actually been pointing at a bowl).

Since Perry (2001) offers the most developed version of the indexical theory, I will take his account as a target. Perry offers a thorough account of what is going on

[68] The indexical, direct reference, and constiutional theories are not exhaustive of all the possible theories of phenomenal concepts one could hold. Lycan, for instance, holds that phenomenal concepts are semantically primitive and essentially perspectival mental representations. But I hope that my extension of my argument to the indexical and direct reference theories makes clear how the basic strategy would be extended.

psychologically when we use indexicals. He proposes that our psychology has a three level structure. (Perry also uses an analogy to a computer, with concepts imagined as computer files.) The top level of our belief architecture is "detached" files, containing information gained from books, lectures, calendars, and the like. These files don't tell us anything about our current perceptual experiences and, by themselves, won't help us to navigate the world. The bottom level of the architecture contains "perceptual buffers" which temporarily store information gained from perception until it can be attached to a concept in the top level. On this picture, when I see a cup and form an indexical concept *that* (cup), the experiences I'm having of the cup are stored in my perceptual buffer. Recognizing the cup *as a cup* involves connecting the perceptual buffer to a detached cup notion. This connection requires there to be some commonality between the information stored in the perceptual buffer and that stored in the detached notion.

Similarly, when Mary has a red experience for the first time, the experience comes to fill her perceptual buffer. On Perry's view, this is what it takes for Mary to have a red phenomenal concept. This is not, by itself, sufficient for recognition of the experience as the same experience she'd read about in textbooks. To get that recognition (and the substantial knowledge that comes with it), Perry thinks that Mary needs some basis for linking the red perceptual buffer to her detached red notion. If simply having a red experience in her perceptual buffer were sufficient for linking the buffer and the detached notion, then Mary could work out the phenomenal-physical identity a priori. Perry, rightly, thinks that Mary could not do this. He suggests that for recognition, Mary would need to e.g. (1) have a red detached notion that included the knowledge that

tomatoes cause red visual experiences (2) see a red tomato, and (3) based on the recognition that a tomato was causing her experience, link up the perceptual buffer with the detached red notion.

But if my earlier arguments are correct, we can easily apply them to Perry's indexical psychology as well. Notice the direct parallels between the psychology Perry sketches and the psychology I sketched in section 7.3. Both Perry's agent and the constitutional agent contain "files" storing descriptive information about phenomenal experiences. On both pictures, there's a file that contains instances of phenomenal experiences themselves—Perry calls this file the 'perceptual buffer'; I called it a 'constituted file'. The only difference between the two pictures is in how this latter type of file/concept is individuated. On the constitutional picture, it would be individuated by reference to the experience contained in the file; on Perry's it isn't. Because of this, the theories differ in what they take phenomenal concepts to refer to in counter-actual worlds. But this is merely a difference in philosophical interpretation, not a substantial difference in the agents' psychologies. Nothing on Perry's picture differs from the constitutional picture that would affect whether the agents could directly match up the files containing instances of the experiences with those describing them.[69]

Perry thinks that match up between the red detached notion and the perceptual buffer requires, e.g. recognition that the perceptual buffer includes an experience of a tomato, together with knowledge that tomatoes are red. But if my argument in section 7.3 is correct, there could be some creature for whom the similarity between the detached

[69] In fact, I originally developed this objection as a response to Perry, and only later transformed it into a reply to the constitutional theorist.

notion and the perceptual buffer was sufficient. So while Perry is surely right that *Mary* could not directly match up the experience in her perceptual buffer to her detached notion, this is again simply due to Mary's psychology. The indexical theory doesn't support the existence of an ideal gap, and so cannot support type-B physicalism.

7.4.2 Direct Reference Phenomenal Concept Strategy

Direct reference theories take phenomenal concepts to be recognitional concepts that refer to phenomenal experiences directly, without employing any modes of presentation at all. To possess a concept of phenomenal redness is simply to possess a brute capacity to recognize instances of the phenomenal experience as "another one of those".

Because phenomenal concepts on this view have no modes of presentation, there is supposed to be no possibility of an a priori derivation of the phenomenal from the physical. Phenomenal concepts effectively function as pointers directed at (and reliably caused by) a certain type of experience. These concepts are supposed to be so "thin" that there's no possibility of using them to derive phenomenal-physical truths; there's just nothing there to use in such a derivation. This generates the epistemic gap.[70]

Tye formerly endorsed a more sophisticated version of the direct reference theory. He thinks that we must distinguish the question "What makes concept C a concept of e.g. phenomenal redness?" from the question "What makes concept C a *phenomenal* concept?" Just as on the simple direct reference theory, Tye thinks that concept C is a

[70] I argue elsewhere that this simple direct reference theory is untenable, on grounds that it cannot account for either the substantialness of our phenomenal knowledge or the explanatory gap.

concept of phenomenal redness, just in case C is reliably triggered by (and because of) phenomenal redness. But this isn't sufficient to characterize what it takes to be a *phenomenal* concept. To be a phenomenal concept, C must be laid down in memory as the result of a phenomenal experience and must tend to trigger appropriate mental images in response to certain mental activities.

But I think both the simple and the complex direct reference theories fail to support the phenomenal concept strategy, for reasons analogous to those I put forward against the constitutional theory. Let's start by considering Tye's theory.

In order for C to tend to trigger the appropriate sort of phenomenal experiences, the agent's brain must contain information guiding the construction of such experiences. Their brain must be able to automatically "unpack" this information, to generate an experience of the relevant type. (They don't have to consult a translation manual to work out how to unpack the information into an experience of the right sort.) But if their brain contains information that can guide the reconstruction of the relevant type of experience, and if (as I argue in section 7.3 having the relevant type of experience is enough for some creature to be capable of immediately determining the phenomenal-physical truths, then the agent's brain must store information such that some creature could (based on it) immediately work out these truths.

Our brains are storing the very same information as on the constitutional view. The only difference is that on the constitutional theory, the information is stored in a format that is identical to a conscious experience; on Tye's theory, the information is stored in such a way that it can (directly) be unpacked into a conscious experience. But

this makes no difference to the argument I gave in section 7.3 It again seems that we find a gap between the phenomenal and the physical, in part, because of our psychology.

The simplistic direct reference theory is similarly vulnerable. Something must be guiding my recognitional abilities, enabling me to reliably discriminate experiences of a certain type. While this might not be consciously accessible, the argument I gave against the constitutional PCS did not rely on the information used in the derivation being consciously accessible. By the same lights as the original argument, we might construct a creature whose brain stored information sufficient to recognize instances of a certain type of experience, who was "wired" such that this information could be immediately linked up to the (differently formatted) information describing phenomenal experiences in physical terms.

7.5 Conclusion

My main aim has been to argue that no version of the phenomenal concept strategy can succeed. Since it offers an independently plausible theory of phenomenal concepts, I've developed my attack against the constitutional version of the phenomenal concept strategy. If the physicalist is right, phenomenal experiences are identical to physical processes. My argument has developed the idea that, if phenomenal concepts involve instances of the phenomenal experiences themselves, it should be possible for the physical system they're processes of to be constructed in such a way that the processes can be matched up to information stored *describing* the processes.

I've described the psychology of such a creature, from the ground up, and have

argued that the conscious result of such a low-level "match up" amounts to a priori knowledge that the descriptive and phenomenal concepts corefer: A creature with the psychology I've described could, given complete physical information, work out the phenomenal-physical truths a priori. I've further argued that if there is a possible creature who could work out the phenomenal-physical truths a priori, then there is not an *ideal epistemic gap*. I've argued that this argument applies to nonfunctionalists as well as functionalists, and have considered several objections, concluding that none of them pose a serious challenge to my argument.

Finally, I've shown that my argument can be extended to the other versions of the phenomenal concept strategy. Even if these theories could offer successful accounts of phenomenal concepts, this would not support type-B physicalism. Once we accept that possessing a phenomenal concept requires one's brain to store an instance of the relevant experience, or information sufficient to guide the creation of such an experience, or information sufficient to guide recognition of such experiences, we open up the possibility of a creature whose brain could directly match this information (or functional state) up to concepts describing the state. By the same lights as before, the result would count as a priori knowledge. The physicalist cannot accept that there's an ideal conceptual gap between the phenomenal and the physical.

