Title
Converse Intentionalism and Experiential Content

Author
Chen, I-Sen

Publication Date
2021

Peer reviewed|Thesis/dissertation

Converse Intentionalism and Experiential Content

By

I-Sen Chen

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Philosophy

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____

Bernard Molynuex, Chair

_____

Adam Sennet, Co-Chair

_____

Zoe Drayson

Committee in Charge

2021

To Dr. Hung-Ming Chen, and Ms. Tsung-Tzu Chen

Abstract

My dissertation is a critical essay on converse intentionalism as embedded in the contemporary philosophy of experience, which says that for any experience, its intentional content supervenes on its phenomenal character. This principle has been deeply involved in a lot of discussions on the content of experience. For example, contemporary phenomenal-intentionality theorists argue for their position by appealing to scenarios in which a brain-in-a-vat (BIV) which is causally disconnected from the external environment, and which has the same phenomenal states as normal humans do, has the same intentional content as normal humans (Horgan and Tienson 2002, Loar 2003, Pautz 2006a, Chalmers 2004). In their view, this BIV type of thought experiment is intuitively compelling and strongly convinces them of converse intentionalism.

So just as in epistemology where converse intentionalism is often tacitly assumed, it is also widely accepted in the philosophy of experience. As we shall see later, whether you are a naturalist or a non-naturalist, a semantic externalist or a narrow theorist, converse intentionalism is a commitment you have strong reason to make. Ironically, however, as in Greek tragedy where one cannot avoid what one wants to, I will argue that a philosopher of experience, whether she is a naturalist or a non-naturalist, an externalist or a narrow theorist, can never embrace converse intentionalism without undermining her own position. Given that converse intentionalism gives rise to tension within their respective theories, I will propose a replacement for philosophers of experience, and I will argue that this new theory is probably the best they can have.

From the naturalist point of view, every property can be understood in terms of the physical sciences. Typically, philosophers naturalize intentionality by holding semantic externalism of mental content i.e., that the content of a mental state is determined by entities in the environment. Then to naturalize phenomenology, they may argue that phenomenology is *identical* to experiential

intentionality. Since identity entails co-supervenience, we can see why naturalists/externalists endorse converse intentionalism.

Meanwhile, narrow content theorists also endorse converse intentionalism. According to almost all such theorists, a semantic theory assigning a content to a mental state should at least do so in accordance with the psychological role of that state (e.g., Loar 1988a, b). Given the strong connection between phenomenology and psychological role—e.g., pain causes avoidance—narrow theorists argue that content is determined by or identical to phenomenology. This is why converse intentionalism generally accompanies narrow theorists' core thesis.

However, as Block's Inverted Earth case (1990) suggests, intentional content as understood by naturalists and externalists is determined by the external environment whereas intuitively, phenomenal character is independent from the external environment. Hence, I will argue in Chapter 2 that we would have an intuition clash if naturalists' identification of phenomenology and intentionality is accepted. On the other hand, the connection between phenomenology and psychological role as presupposed by the narrow theorist to argue for converse intentionalism is questionable. For as Block's (1978) argues, the psychological role of a mental state is relational while the phenomenal character of a state is intrinsic, and hence it is possible to fix the psychological role of a state without fixing its phenomenal character. Conversely, Block (2007) also argues that experiences with the same intrinsic properties may not have the same psychological role. Once again, as I will show in Chapter 3, we would have an intuition clash if narrow theorists' presupposition that they co-supervene on each other is upheld.

This suggests that the crux of the issue is that phenomenal character is internal and intrinsic while psychological role and content are relational. That is why we always have an intuition clash whichever semantic theory is accepted. I suggest we should replace it. To be clear, by *replacing*

converse intentionalism, I aim to find a principle that achieves the following two goals. First, it should address the primary concerns of externalists and narrow theorists. In the former case, the new principle should help externalists naturalize intentionality. In the latter case, it should satisfy narrow theorists' requirement that content assignments should be in accordance with psychological role. Second, the new principle should avoid all the counter-examples discussed in this dissertation, and should not be subject to new counter-examples.

To replace converse intentionalism, in Chapter 4, I propose a new principle that for any two experience tokens, if they have the same phenomenal character *and* psychological role, then they have the same content. I shall call it "converse psychointentionalism". We can sense *prima facie* this principle is less likely to cause such intuition clashes for us since both the antecedent and the consequent involve relational properties, i.e., the former involves psychological role and the latter involves content. I will further argue at the second half of Chapter 4 that converse psychointentionalism can satisfactorily address the primary concerns of naturalists/non-naturalists and of externalists/narrow theorists.

In sum, we should replace converse intentionalism with converse psychointentionalism. In my view, philosophy of experience has been dominated by converse intentionalism, which generates a lot of puzzles and problems for philosophers whether they are naturalists, non-naturalists, externalists or narrow theorists. If my arguments are correct, it is time for us to jettison converse intentionalism, and replace it with a less problematic principle.

Acknowledgements

This dissertation grew out of a question my supervisor, Bernard Molyneux, asked me seven years ago when I was writing a paper on perceptual epistemology. “Don’t you think indistinguishable experiences have the same content?”, Bernard asked. I thought the answer was obvious: No, they don’t. What I did not anticipate, however, was that it took me seven years and more than a hundred pages to justify my answer. At first, I thought it was only an issue in epistemology. Once I started to work on this project, I realized that I was wrong. In order to properly understand the classic arguments and theories, I had to expand my purview, and to study the mind-body problem, informational semantics, narrow content, the metaphysics of propositions, computationalism, philosophy of science, and many other subfields in philosophy. In a sense, this dissertation not only presents my answer to Bernard’s question, but also embodies what I have learned and my training at UC Davis over the eight years.

I first want to thank my committee. I am indebted to Adam Sennet for being my instructor and clarifying my own ideas when this project was at its embryonic stage. I thank Zoe Drayson for teaching me to be a deeper thinker than I was. I am grateful to Cody Gilmore who always urged me to present my arguments in precise ways. Finally, I owe special thanks to Angela Mendelovici at Western University whose works provide great inspirations to my dissertation.

Over the years, I have benefited enormously from the faculty at UC Davis. Their training has helped me build up my toolkit to address the issues in my dissertation. Many thanks to: GJ Mattey for his class on the theory of ideas in the early modern era; Alyssa Ney with whom I studied the mind-body problem and philosophy of science; Hanti Lin whose seminars on formal epistemology helped me better understand the connection between epistemology and cognition;

finally, Rohan French who under my request kindly offered a seminar on inferential role semantics. All the knowledge and skills I learned from them are manifested in my dissertation, directly or not.

I also want to apologize to everyone whom I bothered with my philosophical ideas. I have interruptively knocked on many professors' office doors or sent text messages at midnight to friends just to discuss philosophy. I thank them for tolerating my unconventional behaviors and for sharing their intelligence and friendship with me. I am especially grateful to Aldo Antonelli, Elaine Landry, Roberta Millstein, Jim Griesemer, David Copp, Marina Oshana, Jan Szaif, Tina Rulli, Paul Teller, Jonathan Dorsey, Scotty Dixon, Ted Shear, Chris Healow, Tyrus Fisher, Kyle Adams, Noel Josh-Richard, Rick Morris, Jacob Velasquez, Rachel Boddy, Arieh Schwartz, Patrick Skeels, Jordan Bell, Da Fan, Max Parks, Denise Hossom, Tiernan Armstrong-Ingram, Scott Cole, Danielle Williams, Khang Ton, Selcuk Kaan Tabakci, Ramiel Tamras, Aaron Chavez, Natasha Haddal, Chanwoo Lee, Lauren Viramontes, Stephen Cunningham-Bryant, Jason Mosebach, Stewart Harrison, Jerome Romagosa, Kory Matteolli, Dylan Goldman, Lel Jones, and Ofri Oren.

My greatest debt is to my supervisor, Bernard Molyneux, for everything he has done for me. He is always patient with me as a slow thinker; he improves my English writing as a foreigner, and finally he encourages me to be a brave philosopher.

Finally, I am grateful to my father, Hung-Ming Chen, my mother, Tsung-Tzu Chen, my sister, I-Yun, and my brother I-Lin for their unconditional love. I also want to thank my other family members and friends in my home country for their emotional and material support throughout the past eight years.

Table of Contents

Preface

It has been one of the most worrisome possibilities for philosophers that we could just be brains in vats. In that scenario, we are just a bunch of brains floating in vats. The vats are filled with nutritious fluid to keep us alive and connected to some machines via electrodes. Those electrodes stimulate neural activities to create experiences with the same phenomenal character as the experiences we currently have. According to the standard scenario, there are no flowers, rivers, animals, tables, and chairs. Yet, in this scenario, we still have experiences with the same phenomenal character. So it seems that what we see, hear, or smell—what we experience, in short—are radically mistaken.

To use this thought experiment to argue for external-world skepticism, we need to argue for the premise that a brain in a vat ("BIV" hereafter) whose experience has the same phenomenal character as ours is possible. But this is not enough. We also need something like the following principle:

> *Converse intentionalism*: Content supervenes on phenomenology, i.e., if every two experience tokens are the same with respect to what it is like, then they are the same with respect to what it is about.

This principle is what guarantees that a BIV has an experience of the external object. Only if some link of this sort is assumed can we infer that the experience of a BIV which is the same as my experience with respect to what it is like is about the same thing as mine, and only then that its experience is radically mistaken.[1]

The fact that external-world skepticism has baffled philosophers proves the point that the BIV scenario is compelling. But why is it so? What explains the plausibility of the argument?

[1] Prof. Cody Gilmore reminds me that skeptics do not need to assume converse intentionalism to argue for the possibility of the BIV scenario. All they need is to somehow justify the claim that it is possible that we are BIVs having radically mistaken experiences with the same phenomenology and the same intentional content. I totally agree with his observation. However, my aim here is simply to expose one of the main motivations behind the BIV scenario, i.e., converse intentionalism, that seems to be unnoticed by epistemologists. I am not claiming that it is indispensable to the skeptic's argument. In addition, as we will see, any solution to the skeptic's challenge that violates converse intentionalism is typically taken to be counter-intuitive, which shall strengthen my contention that converse intentionalism is one of the main, albeit not the only, motivations for the BIV scenario.

Certainly, the possibility of a BIV is an empirical claim. Its plausibility arises from empirical sciences. In contrast, converse intentionalism seems to be independent from empirical studies. It is plausible *because it is independently intuitive*. Indeed, it is suggestive to observe that while the phenomenal duplicate in the skeptic scenario has taken different forms over the centuries, e.g., a disembodied ego, then a BIV, and finally a real human living in virtual reality (as described in the movie *The Matrix*), converse intentionalism persists. Hence the intuitiveness of converse intentionalism is isolated from empirical science.

Another way to appreciate how intuitive converse intentionalism is is to see how counter-intuitive it is to reject it. In the literature, Davidsonian/Putnamian strategies for undermining external-world skepticism are considered counter-intuitive precisely they reject converse intentionalism. For example, Putnam (1981) holds the simple causal theory of mental content. According to this view, the content of an experience token is simply its actual cause. If this theory is true, then my experience as of a red apple could mean the vat in which I as a brain lives even though the phenomenal character of my experience remains the same. Davidson (1986) endorses the interpretivist theory of mental content which says that the mental content of an agent S is determined by how the ideal observer of S interprets S's mental states according to the principle of charity. That principle requires the idea observer not to interpret S's mental states in such a way that S's beliefs are systematically false given S's interaction with her environment. Although this theory ensures that S's beliefs are not systematically false, and thereby resolves the skeptic's challenge, it also implies that the contents of S's mental states, including her experiences, do not supervene on the intrinsic properties of S's mental states. This reply to the skeptic argument amounts to rejecting converse intentionalism, and hence is taken to be counter-intuitive. If we are going to reject converse intentionalism, we had better have a replacement that preserves our intuitions.

As I have said, my dissertation is a critical essay on converse intentionalism as embedded in the context of philosophy of experience. In this preface, I have discussed one

reason why converse intentionalism is important for philosophers. It provides the intuitive force for the external-world skeptic argument. In the following chapters, we will see more reasons why almost all philosophers of experience are committed to it. My goal in this project is to convince philosophers that we don't have to accept it. In fact, we have stronger reason to reject converse intentionalism and we have a better option to replace it.

Chapter 1

Introduction: Experience, Strong Representationalism and Phenomenal Intentionality

§1 Introduction

The aims of this chapter are: (i) to provide the background for later discussions, and (ii) to provide a brief overview of my whole dissertation.

My dissertation consists of five chapters. Chapter 1 (i.e., the current chapter) is the introduction. Chapters 2-4 are the main contents of my dissertation. In Chapter 5, I address objections that do not belong to other previous chapters.

Here is the plan for this chapter. In §§2−3, I will provide the background and expositions of the relevant theories and arguments. In §§4−6, I will sketch my arguments in Chapters 2-4. In §7, I will briefly summarize this chapter.