But while the phenomenal concept strategy does not support the existence of an *ideal* epistemic gap, my argument is not at odds with what I take to be the true goal of the type-B physicalist: to offer a way of being a physicalist that respects our "dualist" intuitions. It just seems crazy to deny that captive Mary would be fooled if presented with

a blue banana, to insist that if only we had more physical information, we would find zombies inconceivable, to insist that if only we knew more, we would cease to find phenomenal-physical identities arbitrary.

If what I have argued is correct, the physicalist must accept that an *ideal* agent wouldn't be fooled by the blue banana or find zombies conceivable. But, plausibly, our intuitions were not shaped by consideration of some agent with a completely different psychology from our own. (We surely can't even imagine what it's like to be Mary*.) Rather, our intuitions are shaped by our own psychological restrictions. And the physicalist who accepts my arguments against the phenomenal concept strategy can respect *these* intuitions. Any creature psychologically like us will find an epistemic and explanatory gap, which they cannot bridge, no matter how much physical information they might gain. This means that captive Mary would be fooled if presented with a blue banana, that no amount of physical information would render zombies inconceivable to us, that no amount of physical information could close the explanatory gap. While physicalists must give up on an ideal gap between the phenomenal and the physical, they don't have to give up on an unbridgeable *human* gap. And, plausibly, this is what we were after all along.

I think this is an extremely important result. If my argument is correct, a dominant strategy for defending type-B physicalism is untenable. And furthermore, my argument has the upshot that the physicalist needn't to deny the tight connection between ideal conceivability and possibility—a debate that has pervaded literature about type-B physicalism—in order to respect our intuitions. The phenomenal concept strategy, which

has dominated recent literature on the mind-body problem, must be abandoned. But as a result, we find a way of bringing type-A physicalism and type-B physicalism closer together, reaping the benefits of each without the costs.

So how we think about consciousness can help to defend a plausible, coherent version of physicalism. In the next chapter we turn to explore the impact this has on the debate between physicalism and dualism. Clearly insofar as a compelling physicalist picture of the mind is possible, this reduces the need to be a dualist. But does this new type-A strategy pose any more principled problems to the dualist? One challenge presents itself specifically for the epiphenomenalist: Since epiphenomenalism holds that phenomenology is causally inert, it gives the same explanations of psychological phenomena as physicalists give. But if that's right, then the epiphenomenalist should grant this same explanation of why we have our dualist intuitions that I've sketched in this chapter. This seems to undermine the whole (intuition-based) rational for being a dualist.

In the final chapter, I'll address this and other objections to epiphenomenalism. I'll show that none of these objections poses a significant challenge to epiphenomenalism. In fact, the right story of how we think about consciousness can help to defend a plausible and coherent version of dualism as well as physicalism.

Chapter 8

A Dialectical Stalemate: Making Progress on the Mind-Body Problem

Most of the literature responding to anti-physicalist arguments has centered around the question of whether there's an *ideal* epistemic gap. This question has been thought of as another way of getting at the question of whether we should accommodate or simply reject the epistemic intuitions that ground anti-physicalist arguments. Because of this physicalists have gone in one of two ways: (1) insist that there is no ideal epistemic gap and reject the dualist epistemic intuitions (the standard type-A model), or (2) accommodate the dualist epistemic intuitions by holding that there's an ideal epistemic gap (as type-B physicalists do). The first strategy faces the obvious challenge that it just seems crazy to deny these firmly held epistemic intuitions—we should at very least be able to explain them away! The second strategy faces the challenge of defending an ideal epistemic gap, without collapsing into dualism—a challenge that many have argued cannot be met.

If what I argued in the last chapter is right, the physicalist can walk a middle ground between these two positions, retaining the virtues of each without the problems. The physicalist can accommodate our dualist epistemic intuitions by holding *not* that there's an ideal epistemic gap between the phenomenal and the physical, but that there's an unbridgeable *psychological* gap. Thus they can both respect our most firmly held intuitions and avoid messy debates about the relationship between conceivability and possibility. And, unlike the Phenomenal Concept Strategy, this view is consistent.

So there is a version of physicalism that is not only consistent, but that can

accommodate our dualist epistemic intuitions (by explaining why we have them). Phenomenal concepts help to reveal this version of physicalism. In this chapter, I'll consider the flip side: Is there a consistent and compelling way of being a dualist? And what, if anything, do phenomenal concepts have to say on this matter?

I'll argue that epiphenomenalism is both consistent with our basic assumptions about the mind, and that it is far less at odds with our intuitions than it initially seems. Phenomenal concepts will play a role in this defense. If I am right, phenomenal concepts help to reveal both a compelling way of being a physicalist and a compelling way of being a dualist. The question that remains is how we can adjudicate between these two positions.[71] The second half of this chapter addresses this question and the question of how to go forward making progress on the mind-body problem.

8.1 A Defense of Epiphenomenalism

Many physicalists have offered the following simple argument against dualism: Physics is causally closed. That is, every physical event can be explained entirely by physical causes. Hence if dualism is true—if mental states like *being in pain* are non-physical—then mental states can't causally effect behavior. So if dualism is true, epiphenomenalism is true. But epiphenomenalism is crazy! So dualism isn't true.

In this section, I will argue against this argument, on grounds that epiphenomenalism is not crazy: The epiphenomenalist has compelling responses to each

[71] Note that these are not the only coherent positions. I take it that there is widespread agreement that panpsychism is internally coherent (if a bit odd seeming), and that dualists as well as some physicalists are willing to consider it a version of their own type of theory. I largely omit it from discussion here because my aim is to show that the heated back and forth between dualists and physicalists has not been conducted in a way that can be fruitful. Since panpsychism has been largely left out of this debate (and since I take it that there is little doubt over its internal coherence) I will set it aside.

of the objections that have been raised against them. I'll start by running through some bad objections to epiphenomenalism, which should have been put to rest long ago by Jackson's 'Epiphenomenal Qualia', but which are (in the word of Monty Python) "not quite dead yet". Next I'll move on to some more compelling objections. The first objection concerns the peculiar sort of "psycho-physical bridging laws" that epiphenomenalists appeal to in explaining the existence of phenomenal experiences. The second objection is the problem of how we can refer to, have knowledge of, and remember our phenomenal experiences. Chalmers calls these the 'Paradoxes of Phenomenal Judgment.' Third, I'll consider an objection grounded on the amazing coincidence that our phenomenal experiences integrate with our physical states as well as they do. Fourth, I'll consider the problem that (according to the epiphenomenalist) all our table-thumbing insistences that *there's something it's like to feel pain* are not grounded in the fact that there's something that it's like to feel pain. Zombies would argue just as passionately that physicalism couldn't possibly be true. This, to many, seems bizarre. From my responses to these standard objections against epiphenomenalism will emerge a new problem for the epiphenomenalist: a problem questioning the motivation for being an epiphenomenalist. This final problem brings us to the question of how to move forward on the mind-body debate: What, if anything, can adjudicate between physicalism and dualism? How can we move forward from here?

We now move on to the bad objections against epiphenomenalism.

8.1.1 Objection from Introspection

Doesn't it seem introspectively obvious that when I touch a hot stove, I pull my hand away *because of* the pain that I feel? Doesn't it seem obvious that when I scream at a drunk driver who hit my car, I do so *because of* the anger that I feel raging inside of me? The epiphenomenalist has to deny this introspective datum.

The response to this should be obvious to anyone who has read Hume. "When we look around us at external objects, and think about the operation of causes, we are never able to discover any power or necessary connection, any quality that ties the effect to the cause and makes it an infallible consequence of it. All we find is that the one event does in fact follow the other" (Hume 1748, section 7). We needn't take Hume as arguing that there is no such thing as causation. The relevant point is simply that the experience of causation supervenes on constant conjunction. As a result, experience of causation is indistinguishable from the experience of constant conjunction.

You see billiard ball White moving toward billiard ball Red. White hits Red, and the two balls touch for a moment. Red rolls away. You (correctly) experience this as White *causing* Red to move. But suppose we strapped your head up to a television. We show you a cartoon depiction of the same event. You see White rolling towards Red, the two balls touch, Red rolls away. You will then have an experience as though the white ball *moved* towards the red ball, and then *caused* the red ball to move. But there is no motion inside of the television, no transfer of energy, no causation. There are simply pixels that vary in color, giving the illusion of motion and of causation. Your experience in the two cases is indistinguishable, though in the first case, there was causation, and in

the second there was none: merely a succession of events.

The epiphenomenalist will argue that this is what's going on all the time. I touch the hot stove coil. I feel excruciating pain. I pull my hand away. I holler. It seems to me that the pain causes me to pull my hand away. But this "seeming" isn't able to distinguish between the case where the pain really does cause me to pull my hand away, and the case where pain is simply reliably correlated with my pulling away of my hand.[72] Since the experience of causation is indistinguishable from the experience of constant conjunction, the fact that phenomenal experiences *seem* to play causal roles doesn't tell against the epiphenomenalist.

8.1.2 Objection from Knowledge of Other Minds

One might wonder how we can know that other people—our friends, spouses, strangers on the subway—have conscious experiences if epiphenomenalism is true. After all, we only see the physical behavior of other persons: their vocalizations, their movement when the cross walk lights change, their tears when their spouse says "I'm in love with someone else." If epiphenomenalism is true, we never see peoples' conscious experiences. Moreover nothing that we see is the causal result of peoples' conscious experiences. Because of this it might seem mysterious how we can know that Julie is sad or that Max believes it's safe to cross the street.

But even if the behavior we see isn't caused by conscious experiences, we can still extrapolate from our own case. Let's take for granted that *I know that I have*

[72] This particular example is an especially friendly one for the epiphenomenalist, as it is in fact not the case that any brain-state causes me to pull my hand away. My hand pulls away before any signals have been able to reach my brain. But the point will hold even in cases where brain-states do play a causal role.

conscious experiences. (We'll return to the question of how the epiphenomenalist can explain this shortly.) I know that there's a correlation between tears and sadness in myself, and I know (based on internal models) that hearing that my husband loved someone else would cause me to feel overwhelming sadness. I not only know about correlations between my own physical states and my conscious experiences, I also know that I am generally quite an ordinary human being. I am physically very similar to other members of my species. My body and brain function in a very similar manner to other human beings. Further, laws of nature treat like alike. The laws of nature don't vary across regions of space. So the physical laws—and psycho-physical laws, should they exist—make no special exceptions in the way they treat me. Because of this, it is reasonable to assume that when other people cry or hear that their partner loves someone else, they will both have something similar going on in their brain and will—by virtue of the psycho-physical laws—have a similar phenomenal experience.

8.1.3 Objection from Evolution

If phenomenal experiences are epiphenomenal (if they have no causal effects) then how could they have evolved? We know that we're the product of evolution, and evolution selects for beneficial traits—traits that will *cause* the species to thrive. But epiphenomenalists hold that phenomenal experiences have no causal effects on the survival of the species. So how/why would they have evolved?

As Frank Jackson noted in 'Epiphenomenal Qualia', it is simply not the case that every trait that we evolve to have confers evolutionary advantage. What is true is that

those traits that evolution selects for *either* confer an evolutionary advantage *or* are byproducts of traits that confer an evolutionary advantage.[73]

The function of a lightbulb is to produce light. That is why we produce and buy lightbulbs. But the mechanism by which lightbulbs produce light also makes them produce heat. So lightbulbs produce heat, though this is not what we produce them for nor why we buy them. Byproducts similarly arise as the result of evolution. To borrow Jackson's example, a polar bear has warm fur. This trait was selected for as it conveys an evolutionary advantage to the bear. But a byproduct of the fur being warm is that the fur is also heavy. Having a heavy coat of fur is not an evolutionary advantage. It is simply a byproduct of the way warm coats evolved.

Epiphenomenalists must hold that in addition to the laws of physics, that regulate all of the physical world, there are also psycho-physical bridging laws: laws that say things like *when yada-yada complex physical pattern is realized, yada-yada phenomenal property is generated.* So it should come as no surprise then to find phenomenal byproducts of evolution. It's evolutionarily advantageous for our brains to enter the C-fibers firing state that will cause us to remove our hand from the stove. And—because of the way certain laws of nature work, namely the psycho-physical bridging laws—having your C-fibers firing also generates an experience of pain. This is a perfectly scientifically respectable story, just as the story about why polar bears have heavy coats.

Better Objections to Epiphenomenalism: We now turn to some more compelling

[73] It may be that there is also harmless 'noise' that is passed on from generation to generation in the absence of evolutionarily superior alternatives. If an organism can function equally well with feature P or with feature Q, neither will be selected for.

objections against epiphenomenalism. My responses to several will build on my responses to the "bad objections". The correct understanding of phenomenal concepts will play a role in responding to the Paradoxes of phenomenal Judgments, in particular, and derivatively in responding to other objections.

8.1.4 Psycho-Physical Laws

Epiphenomenalists often talk about "psycho-physical bridging laws": Laws that say that when you've got a physical state of the right kind, you get a phenomenal state of a certain kind. Because of these psycho-physical laws, whenever there's the right sort of physical state, a phenomenal state "arises" or "is generated". But who ever heard of a law that says "whenever you've got this kind of thing, some new and entirely different kind of thing is created"? Laws ordinarily tell you about the behavior of *what you've got*. If something new is created, it's only in the sense that what you've got *constitutes* something new. When you combine a bunch of H_2O molecules, the laws of physics are responsible for them behaving in a certain way: a liquidy way. And so the molecules constitute liquid. But the sort of "generation" that the dualist posits is quite different. It's more akin to there being a law that whenever the sun, earth, and moon are all in alignment, *poof* a mongoose appears! Why think that there could be laws that work like this?[74]

If dualism is true, consciousness is unlike anything else we're acquainted with in that it is non-physical. The objector asks why we should think that the laws responsible for interfacing between physical reality and non-physical reality should fit the mold of

[74] Thanks to Galen Strawson for this objection.

the laws responsible for governing physical reality. The presupposition seems to be that because all the laws that we're familiar with work in this way, this provides inductive evidence for thinking that the psycho-physical laws will.

Now if the only animals you've ever encountered are dogs, and they all bark, it might be reasonable to assume that all animals bark. But suppose you discover some feature that dogs all have in common, V, where V is able to account for barking and V is not possessed by other animals. It would no longer be reasonable to conclude that all animals bark, but only the more limited conclusion that all dogs bark. We're in something like this position concerning laws. For there to be a physical-law that said "whenever this happens, that completely distinct thing is generated", there would have to be a violation of the conservation of energy. This gives a reason for thinking that no physical law can work in this way. But the case of phenomenal (non-physical) generation, requires no such violation. This is reason to think that the psycho-physical bridging laws are relevantly different in this respect. Hence the fact that the *physical* laws we're familiar with work in a certain way doesn't provide inductive evidence for thinking that the *physical-nonphysical bridging* laws work in that way.[75]

8.1.5 Paradoxes of Phenomenal Judgment

If phenomenal experiences are epiphenomenal, then how can I *know* about them? How can I *remember* past experiences? How can my phenomenal concepts *refer* to them? Answering these challenges is where phenomenal concepts become useful to the

[75] For those who are not happy with this reply, but who share the dualist's conviction that phenomenal experiences cannot be reduced to the sort of non-experiential features of the world that physics tells us about, Strawsonian panpsychism offers a coherent alternative.

epiphenomenalist. I'll consider each challenge separately.

How can I know about phenomenal experiences?: Chalmers fleshes out this query into an argument as follows. If epiphenomenalism is true, I have a zombie twin in another possible world. As she looks at her navy blue tea cup, the very same things go on inside of her brain as inside of my brain. These brain processes cause the very same sounds "Navy blue looks like *this*" to come out of her mouth as come out of my mouth. So—the argument goes—my zombie twin has the very same phenomenal beliefs as me ("Navy blue looks like this"), formed by the very same mechanism. But my zombie twin's belief is not only false, it's not justified. If her belief is not justified, and my belief was formed by the same mechanism, then my belief can't be justified either. (Chalmers 1996, 192-193)

If the constitutional theory of phenomenal concepts defended in chapters 3 and 4 is correct, the epiphenomenalist has a simple solution. According to the theory, I am directly acquainted with the phenomenal experience of navy blue. This direct acquaintance with the experience of navy blue enables me to think about it directly, pulling it into my thoughts to feature as a constituent of them: I can form a phenomenal concept NAVY BLUE that contains as a constituent the experience of navy blue. This is how we were able to account for the unique epistemic features of phenomenal belief (in chapter 3).