Conventions: I will use capital letters for mentioning mental concepts. Terms with the prefix "ph-" denote phenomenal characters, e.g., ph-red is the phenomenal character she has when she sees red objects.

§2 Reference-Fixing: Semantics, and Phenomenology

§2.1 Semantics

By "semantics", I mean what Block refers to as metaphysical semantics, i.e., the *metaphysical* study of meaning and content (Block 1998, p. 653; cf., Stanley and Szabó 2000, Speaks 2010/2019, Mendelovici 2013, 2018).[1] Some of its core issues are: what is the ontology of content? Is it a monadic property? A relation? An object? If it is an object, is it abstract, mental or physical? What is the relation between the content and the content-bearing entity? Is it constitution? Or causation?

[1] What I have in mind roughly corresponds to Block's "metaphysical semantics" (1998, p. 653), Stanley and Szabó's "foundational semantics" (2000), Speaks' "foundational theory of meaning" (2010/2019), and Mendelovici's "metaphysical theory of mental representation" (2013) or "theory of intentionality" (2018).

For our purpose, the important issue in semantics is the debate between semantic externalism and the narrow content theory.[2] Strictly speaking, it is the debate between the view that the content of an intentional state does not supervene on the subject's internal states (externalism) and the view that insists that content does supervene on the subject's internal states (the narrow content theory). However, theorists of both stripes typically have stronger commitments than the mere extrinsicness/intrinsicness of intentional states.

For example, most externalist theories say that for a mental state to have some external-physical entity X, e.g., an object or a property, as its intentional content is for it to stand in an appropriate *relation to X*. E.g., according to the causal-informational version of externalism, for John to token a thought that the apple in front of him is red, this thought must stand in an appropriate causal relation to the apple in front of him and to the property red. To be sure, to say that intentional states are extrinsic does not mean that to token an intentional state is to stand in a *relation to the content*. One can hold the view that when I assert the sentence that "Sherlock Holmes is a detective", the term "Sherlock Holmes" means what it does because of the convention that governs the public language I speak and hence it does not supervene on my internal states. But that view does not imply that when I assert that sentence, I stand in a relation to Sherlock Holmes or to something else. This illustrates one way in intentional properties might be extrinsic without being relations. However, since most externalists are relationalists, I will set aside the non-relational views in my dissertation.[3] [4]

---

[2] For an overview of the debate between externalism and the narrow theory, see Fodor 1994, chapter 1.

[3] Notice that relational externalism entails relationalism of intentionality while relationalism does not entail externalism. E.g., Pautz (2006a, b, 2008) argues that to token a color experience is to stand in a *sensory-awareness* relation to the color, and that whether a subject stands in the relation to that color depends on the phenomenal character of the experience.

[4] In the literature, there are many arguments from compositionality against different versions of the monadic view of intentionality, e.g., Jackson's (1977) argument against adverbialism of perceptions, and Fodor's (2010) argument from the compositionality of thought against inferential role semantics. See Garson 2001 for the inferentialist account of compositionality and Mendelovici 2018 for the non-relationalists' reply to this objection.

To sketch a typical argument of this sort, consider a version of long-arm inferential role semantics according to which my concept CAT means what it means because of the linguistic inferential role of the term "cat" in the public language I speak. Given the compositionality of thought, the content of RED CAT is *composed* by the contents of RED and CAT, which according to inferential role semantics under consideration are in turn determined by the inferential roles of "red" and "cat" in English. However, as Fodor argues, there is no well-

Similarly, in addition to the central principle that content supervenes on the internal states of a subject, we should also notice that the narrow theory's central principle is standardly motivated by the methodological constraint that any semantic theory should accommodate the intuition that the content of an intentional state matches its psychological role, i.e., whenever a semanticist assigns contents to a mental representation, she should do so in such a way that her assignment characterizes the agent's psychology. Here the term "psychology" is to be understood broadly: it includes common sense intentional psychology, information-processing, rational reasoning, and behavior-producing processes. To be sure, this is not to deny that there are other motivations for the narrow theory. In fact, some philosophers endorse the narrow theory because they believe that our knowledge of the content of our intentional state should be introspectively accessible (Yli-Vakkuri and Hawthorne 2018, Mendelovici 2018). Since most narrow theorists are motivated by the psychological constraint of content, I will also take it as one of the core theses of the narrow theory.[5]

Here are two kinds of thought experiments for which externalists and narrow theorists typically deliver contradictory verdicts. The first one is *Frege's puzzle*, and the second one *the Twin Earth case*. Let's start with the former. Consider the thoughts expressed by the following two sentences "Hesperus is Hesperus", and "Phosphorus is Hesperus". By the causal-informational version of externalism, the contents of the two thoughts are determined by

---

defined notion of how inferential roles could compose another inferential role. Hence, he concludes that inferential role semantics is objectionable. In contrast, relationalism does not have this problem. By relationalism, the concept RED CAT means what it means because it stands in a *relation* to the property of being a red cat. By compositionality, the content of RED CAT is composed by the properties redness and cathood. Since the compositionality of properties can be easily cashed out using either set theoretic intersection or logical conjunction, relationalism can satisfy the compositionality of thought.

[5] Yli-Vakkuri and Hawthorne (2018) list six theoretical roles that narrow content is expected to serve: (a) to make true our common sense belief-ascriptions (Fodor 1987), (b) to serve as a supervenience base for rationality (Chalmers 2018), (c) to be introspectively accessible (Mendelovici 2018), (d) to serve as explanations of actions (Lewis 1979), and (e/f) to serve as explanations of perceptual and cognitive phenomenology (Pautz 2009, Kriegel 2013a). Frances Egan (1995) argues that (g) cognitive scientists posit narrow content to explain some (subconscious) information-processing. Among these seven theoretical roles, four of them (a, b, d, g) can be subsumed under the methodological constraint I mention in the main text that content assignment should be aligned with psychology. Furthermore, as I will argue later, those theorists who endorse d, e, and f also endorse the methodological constraint. Hence it is fairly uncontroversial to take the methodological constraint to be one of the core principles of the narrow theory.

whatever relevant object(s) causes them, i.e., Venus. Hence, the sentences express thoughts with the same content. However, as Frege points out, these thoughts don't have the same psychological roles: one is cognitively significant, but the other is not. An agent can rationally believe one without believing the other. So narrow theorists would claim that the thoughts do not have the same content.

On the other hand, consider the Twin Earth case proposed by Putnam (1975). Suppose on the far side of the galaxy, there is a planet, called "Twin Earth", on which everything is exactly the same as on Earth except that the transparent odorless liquid that falls from the sky when raining, that exists in rivers and lakes, that is consumed by people when they are thirsty, etc., is constituted by a chemical substance XYZ, not by H_2O. Furthermore, some Twin Earthlings speak a language, Twin English, which is orthographically the same as English on Earth. Suppose an Earthling, Oscar, tokens a thought and expresses the thought by asserting the sentence "water is transparent" in English, and a Twin-Earthling, Twin Oscar, who is physiologically the same as Oscar, tokens a thought and expresses it by asserting the sentence "water is transparent" in Twin English. Do their thoughts have the same content? Once again, externalists and narrow theorists disagree. Externalists, e.g., Putnam (1975), claim that their thoughts have different contents since it is H_2O that causes Oscar's thought whereas Twin Oscar's is caused by XYZ.[6] Narrow theorists, e.g., Fodor (1987), in contrast, insist that their thoughts should be assigned the same content since (i) we should assign content to a thought according to how we type-individuate the thought, and (ii) we should type-individuate a thought by its internal causal role.

[6] As Putnam argues, if sameness of content entails sameness of truth-condition, *since Oscar's thought and Twin Oscar's have different truth-conditions*, their thoughts have different contents. E.g., Twin Oscar's thought is false if it is tokened on Earth whereas Oscar's is true. As Putnam says, the reason why their thoughts have different truth-conditions is because the truth-condition of Oscar's thought depends on H_2O while that of Twin Oscar's depends on XYZ.

§2.2 Phenomenology

By "the phenomenal character of a state", I mean the phenomenal properties that jointly characterize the state. In the literature, the phenomenal character of a state is standardly identified with *what it is like* to be in that state (Nagel 1974), and a state of an object is a phenomenal state iff there is something it is like for that object to be in that state. For example, my state of *being in pain* is a phenomenal state since there is something it is like for me to be in pain, i.e., the painful feeling. If you ask me what it is like to be a certain pain state, I can answer by describing it. E.g., it is a piercing pain, or a burning pain. In contrast, the state of *being a cloud* is not a phenomenal state since there is nothing it is like to be in that state. (If you ask me what it is like to be a cloud, I can only scratch my head, and wonder if that question makes sense.)

There are many features about phenomenal character which are widely though not universally accepted among philosophers. First, *prima facie* every phenomenal state is an access *conscious* state, i.e., it is available for conscious report (Chalmers 1995, Bermúdez 2014). E.g., it seems incoherent for anyone to say "it feels painful now but I am not aware of it." The converse statement however is more controversial. Some philosophers hold that a belief, even an occurrent one, is not a phenomenal state (Kind 2008). Second, phenomenal character is *simple and/or homogeneous*. To say a (mental) property is simple in this context implies that it is does not decompose into type-distinct entities. To illustrate, if we examine water (via a microscope), we see that a sample of water is constituted by a bunch of objects, i.e., water molecules, which are in turn constituted by other type-distinct objects—hydrogen atoms and oxygen atoms—whereas if we examine a pain state (via introspection), we do not find it to constituted by any entities of different types. Furthermore, phenomenal character *prima facie* seems to be a monadic intrinsic property of a mental state, as opposed to a relational property. Notice that to say that a property is intrinsic does not imply that instantiating it does not bring the object to stand in relation to other objects. For instance, being six feet tall is a

monadic property, but instantiating it brings about instantiating a relation, e.g., being taller than an infant. Similarly, to say that pain is intrinsic does not imply that to be in a pain state does not bring the agent to stand in a relation. If I am in pain, I would get away from the source that damages my body. Furthermore, phenomenal character is *subjective* (Nagel 1974, Jackson 1982, Levine 1983, Loar 1997). If some state of me is phenomenal, then it is phenomenal (only) *to me*. My pain state is phenomenal to me, but is not phenomenal to you. In addition, philosophers often associate the notion of subjectivity with the idea of first-personal immediate knowability (Kim 2010, p. 280). From your perspective, by seeing me doing some actions, you can *infer* that I am in pain. In contrast, if I am in pain, then without any inference, I can directly know that I am in pain.

Two remarks before we move on: first, the above features of phenomenal character and phenomenal states are based on *prima facie* observations. All of them are debatable. It may for instance turn out that upon philosophical reflection, phenomenal character is a relation (Dretske 1995). Or, if C-Fiber Physicalism is true, then a pain state may have neural parts, and hence not be simple. Second, the above list of features is not exhaustive. I only listed those features that will be relevant to my later arguments. I have omitted those features of phenomenal character which are independent from my arguments even if they are plausible.[7]

Now I want to introduce a notion I call "the phenomenological gap". By saying that there is a *phenomenological gap* between two states, I mean that there is at least one phenomenal feature that given our phenomenological intuition, we would attribute to one state and would resist attributing to the other one. For example, there is a phenomenological gap between the state of tokening a ph-green experience and the state of tokening a ph-orange

[7] For example, some philosophers say that a concept referring to a phenomenal state is not arbitrarily substitutable with a co-referring concept *salva veritate* (cf., Jackson 1982). Suppose the concept PAIN and the concept C-FIBER FIRING co-refer, and I know that they do. However, even though it is true for me to say that I am directly aware that I am in a pain state, it seems implausible for me to say that I am directly aware that I am in a c-fiber-firing state since I can only be aware of that by inference. I think that the non-substitutivity about (the concepts of) phenomenal states is plausible. Nevertheless, since, as I have said in the main text, it is not relevant to my later arguments, I will not discuss it in my dissertation.

experience. Or there is a phenomenological gap between the pain state and the c-fiber firing. A theory claiming that some two states are identical is said to *have no* phenomenological gap iff if theory is true, there is no such discrepancy of attributing phenomenal features given our phenomenological intuition. For example, if a theory identifies ph-green states and ph-red state, then it has a phenomenological gap. On the other hand, a trivial theory identifying the state of experiencing the temperature of zero degrees Celsius and the state of experiencing the temperature of thirty-two degrees Fahrenheit has no phenomenological gap because there is no phenomenal feature that we attribute to one and resist attributing to the other given our phenomenological intuition if this theory is true.