This story—about how I am able to directly latch onto my conscious experiences and pull them into my thoughts—is important for two reasons. First, it is because my conscious experiences are components of my thought *that I have the thoughts I do*. If the

constitutional theory is right, the reason I'm thinking about what it's like to have a navy blue experience is rooted in the fact that a navy blue experience is a component of my thought. If there were a creature physically just like me, but without any such conscious experience embedded in her thoughts, she would not be thinking the same thing as me. Because of this, when my zombie twin and I both direct our gaze at a navy cup, both have the same things going on inside of our brain, and both (as a result) utter the vocalization "Navy blue looks like this", I am entertaining a belief about the phenomenal character of navy blue experiences. My zombie twin is not. So, contrary to the argument against epiphenomenalism, my zombie twin and I *do not* have the same beliefs. The argument fails.

The second way the constitutional theory of phenomenal concepts is relevant is that it explains how I can be justified in my phenomenal beliefs in the absence of a causal connection to the objects of these beliefs. In most ordinary cases, part of the explanation of how we're able to form reliable beliefs is that our beliefs are caused by the objects of our belief. Why is my belief "It's raining outside" justified? One explanation might be that it was formed through a reliable mechanism, where part of what explains the reliability of this judgment is the fact that I rely on a causal relationship connecting myself to the rain outside my window. If I'm not causally connected to my phenomenal experiences, then this cannot be part of the explanation of my phenomenal beliefs' reliability. But this should not worry us. In the case of phenomenal beliefs, we have an even more reliable basis for forming beliefs. When I believe, e.g., "Navy blue looks like this", my thought is itself *constituted by* the objects of my belief. I do not stand "at a

distance" from its truth-makers, as with a causal relationship. Rather, the nature of the object of my belief—the essence of navy blueness—is directly presented to me as the means by which I think about navy blueness. Because the very phenomenal experience that is the object of my thought is being used to do the thinking, the truth of my judgment is even more secure than in the ordinary case. (Chalmers (2003) makes a similar move.)

While I have used phenomenal concepts to give a specific account of where this so-called 'paradox' fails, I think the problem is fundamentally a more general one. The physicalist asks how—if phenomenal experiences are epiphenomenal—we could possibly know about them. This seems mysterious because a physicalistic account of persons and of belief is lurking in the background. The worry is something like this: I am this physical creature; my beliefs are these physical phenomena. The epiphenomenalist tells us that qualia are non-physical, and that they don't have any causal impact on the physical world. It follows that they don't have any impact on me or on my beliefs. So they are irrelevant to me and my beliefs. So how could I possibly come to latch onto them to think about them let alone *know* about them?

The problem is that this objection is assuming a physicalistic conception of belief and of persons. The physicalist may be happy to grant such conceptions, but she should not push these physicalistic conceptions over onto the dualist. The dualist will insist that my conscious experiences are essential to *me* and to my *beliefs*. It is only when you fail to acknowledge this, when you mismatch a physicalistic conception of belief/persons with a dualistic conception of consciousness that you find that consciousness is irrelevant. But the problem here is not the picture of consciousness, but rather the picture of the rest of

the mind that renders consciousness disconnected from it. We might summarize this as the following mantra: *Dualists should be dualists all the way down.*

I think the desire to avoid making qualia irrelevant to our beliefs is reason enough for epiphenomenalists to hold that consciousness is essential for belief. But there are also independent reasons for thinking that consciousness is essential for beliefs (and for intentionality more generally). Intentionality is at least as mysterious as consciousness. As McGinn puts it, "Nothing we know about the brain, including its relations to the world, seems capable of rendering unmysterious the capacity of conscious states to 'encompass' external states of affairs" (McGinn 1991, 39-40). There's reason to think that the mystery of intentionality is closely tied to the mystery of consciousness.

When my zombie twin's brain whirs, and her mouth voices "Navy blue looks like this", it's not just that she's not aware of the content that I express (what navy blue looks like). She's not aware of any content whatsoever. But arguably what makes intentionality so mysterious is that it not only carries information about the way the world is, but that it *presents the information (as information) to someone* (McGinn 1991). A thermometer does not, on its own, have intentionality. It does not "say" *I'm about the world*. It says something about the world only insofar as someone is aware of it *as* representing the world. Likewise my computer does not *itself* mean anything when a box pops up with lines of the form "Firefox needs to restart." *I* take these words to mean something. But my computer does not *itself* take them to mean anything. Similarly, it's plausible that zombies, though they are more complicated than computers or thermometers, do not *themselves take* their utterances to mean anything. To mean something by their

utterances, they would have to be aware of their utterances as meaning something. And zombies are not *aware* of anything.

If zombies do not have beliefs then consciousness is essential for belief and this so-called 'paradox' poses no challenge to the epiphenomenalist. It is not epiphenomenalism that renders qualia irrelevant. It is the combination of a physicalistic conception of belief together with a dualistic conception of qualia that renders qualia irrelevant. That is the position this paradox warns us against.

How can I remember past phenomenal experiences?: It's more puzzling how epiphenomenalists can account for our memory of past phenomenal experiences, since memory clearly requires a causal connection to a past event.

Let's grant that the epiphenomenalist can explain how we have knowledge of present phenomenal experiences. Given this, how can they explain memory? It can't be that a present conscious experience (a mental image of an orange and purple sunset) stands in an appropriate causal relationship to a past conscious experience (a perceptual experience of an orange and purple sunset), since phenomenal experiences are causally inefficacious. The important causal connection required for memory must come in at the physical level.

Here's the picture: I'm gazing out at a beautiful orange and purple sunset. Light hits my retina, information is sent from my eyes to my brain, my brain processes this information, some sort of physical "memory trace" is laid down. Thanks to the information processing going on in my brain and the psycho-physical laws, I have a conscious experience of a beautiful orange and purple sunset. Later on, the physical

"memory trace" is activated again. This activation—together with the psycho-physical laws—causes me to have a new phenomenal experience: a mental image of orange and purple sunset. There is no causal connection between the original phenomenology and the later mental imagery. The causal relationship is between the physical states that give rise to the conscious experiences. But this—together with the fact that we don't ever *observe* one mental state causing another—is enough to satisfy the intuition that there's some importance to causal connections in memory.[76]

How can I refer to phenomenal experiences?: Finally, you might wonder how it is that I'm able to refer to phenomenal experiences. We're typically causally related to things that we're able to think and talk about. But according to epiphenomenalism, we're not causally related to phenomenal experiences. How then can we 'latch onto them' in thought?

It should be obvious by now how the epiphenomenalist will respond. We can first point out that we aren't causally connected to *everything* that we can think and talk about. Chalmers notes that we can think meaningfully about "the largest star in the universe", though we are not causally connected to it (Chalmers 1996, 201). And I can talk meaningfully about, e.g., the number five, though I don't stand in any causal relationship with the number five. One might think that phenomenal experiences were different: While I'm not causally connected to the largest star in the universe, I am causally connected to stars, to the universe, to instantiations of 'largest'. It's arguable that the only way I'm able to think about the largest star in the universe is by way of a description, where the concepts used in the description are things I can think about because of causal

[76] Chalmers offers a similar explanation in *The Conscious Mind* (Chalmers 1996, 200-201)

relationships I stand in. Phenomenal concepts are not like this.

Likewise, you might think that I am able to talk meaningfully about the number five only because I stand in a causal relationship to instantiations of five. Again phenomenal concepts do not work like this: If epiphenomenalism is true, I do not stand in a causal relationship with any instantiations of redness. But it cannot be that reference always requires us to either stand in a causal relationship to the object of reference, or to components of the description that secures reference, or to instantiations of the object of reference. There are many abstract mathematical and logical objects that we can meaningfully talk and think about, but where it's hard to see how we could be causally connected to instantiations of them. Am I causally connected to an instantiation of disjunction? Of modus ponens? It's hard to see what such a claim could amount to.

So the general principle that reference requires a causal connection to the object of reference is false. But phenomenal concepts give us a story that enables us to go beyond this general point to offer an explanation of how it is that we can think and talk about phenomenal experiences: When we think about phenomenal experiences, these experiences literally become components of our thoughts. When Mary thinks *"This* is what red looks like," she is able to think about the phenomenal experience of redness because the experience is *right there*, in her mind, ready for the taking. This is only mysterious for the epiphenomenalist if we think of the mind and thought in general as a *physical* thing, and then ask how this epiphenomenal experience could have anything to do with that physical thing—how the physical mind could "take" the epiphenomenal experience as an object of thought. But as I've argued, this is a bizarre presupposition that

only a physicalist would make. The epiphenomenalist's picture of the mind is not one on which the mind, the self, beliefs, intentionality are things we can explain physically, and then we need to say how to fit these epiphenomenal what-it's-likes in to the physical picture of the rest of the mind. Rather, the natural position for the epiphenomenalist to take is one that is dualist all the way through. We can think about our phenomenal experiences because they are a part of *us*, because they are there for the taking.

8.1.6 Objection from Coincidence

I wrote in response to the Objection from Evolution that epiphenomenalists hold that in addition to the physical laws of nature, there are psycho-physical bridging laws explain the production of phenomenal experiences and the connection between these experiences and brain-states. The objector might press on these laws. Isn't it an amazing coincidence that when we stick our hand on a hot stove or do something else that causes tissue damage, we feel *pain?* If epiphenomenalism is true, the psychophysical laws could just have well have been set up to make us feel ticklishness or to have a sensation of hearing middle C when we underwent tissue damage. It seems like an amazing coincidence that things that are bad for us should *feel bad* and things that are good for us should *feel good*.

While this looks extremely puzzling, it is difficult to see how to turn this observation into an argument to the conclusion that epiphenomenalism is false. Perhaps the argument is something like this: If epiphenomenalism were true, it would be very unlikely that phenomenal experiences would be "in tune with" the physical world.

Phenomenal experiences are "in tune with the physical world". So it's very unlikely that epiphenomenalism is true. But this argument is not valid. If the lottery number is chosen randomly, then it's unlikely that 7892043 will win. 7892043 in fact is the winning number. It does not follow that it's very unlikely that the lottery number was chosen randomly.

The epiphenomenalist will say that the case of the lottery is in fact very similar to the case of the psychophysical laws. There is a sense in which it's remarkable that the laws of the universe are such that phenomenal states and the physical states are "appropriately tuned". But these laws had to work in some way or other. It just so happened to work out like this. Is it amazing that the winning lottery number was 7892043? In some sense it is. If I had to predict before hand whether that would be the winning lottery number, I would have said no. But now, knowing that it is the winning number, am I surprised? Of course not. There had to be some winning number, and it may just as well be that one.

There's a further question that might worry you here of how we can be justified in believing ourselves to be in the matched-up universe, given the (prior) unlikelihood of it and the fact that regardless of what our qualia are like, we will still *claim* that ours are the appropriately tuned qualia. But this is just to return to the paradoxes of phenomenal judgments. And we have seen that despite their lack of causal influence, we can have *knowledge* of our epiphenomenal phenomenal experiences. So we are in much the position as the lottery worker the day after the winning number has been drawn: We know how the phenomenal and the physical match up, but while this was an antecedently

unlikely way for things to turn out, we should not feel any special surprise at the fact our world has turned out this way.

8.1.7 The Zombie Dualist

Dualists spend a lot of time thumping the table, insisting that physicalism can't explain *what it's like* to feel pain, that the rich and brilliant experience of hearing a Rachmaninov piano concerto can't possibly be reduced to electro-chemical impulses. But if epiphenomenalism is true, then the *what-it's-like* of feeling pain or of listening to Rachmaninov has nothing to do with these insistences. My zombie twin would thump the table and insist the same things with the same forcefulness. Moreover, the explanation for why I say and think these things is no different from why my zombie twin does.

The bizarreness of this—of imagining writing and worrying about consciousness, with consciousness itself playing no role—is what famously lead Frank Jackson to abandon epiphenomenalism. As the objection that converted the most prominent defender of epiphenomenalism, this objection deserves to be taken very seriously.

I can see two ways of interpreting this objection: a call for an explanation of our behavior, and a simple statement that there's an intuitive weirdness. First, it might seem mysterious why I would say "Physicalism can't explain the *what-it's-like* of feeling pain," and why I would spend so much time thinking and arguing about what-it's-likes, if not *because there's something that it's like* to feel pain. But the epiphenomenalist is not the only one who needs an explanation of this. The physicalist needs a (purely physical) explanation of why dualists go about thumping the table and insisting these things. Since

epiphenomenalists think that physics is causally closed, they will tell *the exact same story* about *why* we say and think these things.

I think this story will be no less plausible for the epiphenomenalist than for the physicalist. One might object that the physicalist has an advantage here: Given their view of pain, the pain may be playing a causal role in explaining my behavior. But, though the epiphenomenalist recognizes the same neural patterns as playing a causal role in my behavior, she will not accept that *pain* is playing a causal role. However once we've recognized the earlier point that we can't tell the difference between causation and constant conjunction, we see that this does not tell in favor of physicalism. There is no introspective reason to think that pain *causes* anything. All that we require is an explanation of my dualist-table-thumping behavior, and we—like the physicalist—have that. But this opens the door to a new, and I think more powerful, objection.

8.1.8 Have Epiphenomenalists Debunked Themselves?

Anti-physicalist arguments crucially rely on premises that express dualist intuitions: Mary gains propositional knowledge. A zombie world is ideally conceivable. No amount of physical information can ever explain why there's something that it's like to see red or why it's like *this* rather than like something else. Epiphenomenalists must accept an entirely physical explanation of why I assert these things with such force, of why I spend so much time thinking about what-it's-likes, of why I thump the table and refuse to accept physicalism. In fact, epiphenomenalists accept the very same explanation here as physicalists.

But if there's a physicalistic explanation of why we have these intuitions, then why not take this explanation to debunk the intuitions? If the motivation to be a dualist rested on these intuitions and the intuitions can be—as the epiphenomenalist admits—explained in physicalistic terms, then why be a dualist?

The first thing to note here is that, although epiphenomenalists are committed to holding that the explanation for why we say the things we do, write the things we do, attend to the things we do can be given in purely physical terms, this does not mean that when I say "There's something that it's *like* to feel pain" I do not *know* that there's something it's like to feel pain. This is just the lesson of the Paradoxes of Phenomenal Judgment. But this does not answer the present question of why the epiphenomenalist shouldn't take the physical explanation of our intuitions to debunk the claim that phenomenal experiences are by their nature non-physical.

The epiphenomenalist can't respond "But it just seems that phenomenal experiences couldn't possibly be physical." This is just a statement of an intuition. And the intuition has an explanation that seems to debunk it. Giving more statements of intuition will not get us out of the current puzzle. What we need is an independent explanation for why we *should* believe dualism to be true—for why it's a reasonable belief.

Compare this to my belief that 2+2=4. I believe that 2+2=4. And this belief is surely reasonable, even though *the fact that 2+2=4* does not play any causal role in explaining my belief. Neuroscientists of the future will be able to give a complete story in causal terms of why I have this belief. But this explanation—and the fact that *the fact of*

2+2=4 does not play any role in the story—does not debunk our belief that 2+2=4. It does not give us reason to doubt that two plus two *really does* equal four.

One path the epiphenomenalist could take in responding to the debunking challenge is to say something similar. Neuroscientists of the future will be able to explain why I believe that dualism is true and why I believe that 2+2=4. In neither of these explanations do the truth of dualism or arithmetic play a causal role. But the fact that neuroscientists will be able to explain why we have these beliefs entirely in terms of the electro-chemical circuitry of our brains does not cast doubt on the content of either belief.