I would like to point out that the phenomenological gap provides strong motivation for the challenge of the explanatory gap that C-Fiber Physicalism is claimed to suffer from. Levine (1983) argues that C-Fiber Physicalism leaves an explanatory gap in that the phenomenal character of the state of being in pain is left unexplained if the states of being in pain and of the c-fiber firing are identified. My claim that the phenomenological gap motivates the worry of the explanatory gap can be appreciated in the following way: if one attempts to bridge an explanatory gap by identifying X and Y, but there is a phenomenological gap between them, then we can reasonably question how Y could possibly be identical to X.[8] How, for instance, could a state without any phenomenal character give rise to a phenomenal state? On the other hand, if there is no phenomenological gap, then the worry of the explanatory gap can be significantly lessened though not completely removed. If there is no phenomenal feature that we would attribute to one but not to the other, then we don't need to be worried about how a state without any phenomenal feature could give rise to a state which is inherently phenomenal. This does not imply that if there is no phenomenological gap, then there is no

[8] Of course, the presence of the phenomenological gap does not imply the presence of the explanatory gap since our phenomenological intuition might be mistaken, i.e., Y might actually have the phenomenal feature or X might actually lack it.

explanatory gap. For one might still have some technical problems about explaining the causal role of the pain state in terms of the c-fiber firing activity.[9] In that case, the challenge of the explanatory gap is nonetheless greatly reduced. Hence, although the phenomenological gap is different from the explanatory gap, the former constitutes strong motivation for the latter.

Explaining the nature of phenomenology has been a formidable challenge for naturalists in philosophy of mind. Naturalists want to explain phenomenal character in terms of the natural order of cause and effect explicated by the physical sciences (Rupert 2008). By "physical sciences", I mean fundamental physics, chemistry, biology, neuroscience and arguably cognitive science.

Naturalists generally adopt a divide-and-conquer strategy to naturalizing mental states (Montague 2010). They divide the mental domain into two jointly exhaustive realms: the intentional states and the phenomenal states. Then they naturalize the intentional realm and the phenomenal realm separately. Since the mid-twentieth century, naturalists have made important progress in naturalizing intentionality. There are two leading approaches. One is informational semantics and the other one is naturalistic conceptual role semantics.[10] The former is already mentioned in §2.1 where I call it "the causal-informational version of externalism". The latter holds that the content of a mental state is determined by the psychological role played by the mental state, and hence is a naturalistic version of the narrow theory. Whichever approach you like, the core idea is to identify intentional states of the mind with its causal relations to external physical objects and properties (if you are an informational semanticist) or to other mental states (if you endorse conceptual role semantics).

Unfortunately, as many philosophers point out, inadequacies show up when we apply the same strategy to naturalizing phenomenology (Nagel 1974, Kripke 1980, Jackson 1982,

[9] In fact, some philosophers of the physical sciences point out that there seems to be an explanatory gap between chemistry and fundamental physics. See Hendry 2010.

[10] There are non-naturalistic versions of conceptual role semantics. E.g., Block argues (1986, p. 660) that Searle's theory (1984) explains the content of a state by a psychological role that takes intentionality as primitive, and hence is not naturalistic.

Levine 1983, Chalmers 1995, Mendelovici 2018). Without going into the details, I want to point out that their arguments are motivated by the phenomenological gap between phenomenal states and physical states. As we have seen, phenomenal states are simple and subjective, but the corresponding physical states typically are not simple, and not subjective. E.g., compare the phenomenal state of being in pain with the physical state of having the c-fiber firing. In addition, there seems no plausible physical theory that is capable of bridging the gap as there is for bridging thermodynamics and statistical mechanics. Finally, as has been pointed out, not only do we lack any physical means to identify them, but we also have a strong inclination to *keep them distinct* (Melnyk 2002, Papineau 2002, Molyneux 2011). To illustrate, in ordinary contexts, we would naturally accept the identification of two physical states e.g., *having a high temperature* and *having a high mean kinetic energy* given sufficient evidence whereas we would feel dissatisfied with the identification of a phenomenal state with a physical state—that is, we typically would feel that phenomenology *cannot* be any physical process. In essence, all these problems for naturalism can be summarized as motivated by the phenomenological gap, the phenomenological difference between phenomenal states and physical states.

§3 Experience: Strong Representationalism and Phenomenal Intentionality

In this section, I will introduce four key players in my dissertation: strong representationalism, the phenomenal intentionality theory, converse intentionalism, and the Inverted Earth thought experiment. The first two are theories of experience, and the third one is a common commitment of both theories. I will first discuss strong representationalism (§3.1), and then the phenomenal intentionality theory (§3.2). Along the way, I will explain why both theorists are committed to converse intentionalism. In §3.4, I will discuss the Inverted Earth thought experiment presented by Block (1990, 2003), which has played a pivotal role in the debate between the two theories.

Before we move on, let me clarify what is the intended notion of experience in my dissertation. By "perceptual experience" ("experience" henceforth), I mean a kind of

intentional phenomenal mental state such that (i) the proper tokening of that state requires the functioning of a sense organ, e.g., eyes, ears, and (ii) the proper tokening of that state does not require highly intelligent conceptual/linguistic capacities, e.g., an infant or a dog incapable of speaking a language is still capable of tokening an experience.[11] Notice that condition (i) does not require that when tokening a visual experience S, John has eyes. Rather, it only requires that if S were *properly* tokened, John would have eyes. Hence it is still possible for John to token a visual experience even if condition (i) is true. This point is important because philosophers of experience generally agree that a brain in a vat is capable of tokening a visual experience.

What kinds of entities could be the contents of an experience? Some philosophers of experience, e.g., Byrne (2009), argue that the contents of an experience can only be the "low-level" properties (or in Speaks' (2015) terms, the "sensible qualities") of mid-size objects. For example, we can see the *shape* or the *color* of an object, but not its age or gender (if it has one). In addition, we can experience the weight of a *dog,* but not that of an atom or of a mountain since they are either too small or too big. In contrast, Siegel (2014) holds that experiential contents are richer than what Byrne thinks. Since my arguments later are independent from the debate between Byrne and Siegel, I will only assume that experiential contents include but may not be limited to the low-level properties of mid-size objects.[12]

§3.1 Strong Representationalism

To understand strong representationalism ("SR" hereafter), let's start with the question of how naturalists attempt to fill in the phenomenological gap. I want to stipulate from now on that by "a naturalistic theory in the philosophy of experience" (or "naturalism"

[11] Some philosophers believe that cognitive states, e.g., thoughts, can be a kind of experience. Since my dissertation is exclusively about perceptual experience, I leave it open in my dissertation as to whether cognitive states can be experiences or not.

[12] There are other highly related issues: Could experience have singular content? Could Fregean sense be part of experiential content? Is experiential content propositional? If not propositional, how could experiential content have a veridicality condition? Although all these questions are very interesting, since, to repeat, my arguments are independent from them, I will not discuss them in my dissertation. See Brogaard 2014a, b.

hereafter), I mean any naturalistic theory that attempts to fill in the phenomenological gap between the phenomenal character of an experiential state and the physical state that realizes (or is identical to) that state.[13] As Montague (2010) says, naturalists in the philosophy of experience attempt to fill in the phenomenological gap via a two-step strategy. Typically, they first naturalize experiential intentionality by endorsing informational semantics, i.e., they argue that experiential content is determined by the causal-informational relation between an experience and the physical property in the external environment. Then, at the second step, they naturalize phenomenology by arguing that the phenomenal character of an experience is identical to or co-supervenes with the intentional content of that experience.

How exactly do naturalists argue that experiential phenomenology is identical to or co-supervenes with experiential content? One way is to begin with a very plausible phenomenological observation, called "the transparency of experience". (Harman 1990, Dretske 1995, Tye 2000, Byrne 2009). The observation is that when we introspect our experiences, we invariably experience the properties we are directly aware of as properties of external objects, not as properties of experiences. Metaphorically, experience is transparent to the extent that, like perfectly clear glass, we don't experience its intrinsic properties. We always "see through" it to the properties of external objects. As Harman puts it,

> When Eloise sees a tree before her, the colors she experiences are all experienced as features of the tree and its surroundings. None of them are experienced as intrinsic features

[13] According my stipulation of the idea of the phenomenological gap in §2.2, several theories do not count as naturalistic. For example, if a theory identifies the pain state and the c-fiber firing state, and explains why we attribute a phenomenal feature to the former but not to the latter by showing some computational limitations to our concepts of PAIN and C-FIBER, e.g., by showing that there is no coherent and well-defined algorithm that can process both our concepts of C-FIBER and PAIN (cf., Molyneux 2011), then this theory is not naturalistic despite the fact that it entails that phenomenal states are physical states. For the phenomenological gap still remains, i.e., we still would attribute some phenomenal feature to one but not to the other. Instead of filling in the gap, what this theory does is in fact showing that this gap can never be filled in. Accordingly, given this stipulation, some theorists who call themselves "naturalists" will not be taken as naturalists in my dissertation, e.g., Loar's naturalism by appealing to the phenomenal concept strategy (1997), Horgan's error-theoretic naturalism (1994), Pautz's primitivism (2009), Montague's Brentanian naturalism (2016; cf., Kriegel 2011), Chalmers' naturalistic dualism (2017), and Mendelovici's methodological naturalism (2010, 2018; cf., Mendelovici and Bourget 2014). This stipulation is simply a methodological constraint for delimiting the scope of my project. I don't have any particular view about how naturalism should be understood in the philosophy of mind. For a similar distinction, see Chalmers' type A physicalism vs. type B physicalism (1997).

> of her experience. Nor does she experience any features of anything as intrinsic features of her experience. And that is true of you too….When you see a tree, you do not experience any features as intrinsic features of your experience. Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree…(Harman 1990, p. 39).

As Molyneux (2009) summarizes it, the foundational thesis of transparency is:

> *Transparency*: When we introspect our phenomenal character, we experience the properties we are directly aware of as properties of external objects.

According to Siewart (2004), by inference to the best explanation (IBE), transparency leads those informational semanticists to endorse the identity thesis that phenomenal character is one and the same as representational content. Here is a reconstruction of the argument from transparency to the identity thesis:

1. When we introspect our phenomenal character, we experience the properties we are directly aware of as properties of external objects. (Transparency)
2. The identity thesis explains why transparency holds.
3. If the identity thesis is false, then experience is in general misleading.
4. By 3, any explanation of transparency entailing the negation of the identity thesis entails that experience is in general misleading.
5. Any explanation of transparency which entails that experience is in general misleading is inferior to one which doesn't.
6. By 2, 4, 5, the identity thesis is the best explanation of transparency.
7. Therefore, by 6, IBE, the identity thesis is justified.

To appreciate premise 2, consider a person who asserts a proposition. A straightforward explanation of why she does that is (i) that the proposition is true, and (ii) that the person has the inclination to tell truths. Analogously, a straightforward explanation of transparency is that (i*) the phenomenal is identical to the intentional, i.e., the identity thesis is true and (ii*) experience is reliable. By related reasoning, we can also understand why premise 3 seems plausible. If the identity thesis is false, i.e., the phenomenal is not identical to the intentional, but by transparency, experience tells us that they are, then experience is misleading. Premise 5

also seems plausible. Any explanation entailing that experience is misleading owes us an explanation of how we as a species could survive while experiencing the world incorrectly. It thus seems that the identity thesis is the best explanation.

However, there are many counter-examples to this naïve identity thesis. E.g., experiments involving blindsight patients show that they can token visual representational states without phenomenal character (Bermúdez 2014). So not all representational content is identical to phenomenal character. Hence, we need to modify the identity thesis by restricting phenomenal character to be identical to representational content "that meets certain further conditions" (Tye 2000, p. 45; Siewert 2004, Mendelovici 2010, 2018).[14] This identity thesis entails three supervenience principles:

> *Weak intentionalism*: Necessarily, experience tokens having the same contents have the same phenomenal characters.
>
> *Converse intentionalism*: Necessarily, experience tokens having the same phenomenal characters have the same contents.
>
> *Strong intentionalism*: Weak intentionalism and converse intentionalism are true.

It should be noted that intentionalism by itself is independent of naturalism or informational semantics. As we have seen, it is justified by transparency and the identity thesis. Naturalists can exploit this to fulfill their project. For once we are convinced, based on phenomenological observations, that phenomenology is identical to intentionality, then given that intentionality is naturalizable through informational semantics, phenomenology is also naturalizable.

Strong representationalism as a naturalistic theory of phenomenal consciousness (also known as "reductive representationalism" in Siewert 2004) is the package view that the

[14] There are more counter-examples to the naïve identity thesis than the blindsight experiments. E.g., Dretske (1995) points out that the (absolute) spatial location of an object cannot be identified with phenomenology. As he says, a radar cannot distinguish whether the airplanes it detects are at New York or Chicago. Thus, Tye (2000) explicitly says that the kind of content that can be identified with phenomenology has to satisfy certain conditions (see §3.2 below). I thank Prof. Mendelovici for bringing up this point to me.

phenomenal is identical to the intentional, and the intentional is to be understood in terms of informational semantics. It entails *that strong intentionalism and informational semantics are both true*.[15] Since converse intentionalism is entailed by strong intentionalism which is in turn entailed by SR, SR thus entails converse intentionalism. This is the view defended in Dretske 1995, Lycan 1996, 2000/2015, Tye 1998, 2000, 2009, Byrne and Hilbert 2003, Byrne and Tye 2006, and Byrne 2009. They all presuppose an informational theory of experiential content. Dretske's own semantics is stated in Dretske 1981, 1986, 1988. Lycan endorses evolutionary teleosemantics which is advocated by Millikan (1984, 1989, 2009; cf., Artiga 2013), and which is similar to Neander's (2017). Tye's view has undergone several revisions: sometimes he explicitly holds non-evolutionary optimal-functioning theory (1998), while sometimes he also proposes a non-teleological view which as he says bears an affinity to Fodor's (1990, 2010) causal asymmetric dependence theory (Tye 2000). Finally, in his later book (2009), Tye seems to take evolution to play a critical role.