Another way of responding to this challenge draws more heavily on the response to the Paradoxes of Phenomenal Judgments. I've argued that the most natural way for a dualist to go is to insist that the correct story of belief cannot be given in physical terms. Suppose this is right. Then while there may be a purely physical explanation for why I *say* "There's something that captive Mary is ignorant of," and "No amount of physical information could possibly explain what it's like to see a beautiful sunset," the explanation for why I *believe* these things—why I have these intuitions—is not purely physical. And I not only know that I say these things, I know that I believe them. So, according to the epiphenomenalist, the physicalist's explanation is incomplete, and we can know that it is incomplete. (Although they can explain completely why we *say* that we know this.)

This response, like the first response to the debunking argument, rests on the presupposition that dualism is true. Assuming dualism is true, the correct analysis of belief cannot be given in physical terms, physicalism cannot explain our phenomenal

beliefs (intuitions), and so the physical explanation of our intuitions does not debunk epiphenomenalism. But if dualism is not true, then physicalism gives a complete explanation of why we have the intuitions that we do, and so debunks them. Epiphenomenalism is at least internally coherent. But how we interpret the situation depends on the theoretical commitments we've started with.

8.2 A Return to Mary*

My argument against the phenomenal concept strategy in chapter 7 yielded what I take the be a compelling version of type-A physicalism. On this picture, there is no ideal epistemic gap between the phenomenal truths and the physical truths. But the reason why we cannot work out these truths a priori is not simply because we lack complete physical knowledge. Rather, human psychology is such that *we* find an irremediable gap between the phenomenal and the physical. I suggested that this is a tenable position which, while not living up to the letter of type-B physicalism, matches it in spirit.

As should be clear from our discussion in this chapter, I think that the epiphenomenalist should also adopt this account as an explanation of the physical workings of our phenomenal thoughts and as a (partial) explanation of why we have the mind-body intuitions we have[77]. In what follows, I will spell out the epiphenomenalist's take on Mary*'s psychology. Since the epiphenomenalist cannot hold that Mary*'s phenomenal-physical judgments of the form "phenomenal state Q = physical state P" are knowledge, I will focus on where the epiphenomenalist must hold that Mary*'s beliefs go

[77] A *partial* explanation because intuitions presumably require beliefs. If intentionality requires consciousness, and consciousness is not physically explicable, then while being physically like I am will (assuming I am conscious and have beliefs at all) explain why I have the beliefs that I do, it will not explain what it takes for me to have beliefs at all.

wrong. After addressing this question, I'll step back to look at the broader debate between dualism and physicalism in light of what I've argued.

Nothing I've said in describing Mary*'s psychology hinged on the truth of physicalism, so the epiphenomenalist has to accept that Mary* will directly match up the conscious state she's in with the description of that state. Mary* might—as suggested in my chapter 7—believe that the phenomenal state she's in was of the type described by her functional concept. But the epiphenomenalist must deny that Mary* *knows* that the phenomenal state she's in is the functional state, since they think her belief is false. Where could Mary*'s belief formation be going wrong?

The epiphenomenalist doesn't face any more of a challenge here than in explaining the original Mary case. In the original Mary case, epiphenomenalists must accept that Mary claims to be (and is) surprised *not* because she's having a novel experience, but because of what's going on in her brain. We shouldn't be at all surprised by the possibility that Mary*'s brain should be structured differently, and so yield different vocalizations and different conscious beliefs. We shouldn't be surprised at the possibility of a Mary* who, upon having a conscious experience, immediately intuits that the experience is a particular functional state (the very functional state that her brain is in). Still, we want an explanation of why this immediate belief—formed by a reliable physical mechanism—is not justified and does not amount to knowledge.

The physical process by which the functional state is matched to the description of the functional state is reliably matching up two things based on a genuine similarity. This doesn't seem to be where belief-formation is going astray. But the epiphenomenalist can insist that there is a problem in the *conscious interpretation* of this low-level physical processing. The physical system matches the functional state it's in with the description

of the state, based on natural similarities between the two. But the conscious awareness of the functional state is—by hypothesis—a phenomenal experience. So what the creature consciously concludes from this low-level processing is a matching up of the *phenomenal experience* with the description of the functional state. The epiphenomenalist is going to say that this is where the problem lies. The functional state Mary* is in *is* described by the functional concept. But the conscious interpretation of the match-up goes beyond this, and identifies the *phenomenal experience* (an entirely different thing, generated by psychophysical bridging laws) with the description. And according to the epiphenomenalist, the phenomenal experience *is not* of the same type described by the functional concept.

While the epiphenomenalist will grant that the psychology I've described is possible, they will insist that the result is not a more rational creature (as the physicalist must), but a less rational one: The result is a psychology that leads its possessor to the false beliefs that her *phenomenal experiences* are the sorts of things described by her functional concepts, and correspondingly to a false theory of consciousness.

There could, however be a similar creature that the epiphenomenalist would take to be rationally superior to us: a creature that could directly match up functional states her brain was in with descriptions of those states, but where this low-level match up simply resulted in the conscious belief "I'm in thus-and-so functional state", without taking the *phenomenal experiences* she was having to be of that functional type. (It's worth noting that the possibility of such a creature isn't of help to the type-B physicalist. Even if such a creature is possible, the possibility of the creature I've originally described is enough to thwart their position. They have to accept that such a creature is more rational, and that—if possible—it undermines the possibility of an *ideal* epistemic gap.)

At this point, the debunking challenge rears its head. If I have given a compelling, physicalistically respectable story of why we find an epistemic gap, what reason do we have for rejecting physicalism? Yes, *we* find physicalism difficult to believe. But Mary* would find anything else difficult to believe. Mary* would find zombies utterly inconceivable.[78] Even if there is a story that can be told, on which Mary* is irrational, what *independent reason* do we have for accepting this story and privileging our own intuitions?

Unfortunately—for myself qua dualist—I don't think that we have an independent reason for privileging our intuitions. (Though I also don't think we have independent and compelling reason to privilege Mary*'s intuitions.) What we do have is a vivid a way of undermining the conceivability argument. Dualists and physicalists must both accept the possibility of Mary*, who finds zombies inconceivable. Physicalists will insist that Mary* is has worked out the phenomenal-physical truths a priori; dualists will reject this. The conceivability argument requires us to reject the idea that Mary* has special a priori knowledge that we lack. But unless there is grounds for doing so that doesn't require previous commitment to dualism—either an argument or a pretheoretic intuition—the conceivability argument is impotent against someone with a prior commitment to physicalism. Likewise, one of the presuppositions of the Knowledge argument is that

[78] One might think that there is some sense in which Mary* would be able to conceive of the possibility of zombies. Suppose Mary* is reflecting on her ability to understand the current functional workings of her brain, simply by introspecting on her experiences. Mary* realizes that she has a psychology structured as I've described, and thereby comes to conceive of the possibility of Mary (with our human psychology). Mary* realizes that Mary would be surprised when she had a red experience for the first time, and that Mary would be able to conceive of creatures physically like herself, but without phenomenology. Further, Mary* realizes that there's a hypothesis on which Mary's psychology is actually more rational than Mary*'s. If Mary* could not rule out this hypothesis a priori, then this would show that Mary* finds zombies *conceivably conceivable.* This would entail negative conceivability (she the inability to rule out the possibility of zombies a priori), though not positive conceivability (she's incapable of clearly and distinctly imagining such a possibility). Of course, if Mary* could rule out the hypothesis a priori, then all that she would find conceivable was an irrational creature who—because of her irrationality—thought she could conceive of zombies.

Mary is fully rational. But the possibility of Mary*—together with the physicalist's interpretation of Mary*—is enough to show that only someone starting with dualist commitments has to hold that Mary is fully rational. The Knowledge argument is impotent for the same reason.

8.3 Overcoming the Stalemate

We seem to have arrived at a stalemate. Suppose we begin our theorizing with the supposition that dualism is true—and in particular (since I'm assuming that physics is causally closed) that epiphenomenalism is true. As we have seen in this chapter, we can give dualistic responses to all of the standard objections to epiphenomenalism.