§3.2 Phenomenal Intentionality

By "phenomenal intentionality theories" ("PIT" hereafter), I mean a contemporary narrow theoretic research program of mental intentionality which started in the late twentieth century. The main advocates are Siewert (1998), Horgan and Tienson (2002), Loar (2003), Kriegel (2003, 2007, 2011, 2013a, 2013b), Horgan, Tienson and Graham (2004), Chalmers (2004, 2006), Pautz (2006a, 2006b, 2008, 2013), Bourget (2010), Mendelovici (2010, 2018), Montague (2010, 2016), Horgan and Graham (2012), Farkas (2013), Potrč (2013), Mendelovici and Bourget (2014), Horgan (2014), and Bourget and Mendelovici (2016/2019). The unifying theme of PIT is, as Pautz (2013) puts it, "phenomenology first", in the sense that phenomenology give rise to other things in our mental life, e.g., intentionality.

The central thesis of PIT is that

[15] Accordingly, weak representationalism is the view that both weak intentionalism and informational semantics are true. Since my dissertation only concerns converse intentionalism, I will not discuss weak representationalism.

Weak PIT: There is phenomenal intentionality (Bourget and Mendelovici 2016/2019).

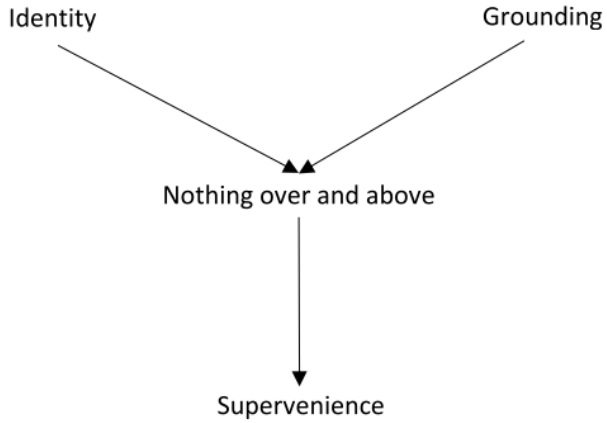
The idea of *phenomenal intentionality* refers to "the kind of intentionality that arises from consciousness" (Mendelovici and Bourget 2014, pp. 329).[16] This notion of "arising from" is diverse, with different theorists offering different accounts of the arising relation, e.g., identity, constitution, or grounding. For stylistic reasons, I will use "determined by" to cover all of these.[17]

Notice that the central thesis of PIT as applied to experiential content implies that experiential content is an instance of phenomenal intentionality which in turns entails converse intentionalism. This is because the relation of arising or determination in PIT is logically no weaker than metaphysical necessitation. Hence, no matter how PIT theorists disagree with each other about the determination relation, the central thesis of PIT entails converse intentionalism.

---

[16] Strong PIT theorists are skeptical of the idea that unconscious states could be genuinely intentional, e.g., Mendelovici 2018, Chapter 8.

[17] Prof. Mendelovici reminds me that for PIT theorists, the "arising-from" relation is a relation the intentional stands in to the phenomenal such that the intentional is *nothing over and above* the phenomenal. Strictly speaking, "nothing-over-and-above" is stronger than supervenience. E.g., the fact that I am identical to myself supervenes on the fact that I am at California since everything is necessarily self-identical. But the property of being self-identical is something over and above the property of being at California. (I thank Prof. Gilmore for this example.) Prof. Gilmore suggests to me that "nothing-over-and-above" is entailed by identity: necessarily if A=B, then A is nothing over and above B. However, the other way round does not hold: it is possible that A is nothing over and above B and A≠B. Furthermore, sometimes in the literature on PIT, the relation between the phenomenal and the intentional is formulated in terms of *metaphysical grounding*, which also entails "nothing-over-and-above". To sum up, we can have a diagram:



(Arrows mean entailing) In my dissertation, when I discuss PIT, by "determination", I mean any relation that is stronger than supervenience. It can mean identity, grounding or "nothing-over-and-above". Whatever it means, it entails converse intentionalism since the latter is only a supervenience claim. In addition, I choose the term "determination" for the sake of felicity. In some contexts, the intransitive verb "arising from" sounds infelicitous. Compare "the phenomenal *strongly/weakly* determines the intentional" vs. "the intentional *strongly/weakly* arises from the phenomenal".

Why do PIT theorists hold that experiential content is determined by phenomenology? As I have said, PIT is a narrow theory. A lot of classical reasons for the narrow theory can be carried over to supporting PIT. For example, one argument for the narrow theory is that (i) we should assign content to an intentional state in accordance with how the state is type-individuated and (ii) we should type-individuate a state by its internal causal role (§2.1). This is the narrow theorist's reply to the Twin-Earth case. Correspondingly, PIT theorists argue that for any two agents, if one is a phenomenal duplicate of the other, then since psychological role supervenes on phenomenology, the two agents' experiential states must have the same content no matter how far apart they are separated. On the other hand, with regard to the Frege case, the narrow theorist holds that the Hesperus thought and the Phosphorus thought, playing different psychological roles, have different contents even if they have the same referent. Analogously, PIT theorists use cases of color inversion, in which two experience tokens with different phenomenal characters are caused by the same external property, to argue that the content of an experience is not determined by its referent, but by its phenomenal character.

Other than converse intentionalism, I want to point out that there is another principle that most PIT theorists are committed to. To begin with, think about Frege's puzzle where two thought tokens have the same referent, but play different psychological roles (§2.1). As we have seen, narrow theorists hold that the two thought tokens have the different contents despite the fact that they have the same referent. Analogously, if two experience tokens are caused by the same property, but have different psychological consequences, then PIT theorists would also say that they have different contents. That is to say, they also hold a principle of the following sort:

> *Psychological supervenience*: For any two experiential state tokens, if their contents are the same, then they play the same psychological role.[18]

[18] Of course, this principle by itself is unreasonably strong. The versions endorsed by PIT theorists are suitably weaker than psychological supervenience formulated here. Since my aim here is simply expositional, it suffices

By saying two states "play the same psychological role", I mean that the states are governed by the same psychological functions. In fact, PIT theorists accept different versions of this principle (Loar 1988a, b, Horgan and Tienson 2002, Pautz 2006a, Mendelovici 2018).[19]

§3.3 A Final Remark on Strong Intentionalism

Strong intentionalism as I have formulated it, underspecifies what kind of content is targeted by those theorists. The only constraint I put is that it includes, but is not restricted to, experiential content. I intend it to be open to different understandings by different theorists. For example, Tye (2000), a prominent strong representationalist, calls the content he intends to capture by strong intentionalism "PANIC". Here are the characterizations of PANIC: first, the content is introspectable by the subject ("the content [of experience]…is poised to make a difference in *beliefs.*" See Tye 2000, p. 63). Second, the content is *abstract*. For example, my experience of a green object is about the universal, greenness, not about any particular trope. Otherwise, any two veridical color experience tokens would always have different contents. Third, the content is *non-conceptual.* (So "PANIC" is the acronym for "poised, abstract, non-conceptual intentional content".) On the other hand, PIT theorists typically take converse intentionalism to characterize *original intentionality* (Bourget 2010, Mendelovici 2018). As Mendelovici puts it, original intentionality is contrasted with derived intentionality "which is

for my purpose now to formulate it this way. We will have more detailed discussions on psychological supervenience in Chapter 3.

[19] Strictly speaking, psychological supervenience is both logically and theoretically independent from the central thesis of PIT. For one thing, some may think that it is metaphysically possible that there is a mental being who is capable of tokening phenomenal states like humans but whose mental states have no causal power so its mental life is consisted of a succession of causally inert phenomenal states. Consider a human and a mental being of this sort, and suppose that they are in experiential states with the same phenomenal character. By the central thesis of PIT, their experience tokens have the same content, but they don't have the same psychological consequence. For another, there may be some PIT theorist who endorses PIT not because of the Frege case. For example, Mendelovici's (2013, 2016) argument from the possibility of reliable misrepresentation for PIT is not motivated by the Frege puzzle. PIT theorists like Mendelovici can coherently remain neutral about psychological supervenience. (Nevertheless, I will argue in Chapter 3 that one of Mendelovici's arguments for PIT does presuppose something like psychological supervenience.) Hence, what I say in the main text could be made more precise in the following way: PIT theorists *who are motivated by the Frege case* are committed to psychological supervenience. Since as we shall see, the classic arguments for PIT are structurally similar to the Frege argument (and/or the Twin-Earth argument) one way or another, I shall set aside those versions of PIT which are independent from the Frege case (or from the Twin-Earth case) in my dissertation.

intentionality that derives from other instances of intentionality (Mendelocivci 2018, p. 22). E.g., linguistic contents are derived from the original contents of mental states such as experiences, beliefs, or Gricean communicative intentions. It thus appears that there is no single principle that I call "converse intentionalism" which is actually endorsed by all parties in the debate. However, *at a certain level of abstraction*, both parties agree that there is a special kind of experiential content (whatever more specific conditions those theorists put on it) that supervenes on phenomenology. And it is exactly my goal in my dissertation to show otherwise.

To summarize, it is interesting to observe that in the philosophy of experiential content, both externalism (i.e., SR) and the narrow theory (i.e., PIT) entail converse intentionalism though for different reasons. As naturalists, strong representationalists fill in the phenomenological gap by the transparency principle which leads to the identity thesis and hence to converse intentionalism. Following the theme of the narrow theory that content is assigned to match psychology, PIT theorists argue that content is determined by phenomenology which entails converse intentionalism. It is interesting to see that converse intentionalism is a common commitment of both theories, and furthermore as we shall see later, it will create problems for both theories too.

§3.4 Inverted Earth

Block's famous *Inverted Earth thought experiment* (1990, 2003) shows that there seems to an inconsistency between strong intentionalism and informational semantics. His original intention is to argue against strong intentionalism. However, many PIT theorists (Loar 2003, Chalmers 2004, 2006, Pautz 2006a, Montague 2016) take it as showing that informational semantics—and hence SR—should be rejected. In my view, what the Inverted Earth case shows specifically is that there is a tension between informational semantics and *converse intentionalism*, and hence it illustrates how converse intentionalism creates problems for SR.

Here is how it goes. Suppose on the far side of the galaxy, there is a planet, called "Inverted Earth", on which everything is exactly the same as it is on Earth except that the colors there are the inverted ones of the colors on Earth. Hence, *e.g.,* the sky on Inverted Earth is yellow. Suppose Rose, an Earthling, is transported to Inverted Earth without knowing. During transit, a pair of inverting lenses is inserted which change every ph-color to its inverted ph-color, e.g., exchange ph-green and ph-red, ph-yellow and ph-blue. Hence when she arrives there, everything she sees looks the same as it does on Earth, i.e., she has experiences with the same phenomenal character. As Block points out, when Rose looks up into the sky on both planets, her ph-blue experience tokens are caused by light of different colors, one being blue light, the other being yellow light. So according to informational semantics, her experience tokens have the correspondingly different contents, despite having the same phenomenal character. This case shows that informational semantics is in tension with converse intentionalism. Hence SR is untenable.

§4 Chapter 2: Informational Semantics and Strong Representationalism

In this chapter, expanding on Block's case, I will present a more general argument showing that there is an integration challenge for SR, i.e., informational semantics and converse intentionalism do not integrate easily.

Two principles are important to my argument in this chapter:

> *Phenomenal internalism*: Necessarily, experience tokens realized by the same internal type have the same phenomenal character.
>
> *Causal contingency*: Possibly, experience tokens realized by the same internal type are normally caused by different external properties.

My integration challenge is divided into two parts. In the first part, I argue that if phenomenal internalism and converse intentionalism are assumed, then (i) necessarily, if two experience tokens are realized by the same internal type, then they have the same content. This is because both phenomenal internalism and converse intentionalism are claims about

supervenience. The former says that phenomenal characters supervene on internal states, and the latter claims that content supervenes on phenomenal character. Since supervenience is transitive, contents supervene on internal states, and thus (i) is true. In the second part, I argue that if informational semantics and causal contingency are assumed, then (ii) possibly, two experience tokens realized by the same internal type have different contents. For informational semanticists identify the content of an experience token with the normal cause of that token, and causal contingency implies that experience tokens realized by the same internal type could have different normal causes. Hence informational semantics and causal contingency together imply that two experience tokens realized by the same internal type could have different contents, and hence (ii) is entailed. Finally, notice that (i) and (ii) are logically contradictory to each other since the former implies that contents supervene on internal l states while the latter denies that. Therefore, summing together both parts of my argument, we get a contradiction.

What this argument shows in effect is that if phenomenal internalism and causal contingency are assumed, then informational semantics and converse intentionalism do not integrate. So the question for strong representationalists is: could they reasonably reject either phenomenal internalism or causal contingency? I will argue that they can't.

As I will argue, the reason why strong representationalists could not reject phenomenal internalism is that, to repeat, they are naturalists. Recall that their strategy to fill in the phenomenological gap between the phenomenal and the physical is to endorse a phenomenologically plausible principle, i.e., concerning transparency. The reason why their strategy is better than traditional physicalism is precisely because, whereas the phenomenal and the physical seem distinct, within the *phenomenological* context of the transparency argument, the phenomenal seems identical to the intentional (which in turn is plausibly reducible to the physical). Hence the strength of strong representationalism as a version of naturalism is its phenomenological plausibility. However, as I will show, rejecting phenomenal internalism brings about some phenomenologically implausible consequences, possibly more

implausible than traditional physicalism. Therefore, although rejecting phenomenal internalism might be a logically coherent option for strong representationalists, it is self-undermining to their own phenomenological motivations.

The next step in Chapter 2 is to argue that rejecting causal contingency is not reasonable for strong representationalists either. To be sure, Block's original argument is premised on phenomenal internalism and some *crude version* of informational semantics. To address the Inverted Earth case, some strong representationalists thus appeal to more sophisticated versions (§3.1). In this part of Chapter 2, I will first go over the literature on informational semantics, and point out that the primary concern of informational semanticists is to solve *the disjunction problem*, the problem that given that there are diverse properties that can cause the same mental state, informational semanticists have to have a criterion to single out one particular cause as the content of that mental state. And then I will argue that if causal contingency is rejected, then informational semantics becomes implausible because it would resuscitate the disjunction problem. Hence rejecting causal contingency is also self-undermining to strong representationalists since informational semantics is an integral part of SR.

One final point: as Lycan (2000/2019) summarizes, to reject Block's original argument, strong representationalists, in the literature, typically either reject phenomenal internalism or appeal to sophisticated versions of informational semantics. Hence, my strategy in Chapter 2 effectively covers all possible moves of strong representationalists in the literature, which should provide us good reason to conclude that informational semantics and converse intentionalism do not integrate easily.

§5 Chapter 3: Phenomenal Intentionality and Psychological Role

As I have said, PIT theorists have used the Inverted Earth case to show that SR violates converse intentionalism. People might thus reasonably think that converse intentionalism goes along well with PIT. In Chapter 3, however, I will show that, in conjunction

with converse intentionalism, some premises in the classic arguments for PIT are susceptible to counter-examples, and hence PIT may not be as well-founded as it seems.

To anticipate, given that PIT is a narrow theory, there are at least two standard types of arguments for the position. First, *the Fregean-style argument* for PIT parallels Fregeans' arguments by starting from some perceptual versions of Frege's puzzles in which the experiential tokens are caused by the same properties but have different phenomenal characters, and then establishing the existence of phenomenal intentionality. Second, *the Twin-Earth-style argument* proceeds from a perceptual version of the Twin Earth case in which the experiential tokens have the same phenomenal character but have different referents. Then the advocates of the Twin-Earth-style arguments argue that there is a special kind of content shared by both tokens, which, as the advocates argue at the end of the argument, is exactly phenomenal intentionality.

Let's start with the observation that converse intentionalism and psychological supervenience (§3.2) together entail the following principle:

> *Phenomenology/psychology supervenience* (PPS): Psychological role supervenes on phenomenology, i.e., every two experience tokens having the same phenomenal character play the same psychological role.

Call this principle "PPS" for short. Accordingly, I will call the kind of classic arguments for PIT that involves PPS "phenomenal-psychological supervenience arguments" or "PPAs". I will argue that PPS is false, and thereby show that PPAs are unsuccessful.

At beginning of Chapter 3, I will classify PPAs into three categories:

> *Type 1*: PPS is entailed by the conclusion of the arguments, e.g., Pautz 2006a; cf., Montague 2016.
>
> *Type 2*: PPS is an explicit premise in the arguments, e.g., Horgan and Tienson 2002, Loar 2003.
>
> *Type 3*: A significantly weakened version of PPS is entailed by the conclusion of the arguments, e.g., Mendelovici 2018.

(According to my analysis in Chapter 3, Type 1 arguments correspond to the perceptual variants of the Fregean-style argument for the narrow theory in §3.2 of this chapter. Similarly, Type 2 arguments correspond to the perceptual variants of the Twin-Earth-style argument. Type 3 arguments can be seen as a variant of Type 1 PPAs.) Next, I will present a counter-example to PPS by revising Block's own case. Therefore, if my argument succeeds, i.e., PPS is false, then the advocates of PPAs either are not entitled to their theory (because their arguments for the theory entail a falsehood) or are not entitled to their arguments (because their arguments are premised on or presuppose a falsehood).

We can get a brief synopsis of my counter-example to PPS by observing that some colors look like mixtures of two different colors. E.g., orange looks like the mixture of red and yellow, and purple looks like the mixture of red and blue. Some colors don't look like mixtures, e.g., red, green, yellow, and blue. We call the former kind "binary colors", and the latter kind "unitary colors". Consider a possible world where everything is the same except that the evolutionary history of human is different so that the phenomenal character of a unitary color and the phenomenal character of a binary color are switched E.g., yellow looks ph-orange and orange looks ph-red. Let Clone be the counterpart in that world of an actual human Jones and so Clone's ph-red experience is caused by orange objects.[20] I argue that the ph-red experiences of Jones and of Clone have different narrow psychological roles, and hence we have a counter-example to the conjunction of converse intentionalism and psychological supervenience. For example, imagine Jones sees a monitor that looks red to him. Since he is normal, that means that the monitor is illuminated with red pixels. On the other hand, Clone also sees a monitor that looks red to him. But since his spectrum is shifted from Jones', the monitor he sees is purple. It follows that the monitor is illuminated with red and blue pixels. Presumably, when both persons zoom in, Jones will continue tokening ph-red experiences while Clone will token

[20] Cf., Pautz 2006a. He argues that it is possible that humans evolve to have shifted spectra, and uses this case to argue against strong representationalism.

an experience which is partly ph-orange and partly ph-purple since according to his spectrum, red looks orange and blue looks purple to Clone. Accordingly, if they both have ph-red experiences, then if both imagine or simulate in their minds what would happen if they were to zoom in, one of them would token mental imageries with the same phenomenal character while the other one wouldn't (Chapter 3, §5). Hence they would have experiences with the same phenomenal character but different psychological consequences.

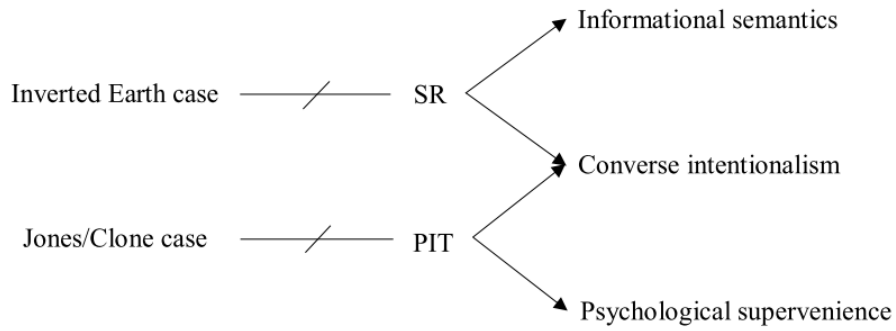
Let me summarize my argument in Chapter 3 by highlighting its relation to converse intentionalism. Crudely put, any advocate of PPAs can either (a) choose Type 1 arguments and/or (b) choose Type 2 (with Type 3 being a variant of Types 1). Since Type 1 arguments are premised on psychological supervenience, in conjunction with converse intentionalism, they entail PPS. Since converse intentionalism is entailed by the central thesis of PIT, the advocate can only avoid the Jones/Clone case by weakening psychological supervenience. However, as I will argue, it would render her argument too weak to support PIT. On the other hand, as I have pointed out, the strategy of Type 2 arguments is to first establish converse intentionalism by using PPS, and then to justify PIT by an inference to the only available explanation. But then again, I will show that if the advocate of PPAs wants to avoid my counter-example by weakening PPS, this weakened PPS is not strong enough to establish converse intentionalism in the first place. All in all, the classic arguments for PIT may not be as compelling as they are standardly received.

§6 Chapter 4: Psychointentionalism

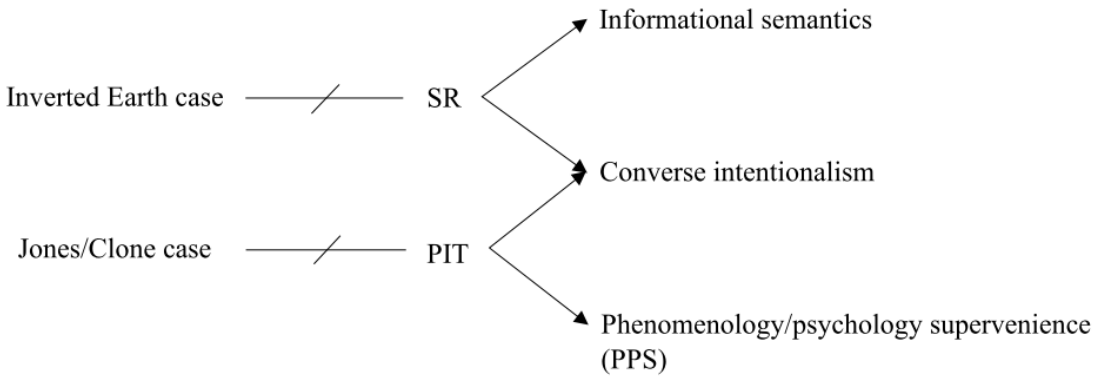
We can now see that converse intentionalism causes problems for both SR and PIT. For SR, the problem is that content changes with environments according to informational semantics while phenomenology is narrow according to phenomenal internalism. By endorsing converse intentionalism, strong representationalists make content invariant with respect to environments, putting it at odds with informational semantics. For PIT, the problem is that converse intentionalism is so strong that either (i) using a principle as strong as PPS or

psychological supervenience to argue for converse intentionalism/PIT is susceptible to the Jones/Clone case as a counter-example or (ii) using any weaker principle is not enough to establish converse intentionalism/PIT. Hence, paradoxically, both strong representationalists (externalists) and PIT theorists (narrow theorists) want to have converse intentionalism, but neither of them can. In Chapter 4, I will argue for a replacement of converse intentionalism that can satisfy strong representationalists' and PIT theorists' needs.

We can have a diagram to represent the dialectic:



Arrows mean entailing. Crossed lines mean contradicting. This diagram assumes that the PIT theorist considered is an advocate of Type 1/Type 3 PPAs. If the PIT theorist is an advocate of Type 2 PPAs, then the dialectic is:



In Chapter 4, I will only consider PIT theorists who advocate Type 1 PPAs. However, my replacement and arguments equally apply to Type 2 advocates.

I propose the following as my replacement of converse intentionalism:

> *Converse psychointentionalism*: Every two experience tokens having the same phenomenal characters *and psychological roles* have the same content.

Logically, converse psychointentionalism is weaker than converse intentionalism. Let's use "SR*" and "PIT*" to refer to the theories resulting from replacing converse intentionalism in SR and PIT with psychointentionalism.

In Chapter 4, I will show that in any case of color inversion, or any case involving shifting the spectrum around the hue circle, the color inverts' experiences are governed by different functions and hence their color experiences do not play the same psychological role. Thus, the antecedent of converse psychointentionalism is not satisfied, and hence these cases are not counter-examples to converse psychointentionalism. After showing that there is no counter-example by color inversion to converse psychointentionalism, I argue that the Inverted Earth case is not a counter-example to SR*. Nor is the Jones/Clone case a counter-example to PIT*. Notice that *my arguments are independent from any particular theory of experiential content*. My solution is a *psychological* solution since instead of giving a new meaning theory, I present a different account of the nature of experience as a psychological state. Since my arguments are actually compatible with each semantic theory, my account can fit into SR's and PIT's foundational programs respectively, which should increase the attractiveness of my account.

In the later part of Chapter 4, I argue that SR* and PIT* can satisfy strong representationalists' and PIT theorists' needs. In particular, SR* can satisfy strong representationalists' need to fill in the phenomenological gap, i.e., psychointentionalism coheres with transparency as well as strong intentionalism does. On the other hand, PIT* can also address various concerns of PIT theorists': First, in conjunction with psychological supervenience, converse psychointentionalism is not susceptible to the Jones/Clone example. Second, since converse psychointentionalism is weaker than converse intentionalism, PIT theorists can use appropriately weak principles to justify it. Thirdly, and most importantly, in conjunction with some plausible assumptions, converse psychointentionalism can explain the

intuition, insisted on by narrow theorists, that the experiences of a BIV and a normal person, if they are phenomenally the same, have the same content.