But now suppose we begin from the assumption that physicalism is true. We can – as we saw in chapter 7 – give a physicalistic explanation of why we have the intuitions that zombies are possible, that Mary learns something when she sees red for the first time, and that no amount of physical information can explain what it's like to see red. The explanation is that we have certain cognitive limitations that prevent us (no matter how much physical information we may possess) from matching this information up to the phenomenal information, and that our intuitions about what's possible in hypothetical cases was formed by thinking about creatures that are psychologically similar to ourselves. Because of this, we have a simple way of diffusing the standard anti-physicalist arguments: *Assuming that physicalism is true*, then Mary* can work out the phenomenal-physical truths a priori. Hence zombies are not ideally conceivable, and Mary in the original knowledge argument is not fully rational. Finally, this way of being a physicalist doesn't require denying that there's a tight connection between conceivability and possibility—something that might have posed internal difficulties to defend. So if we

start from the assumption that physicalism is true, we find a way of being a physicalist that (i) is internally coherent, (ii) gives a physicalistic explanation of *why we have* the dualist intuitions that anti-physicalist arguments rest on, and (iii) can respond to the anti-physicalist arguments as presupposing a certain (dualistic) conception of the rational agent.

There's an important parallel here between the situation that dualists and physicalists find themselves in. A complete defense of dualism can only be given if we begin with the presupposition that dualism is true. This comes out most clearly in the response to the debunking challenge. The physicalist asks "If you accept that we can give a purely physical explanation of our behavior, doesn't that entail a purely physical explanation for why we have the intuitions that motivated arguments for dualism, and hence doesn't it abolish your reason for being dualists?" The epiphenomenalist responds either by appealing to cases like arithmetic—where the fact that we can give a purely electro-chemical explanation for our belief that 2+2=4 does not debunk the belief—or she argues that a purely physical explanation cannot be given of our intuitions (our beliefs), but only of our *behavior*. Both these responses are entirely adequate at responding to the challenge, but both only work if you begin by presupposing the truth of dualism. The electro-chemical explanation for why we believe that 2+2=4 doesn't give us reason to abandon the belief in part *because two plus two really does equal four.* And a physical explanation cannot fully capture what it takes to have a belief only if physicalism is false.

But likewise, a complete defense of physicalism can only succeed given the presupposition that physicalism is true. The physicalist's response to the conceivability and knowledge arguments rested on the claim that Mary* is working out phenomenal-physical *truths* a priori. (If Mary* is not working out *truths*, then she is not a more ideal

agent, but a confused one. We lose our basis for rejecting the ideal epistemic gap.) This quite blatantly assumes that physicalism is true!

So there's no theoretically neutral way of defending either dualism or physicalism. What theory we wind up with will depend on where we start. And if we try to start "from nowhere"—from no assumptions about the nature of the mind—then we will not be able to give a complete defense of either position.

Where do we go from here? How *can* we go on from here? We might try to find independent points in favor of either side to sway the balance.

Both physicalism and epiphenomenalism are equally compatible with all current and future physical evidence that might be discovered, so a posteriori considerations will not help break the tie. The physicalist might appeal to the comparative simplicity of her theory as a tie-breaker. The dualist might here respond with an allusion to Alonzo Church's satirical "The Ontological Status Of Women and Abstract Entities", which suggests that just as the nominalist might appeal to the simplicity of her theory as a deciding virtue, so an "ontological misogynist" might appeal to simplicity to support the theory that women are not independently existing entities, but merely shadowy "secondary presences" of their husbands or fathers. Clearly even if such a theory was just as explanatorily powerful as the True Ontological Theory of Women, simplicity would not be enough to sway in its favor. The dualist will say "likewise in the mind-body case."

The dualist may also question why – if simplicity is such a great concern – we should be swayed towards physicalism, rather than to idealism. Idealism deals with the mind in the most straight-forward, intuitive way. And it has all the simplicity advantages of physicalism. Yet many philosophers, confronted with a simplicity argument for idealism, balk, claiming that simplicity concerns are incapable of swaying us towards

anything so intuitively bizarre as idealism. But this is precisely the position dualists see us in regarding physicalism.

Further, the dualist might point out that – unlike the physicalist who deals with our intuitions by explaining them away – she is able to *accommodate* our intuitions. Surely, she says, it's better to respect our most deeply held intuitions than to systematically explain them away out of a commitment to a theoretical metaphysical viewpoint (physicalism). The physicalist responds that we have no independent reason to privilege our intuitions over those of Mary*. The dualist insists that this is only true if you presuppose the truth of physicalism, the physicalist responds that the contrary position is only true if you presuppose the truth of dualism, and we're back at it again! It is not clear that they will be able to say anything that could capable of breaking the stalemate, or that will be taken seriously by those on the other side.

If looking narrowly at the mind-body problem, neither side can get a grip against the other, I propose that we need to broaden our scope. The problem of consciousness isn't isolated from the rest of philosophy of mind, or from metaphysics and epistemology more broadly. What position you take on the nature of consciousness will have implications for the rest of philosophy of mind. (We've already seen one instance of this arising from the "paradoxes of phenomenal judgment", where dualism about qualia commits you to a dualistic view of personal identity and beliefs.) Since we cannot overcome the stalemate looking narrowly at consciousness, I propose that we must expand the scope of our discussion. Each side must aim to give a *total theory of mind:* a theory of consciousness, intentionality, personal identity, and perception. Once we have these competing total theories on the table, we can assess the merits of each. Each side may be able to tell an internally coherent and compelling story about consciousness, but

ultimately we want something more. We want a coherent and compelling story about the mind as a whole.

This points to a second, perhaps more satisfying, reason why simplicity cannot (at present) tip the scale in favor of physicalism. If theories P and Q are both able to accommodate all of the data points, and theory P is more simple than theory Q, this may tell in favor of theory P. But simplicity only comes in as a tie-breaker once we know that both theories can accommodate all of the data points. If P can only account for a small subset of what needs explaining, then the fact that it is simpler doesn't tell in its favor. It may be that once we have worked out a total physicalistic theory of mind and a total dualistic theory of mind, we will find that both can accommodate all of the points that need explaining. In this case, simplicity may tip the balance in favor of the physicalist. But we are not in this position yet. We must first see whether we can construct Total Theories of Mind consistent with physicalism and with dualism, and see what implications these Total Theories of Mind have.

What would a Total Theory of Mind look like? It will tell a story about personal identity at a moment and over time. It will explain what it takes for there to be meaningful thoughts *about* the world. It will explain how perception can give us direct access to reality. Some of these tasks pose challenges for the physicalist in particular. To give a physicalistic theory of intentionality, one must be capable of constructing a relationship of *aboutness* simply by putting together increasingly complex physical relations. How could this work? Can the physicalist tell a story about intentionality that prevents intentionality from pervading the world in an unintuitive way? If intentionality is simply a matter of carrying information about the world, say through causal covariance of the sort thermometers and tree rings covey, then is intentionality everywhere?

Likewise can the physicalist prevent consciousness from pervading the world? If being conscious is simply a matter of being in the right sort of functional state, can we flesh out what it takes to be the right sort of functional state in a way that doesn't give a continuum from me to computers to calculators to soda machines? On the topic of personal identity, can the physicalist account for the intuitive "strong unity" of minds? Minds are intuitively unified—candidates for objecthood—in a much stronger way than any other aggregate of matter. But can the physicalist account for this intuitive datum if minds simply are complex aggregates of physical matter?

Some of these challenges extend to the epiphenomenalist. If the epiphenomenalist thinks that complex functional states together with psycho-physical laws are responsible for generating phenomenal experiences, then she will face the same challenge of how to avoid consciousness (and consequently intentionality) from pervading the world.

There is some reason to be skeptical of whether expanding our view can help to break this stalemate. Perhaps the disagreement over consciousness will simply pervade all of these other topics. Suppose that the dualist, unsatisfied with physicalistic accounts of intentionality, tries to tell a story on which consciousness is essential for intentionality. The physicalist might try to co-opt this story, saying that they too think that consciousness is essential for intentionality (although they think that consciousness is a purely physical phenomena). They might even introduce an appeal to the distinctive way we think about consciousness to overcome the intuition that their physicalistic account failed to get to the heart of genuine intentionality. We would simply have the same argument repeating itself in a different sphere.

There is much to this skepticism. It may be naïve optimism to think that expanding our horizons will help to resolve anything. But there is some reason to think

that having parallel debates in other areas of philosophy of mind might add value. It is one thing for a theory to reject and explain away a single very strongly held intuition. It is another for that theory to explain away *every* strongly held intuition we hold about its domain. If it turns out that physicalists need to explain away not only our most deeply held intuitions about consciousness, but also our most firmly held intuitions about intentionality and personal identity, the cost might eventually seem to be too high. We have very little to go on in our theorizing about the metaphysics of mind apart from intuitions. The more intuitions we must explain away, the more me must question what we are getting in return.

8.4 Conclusion

We have seen that the correct understanding of how we think about conscious experiences has huge implications for our metaphysics of mind. Phenomenal concepts can help to reveal a compelling version of physicalism. This view accepts the datum of an epistemic and explanatory gap, which cannot be bridged with more knowledge, more caffeine, or more training. The epistemic and explanatory gaps are explained away as features of how creatures psychologically like us think about consciousness.

Phenomenal concepts can also help to reveal a compelling version of dualism: a version of epiphenomenalism that can avoid the paradoxes of phenomenal judgment and other epistemic worries. Each of these mind-body positions is internally defensible, not just in that a committed member of either camp cannot be persuaded to abandon their position (ignoring intuitive data where necessary to remain on their favored side), but in the stronger sense that both sides offer compelling explanations of all the intuitive data that the other camp tries to use against them. We face a stalemate.

I have suggested that because the problem of consciousness is so interwoven with other debates in the philosophy of mind, expanding our horizons to give a Total Theory of Mind may prove fruitful. In addition to these further questions in the philosophy of mind, this ending point leaves us with some pressing metaphilosophical questions: First, what is (and what should be) our aim in debates like this? Is a good argument one that convinces committed rational proponents of the opposing side? One that convinces rational agents who are neutral on the topic? What would it take to be neutral on the topic? Simply not to have given the issue enough thought to have formulated a position? Or do one's intuitions in some sense need to be "up for grabs"? Is the good argument one that will work against agents psychologically like us – with human-like intuitions? Or are we mistaken to privilege human intuitions in this way? And if this privileging is mistaken, how can we hope to make progress in philosophy?

Bibliography

Alter, Torin. (2007). On the Conditional Analysis of Phenomenal Concepts. *Philosophical Studies,* 131(3): 777-778.

Balog, Katalin. (2012a). Acquaintance and the mind-body problem. *The Mental, the Physical.* Eds. Hill, C. & Gozzano, S. Cambridge: Cambridge University Press.

Balog, Katalin. (2012b). In Defense of the Phenomenal Concept Strategy. *Philosophy and Phenomenological Research,* 84(1): 1-23.

Berkeley, George. (1710). *A Treatise Concerning the Principles of Human Knowledge.*

Bigelow, J. & Pargetter, R. (2004). Acquaintance with Qualia. *There's Something About Mary*. Eds. Ludlow, Nagasawa, Stoljar. MIT University Press.

Block, Ned. (2006). Max Black's Objection to Mind-Body Identity. *Oxford Studies in Metaphysics*. Vol. 2, Ed. D. Zimmerman. Oxford University Press.

Braddon-Mitchell, David. (2003). Qualia and Analytic Conditionals. *Journal of Philosophy,* 100(3): 111-135.

Braddon-Mitchell, D. & Jackson, F. (2007). *The Philosophy of Mind and Cognition*. Malden, MA: Blackwell Publishing.

Burge, Tyler. (1988). Individualism and Self-Knowledge. *Journal of Philosophy*, 85: 649-663.

Chalmers, David. (1996). *The Conscious Mind.* New York: Oxford University Press.

Chalmers, David. (2002a). Consciousness and Its Place in Nature. *Philosophy of Mind: Classical and Contemporary Readings.* Ed. Chalmers. Oxford University Press.

Chalmers, David. (2002b). Does Conceivability Entail Possibility? *Conceivability and Possibility.* Eds. Gendler & Hawthorne. Oxford University Press.

Chalmers, David. (2003). The Content and Epistemology of Phenomenal Belief. *Consciousness: New Philosophical Perspectives.* Eds. Smith & Jokic. Oxford University Press.

Chalmers, David. (2006). Two-Dimensional Semantics. *The Oxford Handbook to the Philosophy of Language*. Eds. Lepore & Smith. New York: Oxford University Press.

Chalmers, David. (2007). Phenomenal Concepts and the Explanatory Gap. *Phenomenal Knowledge and Phenomenal Concepts.* Eds. Alter and Walter. Oxford University

Press.

Chalmers, David. (2010). Two-Dimensional Argument Against Materialism. *The Character of Consciousness.* Oxford University Press.

Chalmers, David. Responses to Articles on My Work. http://consc.net/responses.html#braddon

Church, Alonzo. (1958). The Ontological Status Of Women and Abstract Entities. Lecture reprinted: http://cs.nyu.edu/pipermail/fom/2005-September/009079.html

Galton, Francis. (1880). Statistics of Mental Imagery. *Mind* (5): 301–318. Reprinted: http://psychclassics.yorku.ca/Galton/imagery.htm

Haukioja, Jussi. (2008). A Defense of the Conditional Analysis of Phenomenal Concepts. *Philosophical Studies,* 139(1): 145-151.

Hawthorne, John. (2002). Advice for Physicalists. *Philosophical Studies,* 109(1): 17-52.

Hume, David. (1748). *An Enquiry Concerning Human Understanding.*

Jackson, Frank. (1982). Epiphenomenal Qualia. *Philosophical Quarterly,* 32: 127-136.

Jackson, Frank. (2000). *From Metaphysics to Ethics.* Oxford University Press.

Johnston, Mark. (2007). Objective Mind and the Objectivity of Our Minds. *Philosophy and Phenomenological Research,* 75(2): 233-268

Kripke, Saul. (1972). *Naming and Necessity.* Blackwell Publishing.

Levin, Janet. (2007). What is a Phenomenal Concept? *Phenomenal Knowledge and Phenomenal Concepts.* Eds. Alter & Walter. Oxford University Press.

Levine, Joseph. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly,* 64: 354-361.

Levine, Joseph. (1993). On Leaving Out What It Is Like. *Consciousness: Psychological and Philosophical Essays.* Eds. Davies & Humphreys. Blackwell Publishing.

Loar, Brian. (1997). Phenomenal States (Revised Version). *The Nature of Consciousness*. Eds. Block, Flanagan, Güzeldere. MIT University Press.

Locke, John. (1690). *An Essay Concerning Human Understanding,* Book III.

McGinn, Colin. (1991). Consciousness and Content. *The Problem of Consciousness: Essays Towards a Resolution.* Oxford: Blackwell Publishers.

McGinn, Colin. (2004). Images and Percepts. *Mindsight: Image, Dream, Meaning.*

Cambridge, MA: Harvard University Press.

Mendelovici, Angela. (2010). *Mental Representation and Closely Conflated Topics.* (Doctoral Dissertation).

Nida-Rumelin, Martine. (2004). What Mary Couldn't Know: Belief about Phenomenal States. *There's Something About Mary*. Eds. Ludlow, Nagasawa, Stoljar. MIT Univeristy Press.

Papineau, David. (2002). *Thinking about Consciousness*. Oxford University Press.

Papineau, David. (2007). Phenomenal and Perceptual Concepts. *Phenomenal Knowledge and Phenomenal Concepts.* Eds. Alter & Walter. Oxford University Press.

Perry, John. (1977). "Frege on Demonstratives." *Philosophical Review,* 86(4): 474-497.

Perry, John. (2001). *Knowledge, Possibility, and Consciousness.* MIT University Press.

Pitcher, George. (1977). Abstract Ideas. *Berkeley*. New York: Routledge Press.

Schroer, Robert. (2010). Where's the Beef? Phenomenal Concepts as Both Demonstrative and Substantial. *Australasian Journal of Philosophy,* 88(3): 505-522.

Stoljar, Daniel. (2005). Physicalism and Phenomenal Concepts. *Mind and Language* 20: 469-494.

Taylor, C. C. W. (1978). Berkeley's Theory of Abstract Ideas. *The Philosophical Quarterly,* 28: 97-115.

Tye, Michael. (2003). A Theory of Phenomenal Concepts. *Minds and Persons*. Ed. O'Hear. Cambridge University Press.

Tye, Michael. (2009). *Consciousness Revisited.* MIT University Press.