In sum, converse psychointentionalism performs well all the theoretical roles assigned to converse intentionalism by SR and PIT, and it is weak enough to avoid the problems to SR and PIT caused by converse intentionalism. If my argument is correct, then psychointentionalism should be recommendable to both SR and PIT.

§7 Summary

In this chapter, I give an overview of the general plan of my dissertation. In my dissertation, I will argue that (i) given SR's fundamental aim to fill in the phenomenological gap, its two pillars—informational semantics and converse intentionalism—do not integrate, (ii) converse intentionalism is too strong for PIT theorists to justify since their standard arguments for converse intentionalism are either unsound or too weak to justify it, and (iii) converse psychointentionalism is a good replacement for converse intentionalism because it can help strong representationalists and PIT theorists. I.e., for SR, it can fill in the phenomenological gap, and for PIT, it is both weak enough to argue for and strong enough to explain the BIV case. In §§2−3, I clarify the key notions in my discussions, including the notions of semantics and of phenomenology, and provide expositions of strong representationalism and the phenomenal intentionality theory, and why they both are committed to converse intentionalism. In §§4−6, I provide sketches of my arguments for claims (i) to (iii), which I will elaborate on in Chapters 2 to 4 accordingly.

Chapter 2

Informational Semantics and Strong Representationalism

Abstract

By informational strong representationalism ("strong representationalism" or "SR" hereafter), I mean a version of naturalistic philosophy of mind, which first naturalizes intentionality by identifying it with causation to physical properties, and then naturalizes phenomenology by identifying it with intentionality or making them co-supervene on each other (Montague 2010). In particular, SR will be taken as the conjunction of informational semantics and the intentionality-phenomenology identity thesis, the latter of which entails what I call "converse intentionalism", the principle that experiential content supervenes on phenomenology. Because of this identity thesis, SR enjoys some phenomenological plausibility which is absent from traditional physicalism of mind. However, in this chapter, I shall raise an *integration challenge* to SR by arguing that its two foundational components, informational semantics and converse intentionalism, do not integrate easily. I will also explore some strategies open to SR for addressing my challenge, and argue that by invoking those strategies, SR either loses its phenomenological plausibility or undermines informational semantics. I conclude that if my argument is correct, it provides us reason to search for new principles to replace SR's foundations.

§1 Introduction

By informational strong representationalism ("strong representationalism" or "SR" hereafter), I mean a version of naturalistic philosophy of mind, which first naturalizes intentionality by identifying it with causation to physical properties, and then naturalizes phenomenology by identifying it with intentionality or making them co-supervene on each other (Montague 2010).[1] Specifically, I will treat SR as the conjunction of the intentionality-

[1] In this chapter, I follow Montague's (2016) distinction between different versions of representationalism. What I call "intentionalism" later corresponds to her general representationalism, and what I call "(informational) strong

phenomenology identity thesis (or co-supervenience) and informational semantics.[2] Because of this co-supervenience, SR gains some strengths over traditional physicalism. However, in this chapter, I shall raise an *integration problem* for SR by arguing that its two foundational components, informational semantics and converse intentionalism, are in tension with one another, which gives us reason to search for new principles to replace its foundations.[3]

Given the above understanding, strong representationalism entails the following:

> *Converse intentionalism*: Necessarily, experience tokens having the same phenomenal characters have the same contents.
>
> *Informational Semantics*: a mental state type X means a property P iff P is the normal cause of X.

In addition, here is a principle which will play a crucial role in our later discussion:

> *Causal contingency*: It is possible that internal state tokens of the same experience type are normally caused by external properties of different types.

The idea of normalcy is left underspecified since I intend to make the principle subsume as many informational semantic theories as possible.[4]

---

representationalism" is roughly equivalent to her standard representationalism.

[2] Prof. Mendelovici suggests to me that strong representationalists might say that the phenomenal is *nothing over and above* the intentional. As I have said in Chapter 1, note 17, saying that A is nothing over and above B entails that A supervenes on B, but not the other round. So "nothing-over-and-above" is strictly stronger than supervenience. What they mean by "nothing over and above" is not completely clear. Is it identity? Or is it grounding? As far as I know, Dretske, Lycan, Tye, and Bryne all endorse the identity view because of the transparency argument (Chapter 1, §3.1). Since my target is converse intentionalism which is an implication of SR, whether it is formulated as the identity view, the grounding view, or simply the co-supervenience view, I will ignore the difference.

[3] Since the mid-twentieth century, naturalists have made important progress in naturalizing intentionality. There have been two major approaches, internal causal role semantics, and informational semantics. Following Dretske (1981), and Fodor (1990), in this chapter, I take informational semantics to be the more promising approach to naturalizing semantics, and I will focus on that theory rather than internal causal role semantics in the later discussions. A view could be developed that resembles SR, but incorporates the internal role semantics, instead of informational semantics. Such a view may solve the integration problem I propose in this chapter. Without a view in hand, I will not discuss this approach in this chapter.

[4] Fodor explicitly argues against incorporating the notion of normalcy into informational semantics given that informational semanticists intend to naturalize semantics which is normative (see in §3). His reason is that the notion of normalcy is also normative. Neander (2017) on the other hand argues that Fodor's own version of informational semantics cannot dispense with normalcy. For my purpose, the notion of normal cause simply means that which in virtue of causing the mental state constitutes the content of the state. This is a benign stipulation since my purpose here is to have a principle that subsumes as many informational semantic theories as possible, not to reduce the notion of normal to something non-normative. What's really important is whether the theories respectively are committed to causal contingency or not. And in §3, I will show that they are.

Given these principles, let me formulate my main argument:

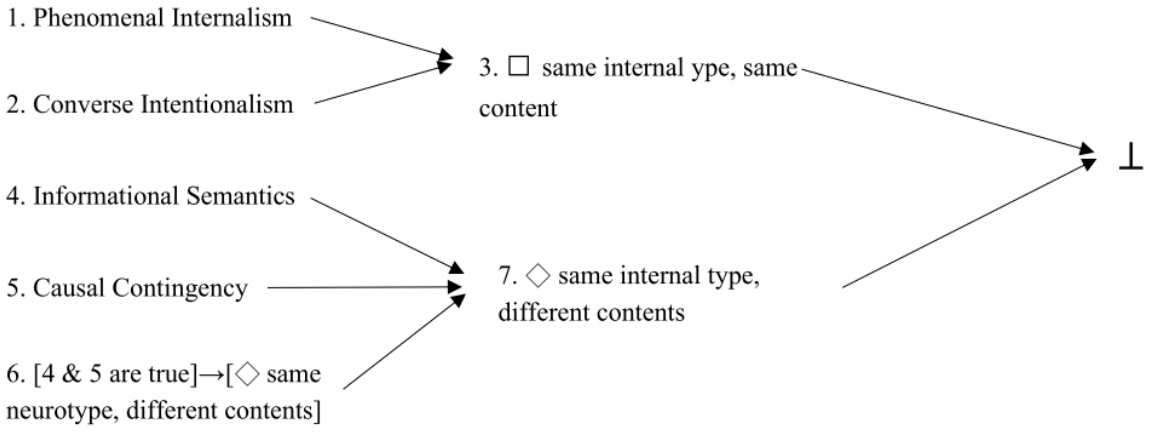
Sub-Argument A:

1. Necessarily, two experience tokens realized by the same internal types have the same phenomenal characters (phenomenal internalism).
2. Necessarily two experience tokens having the same phenomenal characters have the same content (converse intentionalism).
3. By 1, 2, necessarily two experience tokens realized by the same internal types have the same content.

Sub-Argument B:

4. A property is the content of an experience token iff it is the normal cause of the internal type which realizes the experience token (informational semantics).
5. It is possible that two experience tokens realized by the same internal types are normally caused by different external properties (causal contingency).
6. If causal contingency and informational semantics are true, then it is possible that two experience tokens realized by the same internal type have different contents (this is true by the formulations of causal contingency and of informational semantics).
7. By 4, 5, and 6, it is possible that two experience tokens realized by the same internal type have different contents.[5]

Please see the diagrammatic representation of this below (diverging arrows mean *separately entailing*, and converging arrows mean *jointly entailing*). Notice that 3 and 7 are contradictory to each other. Notice also that 6 is analytically true because according to informational



[5] There is an alternative formulation of my argument. E.g., an alternative formulation of premise 4 is that a property P is the content of an experience token *e* iff P is the normal cause of the internal type N such that there is an internal state token *x*, *e* is realized by *x*, and *x* is a token of N. Other premises can be adjusted accordingly. If one insists that realization is a relation between an experience token and an internal state token, then this alternative formulation is recommended. If we allow that an experience token is realized by an internal type, then the formulation in the main text can be accepted. Differences in the formulations would not affect the essential of my argument. I thank Prof. Gilmore for this point.

semantics, the content of a mental state is its normal cause, and hence two state tokens have the same content iff they have the same normal cause, and because according to causal contingency, it is possible that two tokens have different normal causes. It follows that it is possible that two mental tokens have different contents.

This argument effectively shows that phenomenal internalism (premise 1), converse intentionalism (premise 2), informational semantics (premise 4), and causal contingency (premise 5) are inconsistent. While I don't deny that there might be a way to make them coherent via reformulations without losing the essentials of each, in this chapter, I will argue that it is no easy task. I call this "the integration challenge".

In §2, I will provide the background for discussion, and explain why strong representationalism, given its two foundational principles, informational semantics and converse intentionalism, possesses strengths over traditional physicalism. Along the way, I will discuss why rejecting phenomenal internalism is not an attractive option for strong representationalists. In §3, I will provide a brief overview of informational semantics, and argue that informational semantics is implausible if causal contingency is rejected. This implies that rejecting causal contingency is not attractive to strong representationalists either. In §4, I will attempt to anticipate how strong representationalists might apply informational semantics to solving the challenge, and I will argue that these attempts are unsuccessful. All in all, if my argument is correct, we need to reject converse intentionalism, and, with it, strong representationalism in its current form. If something resembling strong representationalism to be salvaged, the new theory must replace converse intentionalism with some appropriately repaired principle.

Conventions: I will use capital letters for mentioning mental concepts. Terms with the prefix "ph-" denote phenomenal characters, e.g., ph-red is the phenomenal character she has when she sees red objects. Finally, "A→B" designates the nomic law that property A causes property B.

§2 Strong Representationalism and Causal Contingency

§2.1 The Naturalization Project

Naturalism in the philosophy of mind is a research program which argues that there is "nothing nonphysical, immaterial, or otherwise unnatural, i.e., nothing that cannot, at least in principle, be understood in terms of the natural order of cause and effect described by the physical sciences." (Rupert 2008) Here, the physical sciences range from fundamental physics all the way up to cognitive science. Naturalism has become a predominant position in contemporary philosophy of mind. Naturalists generally adopt a divide-and-conquer strategy to explain mental phenomena: they divide the whole mental domain into two parts, the intentional and the phenomenal, and then try to conquer both subdomains respectively.

Unfortunately, naturalists encounter formidable challenges when they try to naturalize phenomenology. Objectors (Nagel 1974, Kripke 1980, Jackson 1982, Levine 1983, Mendelovici 2018) base their arguments on phenomenological observations, showing that phenomenal states have phenomenal characters, what-it-is-like-ness which is absent in, or resist being explained in terms of, physical states. Moreover, it is hard for naturalists to get around our intuitive resistance to attribute phenomenal features to physical states. To illustrate, before astronomers showed that Phosphorus was identical to Hesperus, we attributed some feature to one without attributing it to the other, e.g., *being the brightest star in the morning*. After they are identified, we accept the identification without resisting attributing that property to Hesperus. In contrast, if scientists found the strong correlation between the pain state and the c-fiber firing state, and proposed that they were identical, we would still find that identification questionable because we would still resist attributing any phenomenal feature to the c-fiber firing state (Melnyk 2002, Papineau 2002, Molyneux 2011). Since we intuitively attribute phenomenal features to phenomenal states but resist attributing them to physical states, I will say that there is "a phenomenological gap" between phenomenal states and physical states.

As I pointed out in Chapter 1, §2.2, the phenomenological gap provides strong motivation for arguments against naturalism. Consider the first astronomer who identified Phosphorus and Hesperus, and a skeptic who rejected his identification by pointing out that we do not attribute the property of being the brightest star in the morning to Hesperus. We would typically think that her rejection is under-motivated. In contrast, Nagel's bat argument (1974), Jackson's Mary argument (1982), and Mendelovici's argument (2018) from the mismatch between phenomenal states' and physical states' superficial characters are intuitively forceful, and they all motivate their arguments from some particular phenomenal feature that we attribute to phenomenal states and resist attributing to physical states. On the other hand, Levine (1983), and Chalmers (1995) motivate their arguments by *modalized* phenomenological gaps. In particular, we would attribute the property of being *necessarily* phenomenal to phenomenal states, and would resist attributing it to physical states. As Levine says, the statement that pain is the firing of c-fibers sounds inherently contingent. Again, this phenomenological gap provides the fulcrum to leverage Chalmers' zombie argument and Levine's argument from the explanatory gap. All in all, the classic challenges to naturalism in my view are motivated by the phenomenological gap, and thus unless they can somehow fill in the phenomenological gap, naturalists' hope seems bleak.

Nevertheless, naturalists may still be able to achieve their aims by identifying the phenomenal character of a mental state with its intentional content. In particular, many people find the following phenomenological principle regarding perceptual experience very plausible:

> *Transparency*: When we introspect our phenomenal characters, the only things we are directly aware of are the objects represented as in the environment, and the properties they are represented as having.[6]

As Siewart (2004) points out, by inference to the best explanation, transparency leads many people to endorse the identity thesis that at least in the experiential cases, "phenomenal

[6] See Harman 1990, Molyneux 2009, Speaks 2009, 2015, Tye 2009.

character is one and the same as representational content that meets certain further conditions" (Tye 2000, p. 45, Speaks 2015, Mendelovici 2018). This identity thesis entails three supervenience principles:

> *Weak intentionalism*: Necessarily, experience tokens having the same contents have the same phenomenal characters.
>
> *Converse intentionalism*: Necessarily, experience tokens having the same phenomenal characters have the same contents.
>
> *Strong intentionalism*: Weak intentionalism and converse intentionalism are true.

It should be noted that intentionalism by itself is independent of naturalism or informational semantics. As we have seen, it is justified by transparency and the identity thesis. Naturalists can exploit this to fulfill their project. As we know, the main obstacle to naturalizing phenomenology is the phenomenological gap. However, once we are convinced, based on phenomenological observations, that phenomenology is identical to intentionality, then given that intentionality is naturalizable through informational semantics, phenomenology is also naturalizable.

Strong representationalism is the package view that strong intentionalism and informational semantics are both true. This is the view defended in Dretske 1995, Lycan 1996, 2000/2015, Tye 1998, 2000, 2009, Byrne and Hilbert 2003, Byrne and Tye 2006, and Byrne 2009. They all presuppose an informational theory of experiential content. Dretske's own semantics is stated in Dretske 1981, 1986, 1988, 1995. Lycan endorses evolutionary teleosemantics which is advocated by Millikan (1984, 1989, 2009; cf., Artiga 2013), and which is similar to Neander's (2017). Tye's view has undergone several revisions: sometimes he explicitly holds non-evolutionary optimal-functioning theory (1998), while sometimes he (2000) also proposes a non-teleological view which as he says bears an affinity to Fodor's (1990, 2010) causal asymmetric dependence theory ("CAD" hereafter). Finally, in his later book (2009), Tye seems to take evolution to play a critical role (2009).

Despite the objection I raise in this chapter, I find strong representationalism to be one of the most promising theories of what I call "the Naturalization Project", which refers to the research program of naturalizing all mental phenomena by the divide-and-conquer strategy mentioned in the beginning of this section. They first account for intentional properties as causal relations to physical properties, and thereby naturalize the intentional domain. Then without reducing experiential phenomenology to any physical state directly, they identify experiential intentionality with it. That way, they not only get around the phenomenological gap but also honor the phenomenological data of transparency. Strong representationalism thus has three virtues at the same time: it is scientifically respectable, metaphysically naturalizable, and phenomenologically plausible.[7]

§2.2 Phenomenal Internalism

As I said in §1, the following four principles do not integrate easily: informational semantics, converse intentionalism, causal contingency, and phenomenal internalism. In §2.1, I have explained why the former two are the two foundational pillars of SR. In this subsection, I will discuss why *prima facie* causal contingency creates a problem for SR, and why rejecting phenomenal internalism is not attractive to SR.

To illustrate how causal contingency creates a problem for SR, I would like to revisit Block's famous "inverted earth argument" (1990, 2003). Suppose on the far side of the galaxy, there is a planet, called "Inverted Earth", on which everything is the same as is on Earth except

[7] I take Mendelovici's argument (2010, 2018) against SR from the mismatch problem as showing that the putative progress made by SR to fill in the phenomenological gap is at best illusory. Surely, there is no phenomenal gap in identifying the phenomenal and the intentional. However, if the intentional is identified with the information-carrying relation, then by the transitivity of identity, the phenomenal is identical to the information-carrying relation. But then as she points out, there are differences with regard to the characters between the phenomenal and the information-carrying relation. Thus, in my terms, the phenomenological gap is still open. Assessing Mendelovici's argument goes beyond the scope of my dissertation. And I am not claiming that SR has already filled in the gap. All I want to argue in this section is that (a) SR is more phenomenologically plausible than traditional physicalism (§2.1), and (b) if strong representationalists want to solve my integration challenge by rejecting phenomenal internalism, then they would lose their advantage over traditional physicalists (see §2.2). Anyone convinced by Mendelovici's argument may directly reject (a) in the first place (so we don't need to bother thinking about my integration challenge at all). In that case, we can say that Mendelovici's argument is an *external* challenge to naturalism whereas my integration challenge is internal. The two projects can work in parallel, and need not be taken as incompatible.

that the colors there are the inverted ones of the colors on Earth. Hence the roses on Inverted Earth are green, and the leaves are red. Suppose Rose, an Earthling, is transported to Inverted Earth without knowing that she is transported. During transit, a pair of inverting lenses is inserted to her which would change every color to its complementary color, i.e., the lenses invert Rose's spectrum. Hence when she arrives there, everything she sees looks the same as she does on the earth. As Block points out, when Rose looks up into the sky on both her ph-blue experiences are caused by lights with different colors, one being blue light, the other being yellow. Right after Rose arrives at Inverted Earth, when she looks up the sky, intuitively, we would think that the content of her experience is blueness. However, as Block argues, after living on that planet many years later, when she looks the sky again, the content of her ph-blue experience is yellowness since her experience now is normally caused by yellow light. Given that the contents of Rose's experiences on Earth and Inverted Earth years later are different, strong representationalism is false.

To stay consistent with converse intentionalism, strong representationalists may choose to reject phenomenal internalism (Harman 1987, Dretske 1995, Lycan 1996, Tye 2009). This position allows for the possibility that experiences are realized by the same internal type, but have different phenomenal characters due to the difference in their contents. Between the first ph-blue experience token Rose has on Inverted Earth and the first experience she has whose content is yellowness, she presumably sees the sky many times and tokens many experiences. As Block plausibly argues, these experiences should be indistinguishable. For the colors are inverted with respect to colors on Earth, so Rose's psychological states and behaviors are functionally symmetric to the inverted Earthlings. E.g., both use the term "red" for the color of roses although inverted Earthlings mean green while Rose means red right after her arrival. But since the content of the experience eventually changes, to stay consistent with converse intentionalism, phenomenal externalists have to say that the phenomenology also changes. This implies that during the time between the first experience token on the planet and the first token

having yellowness as content, the phenomenal characters of Rose's experiences of the sky gradually change from ph-blue to ph-yellow, presumably passing through different phenomenal hues on the way.[8] This consequence is phenomenologically implausible (Tye 1998, Bartlett 2008).[9]

Of course, phenomenal externalists can bite the bullet and reply that phenomenological implausibility does not imply falsity, but this reply is self-undermining to strong representationalists. For as I have said in the previous subsection, a supposed virtue of strong representationalism is its phenomenological plausibility: unlike traditional physicalist theories, it respects our phenomenological intuitions. If however in order to maintain physicalism, phenomenal externalists are allowed to violate our phenomenological intuitions and assert the implausible possibility here, why can't traditional physicalists, e.g., C-Fiber Physicalists defy our intuitive resistance of attributing phenomenal features to physical states, i.e., the phenomenological gap, and claim that the firing of c-fibers is (necessarily) phenomenal? In my view, allowing for those phenomenological consequences of phenomenal externalism is much more phenomenologically implausible than denying the phenomenological gap.[10] Phenomenal externalism thus seriously undermines the motivation

[8] Which phenomenal hues? The externalist has no way to say. They would have to be ones between, and therefore different from, ph-blue and ph-yellow, and hence not reflections of the colors in the environment. Thus phenomenal externalists must undercut phenomenal externalism by attempting to solve the problem this way. I thank Prof. Molyneux for raising this point to me.

[9] See Tye 1998, 2000 for detailed discussions on those phenomenological implausible consequences (cf., Speaks 2015). One of those is that ph-blue experience and ph-yellow experience would become indistinguishable to Rose.

[10] Prof. Gilmore reminds me that strong representationalists may think that the phenomenological gap is weightier than phenomenal internalism, so if the choice is between denying the phenomenological gap and denying phenomenal internalism, they can justifiably deny the latter. This is a reasonable reply, on which I have three remarks however. First, strong representationalists, at least those philosophers I discuss in this chapter, do not explicitly address this worry I raise. Neither do they have a clear way to measure the weights of phenomenological intuitions. Any justification for denying phenomenal internalism instead of the phenomenological gap so far is under-developed. Second, the phenomenological intuitions violated may be weightier than it first appears. For example, consider the phenomenal character of Rose's visual experience of the sky which gradually changes from ph-blue to ph-yellow. Let's ask this question: what is the phenomenal character during the transition? Could it be ph-red? No! Because by strong intentionalism, if it is ph-red, then that means that the content of Rose's experience is the color red. But according to informational semantics, it means that her experience is caused by red, which cannot be right because the sky is not red. Hence by the same token, during the transition, the phenomenal character of her experience of the sky cannot be any of the ph-colors (other than ph-blue and ph-yellow). It follows that (*) during the transition, her visual experience of the sky has an *indeterminate* phenomenal character. In this case, I have to admit that (*) is not conceivable to me. Thus phenomenal externalists respect our intuitive resistance

for representationalism.[11] Phenomenal internalism is the only viable choice for strong representationalists.

Some people might object to my claim that phenomenal externalism undermines the motivation for representationalism in the above paragraph.[12] To elaborate on this objection, let me distinguish two versions of representationalism: the content-based version and the vehicle-based version (Thompson 2008). So far I have only talked about the content-based version, i.e., phenomenology is identical with content. In contrast, "the vehicle-based view says that phenomenology is identical to the property of representing p where p is the content" (Thompson 2008, p. 401). As Thompson argues, the vehicle-based view implies that phenomenology is a property of experience while the content-based view implies that it is a property of external objects. If so, then given the content-based view, phenomenal externalism just falls out naturally. Since all the representationalists, e.g., Dretske, Tye, I discuss in my dissertation endorse the content-based view, I should not hold them accountable for failing to meet phenomenal internalism.

In response, I have three replies. First, strictly speaking, the content-based view does not imply that phenomenology is a property of external objects. All it says is that phenomenal characters are identical to contents, i.e., the way we characterize an experience-type phenomenally is identical to the way we characterize it by its content. By itself, the content-

---

of attributing phenomenal characters to physical states (owing to the phenomenological gap) by attributing an inconceivable phenomenal character to a human! I am not sure what we gain here outweighs what we lose. At least, it seems to me that phenomenal externalism is as problematic as traditional physicalism. Finally, even if phenomenal externalists can somehow justify their claim that the intuitions they respects are weightier than the ones they violate, this view is not without deep problems. Since my aim is to provide an *integration challenge* to SR, not to show that SR is inconsistent, it already suffices for my project to expose those internal tensions within SR without having to show that they are unsolvable.

[11] Perhaps, phenomenal externalists can explain away the implausible consequences by some postulations. For example, they might explain the consequence that ph-blue and ph-yellow are indistinguishable throughout by arguing that Rose's memory system is malfunctioning on the inverted earth. (See note 8) But this is irrelevant. This issue is not whether externalists can have a coherent account of this implausible consequence. Rather, it is about the fact that externalism implies a *phenomenologically implausible* consequence. As we have seen, strong representationalism in conjunction with phenomenal externalism fares no better than traditional physicalism in respect of phenomenological plausibility.

[12] I thank Prof. Zoe Drayson for raising this worry to me.

based view does not say anything about phenomenal externalism. Second, it is not the case that all representationalists who hold the content-based view embrace phenomenal externalism *whole-heartedly*. For example, Michael Tye in his early works (1998, 2000) strongly criticizes phenomenal externalism. Indeed, in my view, he sharply points out some of the most implausible consequences of phenomenal externalism. Finally, regardless of representationalists' actual motivations, I only want to point out that embracing phenomenal externalism fares no better than embracing traditional physicalism in terms of phenomenological plausibility. To naturalize phenomenology, which is *the* fundamental concern of representationalists, phenomenological plausibility is what matters. Thus, if a strong representationalist theory of phenomenal consciousness solves the integration challenge by incurring another phenomenological implausibility, it only captures the letter of representationalism as a naturalistic theory of mind, not its spirit.[13] (As we shall see in Chapters 4 and 5, I will provide a new representationalist account that is both immune to the integration challenge and compatible with phenomenal internalism, and hence fully captures the spirit of representationalism.)

§3 Informational Semantics and Causal Contingency

Recall that the integration challenge for SR is that the following four principles do not integrate easily: (a) informational semantics, (b) converse intentionalism, (c) phenomenal internalism and (d) causal contingency. I have explained why strong representationalists are committed to (a) and (b). Now I have explained why rejecting (c) is not attractive for strong representationalists (§2.2). In this section, I will argue that rejecting (d) is not attractive either. In §3.1, I will provide an overview of informational semantics, and its ultimate concern to explain how misrepresentation is possible (called "the disjunction problem" in the literature). I shall follow Rupert's (2008), and Mendelovici's (2013) classification of informational

[13] According to my stipulation of naturalism (Chapter 1, §3.1), a naturalistic theory should fill in the phenomenological gape and hence take phenomenology seriously.

semantics into three groups: Dretske's indicator semantics, Fodor's CAD theory, and the Millikan-Neander teleosemantics. In §3.2, I will argue that all three versions of informational semantics are implausible if causal contingency is rejected.

§3.1 Informational Semantics

Let's start with the idea of information. How does any state carry information about another state? Think about an ordinary example, since the temperature of the air causally covaries with the size of the mercury, a thermometer carries information about the temperature of the ambient air. Dretske (1981) thus suggests that the nature of information can be explained in terms of causation.

In Dretske's view, *meaning* is different from information. The usual examples of meaning are public languages and mental concepts. According to Dretske, the most important difference between information and meaning is that the latter, not the former, is *normative*: a sentence can be true or false, and a concept can be correctly applied or misapplied. For example, we can fool a person into thinking of a sheep as a cow by disguising the sheep, so she *mis*applies the concept COW to the sheep. On the other hand, we cannot fool a thermometer. If we put it into hot water, we cannot say it *mis*represents the temperature of the air since in this case it carries information about the water, not the air, which is still accurate. Similarly, since there exist the law that the property of being a cow→COW and the law that the property of being a horse-in-the-dark→COW, if we don't distinguish between the laws, then the concept COW means *cow or horse-in-the-dark*. The basic idea of informational semantics then is that because many different external properties, mediated by various nomic laws, may play a causal role in the tokening of a given mental state, what we need is a criterion picking out whichever among those laws is the one that determines the content. To specify the criterion is what is called "the problem of misrepresentation" or "the disjunction problem" in the literature. Different theorists thus offer different accounts for this criterion. I will call this criterion "the content-determining condition".

Schematically, all informational semantic theories thus have the following structure:

> For any mental state X, any property P, X means P iff (i) there is a nomic law L that P→X, and (ii) L satisfies the content-determining condition.

Thus, different informational semanticists differ in their accounts of (ii). I will lay out in detail their accounts of (ii) in §3.2.

§3.2 Causal Contingency

As we have seen, an informational semantic theory has two components: (i) the nomic law connecting the content and the state, and (ii) the content-determining condition. Accordingly, to reject causal contingency (i.e., to claim that the relevant causality is necessary) implies that (i′) the law L connecting the content and the state is *necessary*, and that (ii′) L *necessarily* satisfies the content-determining condition. In §§3.2.1−3.2.3, I will show that if (ii′) is true, then informational semantics would become either implausible or vacuous.

§3.2.1 The Content-Determining Condition: Fodor

Let's start with Fodor's CAD theory. Fodor's (1990, 2010) CAD theory says that a causal law L satisfies the content-determining condition iff other laws connecting the state to other properties are *asymmetrically dependent* on L, i.e., if L did not hold, then others wouldn't either, but L would still hold even if others didn't. For example, both the property of being a cow and the property of being a horse-in-the-dark cause the concept COW. But the law L1 that horse-in-the-dark→COW is asymmetrically dependent on the law L2 that cow→COW: if L2 didn't hold, then L1 wouldn't and if L1 didn't hold, L2 would still hold. That's why COW means cow, and why when someone applies COW to a horse in the dark misapplies the concept. In our terms, according to Fodor, a law L that P→X satisfies the content-determining condition iff other laws connecting some property Q to X are asymmetrically dependent on L.

How can we know that it is a contingent matter whether a law L satisfies the content-determining condition as specified by Fodor? We can gather evidence from our daily life. It is easy to observe that one's cognitive system changes over time. A law may satisfy the content-

determining condition at one time, but become asymmetrically dependent at a later time. Or it may be asymmetrically dependent at one time, but then satisfies the content-determining condition later. For instance, consider the visual experience of a 20-year-old agent Lisa with the phenomenal character she experiences when an "E" mark is 2 meters away from her. Suppose that Lisa's visual capacity functions normally. Call the neural state type realizing this experience "N". Let's call the law that the property of E's being 2 meters away→N "L1". I claim that L1 satisfies CAD's content-determining condition. That's because *ceteris paribus* in her 20s, given the curvatures of her lens, any optic trick (e.g., in the Ames room illusion)[14] to the effect that E causes N where E is not 2 meters away is asymmetrically dependent on the law L1 that the property of E's being 2 meters away→N. I.e., if L1 did not hold, then the trick would not work, and L1 would still hold even if the trick did not occur. So L1 is the content-determining law when Lisa is 20 years old.

Now consider Lisa at 40 years old. Since the curvatures of the lenses in the eyes of a human change over time, this in turn causes the focal distance to vary too. For everything to look as clear and discernable as it was, it has to be farther away from Lisa. For example, the E mark has to be 4 meters away for it to appear as clear and discernable as it was. However, in this case, the E mark appears smaller than it was. Suppose one day, the E mark is enlarged correspondingly so that the (enlarged) E mark being 4 meters away appears to be phenomenally indistinguishable to Lisa's experience when she saw it 2 meters away in her 20s.[15] So, an E mark 2 meters away causes Lisa to token N in her 20s iff an enlarged E mark 4 meters away causes her to token N in her 40s. Let's call the law that the property of E's being enlarged and 4 meters away→N "L2". Suppose Lisa is adapted to the environment in accordance with L2. Then by reasonings similar to the one in the last paragraph, we can conclude that L2 satisfies

[14] One optic trick is that designer uses the brighter light to illuminate the farther corner of a room, and uses the darker light for the nearer corner so that the two corners appear to be equally far away from the observer.

[15] In reality, human perception of size is influenced by the angle of the eyes when they are focused on the object. To simplify our discussions, let's assume that Lisa is a monocular agent.

CAD's content-determining condition. So L2 is the content-determining law when Lisa is in her 40s. Therefore, whether a law satisfies the content-determining condition changes through time under the CAD framework, which implies that whether a law satisfies the content-determining condition under the CAD framework is a contingent matter.

Perhaps, strong representationalists may argue that in the above example, there is a law L3 that is asymmetrically depended on by the other laws all through Lisa's life and hence L3 is the content-determining law throughout. If this suggestion works, my argument for causal contingency is undermined. Notice that whatever the law may be, it had better deliver the correct verdicts about the contents of Lisa's experience tokens in her 20s and 40s. Thus L3 has to be a law that E's instantiating a particular property→N, and whatever that property is, it is co-extensive with the property of *being 2 meters or 4 meters away*.

However, this suggestion simply brings us back to the disjunction problem (§3.1). Recall that the disjunction problem is the ultimate concern of informational semanticists. It is to solve this problem that they propose those sophisticated epicycles to account for our intuition that content could be non-disjunctive. Yet, L3 actually renders CAD impotent to solve the disjunction problem. To see this, suppose L3 is asymmetrically depended on by other laws, and Lisa in her 40s tokens N. However, due to some optic trick, the E mark is actually 2 meters away. According to the supposition, Lisa correctly represents the distance of the mark, and thus correct representation becomes trivial.

Alternatively, strong representationalists may rebut my argument by saying that the content of Lisa's experience in her 40s is still the property of being 2 meters away, not the property of being enlarged and 4 meters away, and so the law that persists throughout Lisa's life is actually L1. Thus, as they may say, there is no disjunction problem in this case. Furthermore, they may support their claim by showing that there is a sense in which L2 is asymmetrically dependent on L1: according to the developmental trajectory of human physiology, if L1 didn't hold for Lisa in her 20s, then L2 would not hold in her 40s whereas if

L2 didn't hold in her 40s, L1 still would still hold in her 20s.

However, I would like to point out that there are two reasons why this reply doesn't work. First, recall that we are working under Fodor's CAD framework to see if it helps SR. Fodor (1987, p. 109) distinguishes two kinds of asymmetric dependency, the synchronic and the diachronic. The former involves two laws instantiated by the subject at the same time while the latter involves laws instantiated by the subject at different times. He insists that a law is content-determining iff other laws asymmetrically depend on it *synchronically*, not diachronically. Hence, though it may be right in that L2 asymmetrically depends on L1 given the developmental trajectory of human physiology, it is irrelevant since this dependence is not synchronic.

Second, it is independently intuitive that the content of Lisa's experience in her 40s is the property of being enlarged and 4 meters away. Suppose after her 20s, the curvatures of Lisa's lenses start to become far-sighted gradually, which causes tiny errors. For example, when she stretches out her arm to get something in front of her, she sometimes slightly misses it because she does not stretch her arm far enough. Or she sometimes has the wrong grip aperture because she misjudges the size of the object. But assume that, as time goes on, each slight change is compensated for, her behaviors at 40 should be generally successful. At this stage, it is intuitive to say that the content of her experience is now the property of being enlarged and 4 meters away (Harman 1987). Indeed, if one asks Lisa how far away the object is when she tokens the experience, she would answer: "4 meters." In the absence of counter-evidence, it is groundless to claim that her report is mistaken and insist that the content of her experience is still the property of being 2 meters away.

Why does Fodor insist that the content-determining law is only synchronically depended on by other laws? The answer is precisely to allow for the possibility of Lisa's case (or cases of that sort). If the non-synchronic dependence, e.g., L2's dependence on L1, is relevant to semantics, then Lisa in her 40s always visually misrepresents the distances and the

sizes of objects she sees even though her behaviors are generally successful and habitual.[16] Hence if our semantics under the CAD framework is to accommodate the intuition that Lisa in her 40s does not visually misrepresent the distances and the sizes of objects, then we should agree with Fodor that the only relevant kind of dependence is synchronic, which implies that causal contingency is true.

In sum, I presented Lisa's case in which the content of her experience changes through time as an instance of causal contingency under the CAD framework. Given that in her 20s, the content of Lisa's experience is the property of being 2 meters away, strong representationalists have two jointly exhaustive replies: (a) showing that L2 is not the content-determining law for Lisa in her 40s, e.g., by appealing to some non-synchronic notion of dependence, or (b) showing that there is a third law L3 which is asymmetrically depended on by L1 and L2 throughout her life, which delivers verdicts that match our intuitions about the contents of Lisa's experience at different ages. I argued that strategy (a) leads to the implausible conclusion that Lisa in her 40s misrepresents the distances and the sizes of objects, and is actually unsupported by Fodor's CAD, while strategy (b), in order to deliver the correct verdicts, brings back the disjunction problem.

§3.2.2 The Content-Determining Condition: Dretske

Dretske's indicator semantics (1986, 1988) argues that a mental/neural state means a property iff it has the *function* of carrying the information of that property. In Dretske's terminology, A *carries the information* of B iff A indicates B. What concerns us here is how a mental state acquires the function of indicating B. According to Dretske, it acquires the function through *conditioning*. Consider an animal which has a mental state C that indicates an external

[16] Fodor's own example focuses on a child learning the concept LION. Typically, the teacher would show the picture of a lion, and tells the child this is a lion. At this stage, the law L1* that a picture of a lion→LION is formed in the child. Then after the child knows the difference of an object and the picture of the object, the law L2* that lion→LION is formed. At this stage, intuitively, we think that the child successfully acquires the concept. As we can see, L2* is asymmetrically dependent on the law L1* in the diachronic sense. So if diachronic dependence is relevant to semantics, then the content of the child's concept LION is the picture of a lion, not the lion even though she can distinguish a lion and a picture of a lion, which is implausible.

features F and G. At first, at a few occasions, when seeing F, the animal tokens C and C causes the animal to do M, say, reaching out getting the F, which given its desire at that occasion satisfies its needs. After this pattern repeating many times, the causal link between C and M becomes reinforced. Then the property type C of the animal generally causes the property type of doing M. As Dretske says, "Once C is recruited as a cause of M….C acquires, thereby, the function of indicating F….C acquires its semantics, a genuine meaning, at the very moment when [the information] acquires an explanatory relevance". (1988, p. 84) In contrast, G is not explanatorily relevant, so C does not have the function of indicating G, i.e., C does not mean G. In our terms, Dretske's account of the content-determining condition can be stated in the following: the law L that P→X where X is a mental state of an agent S satisfies the content-determining condition iff there is a behavior M of S such that there is causal law L* that X→M, and L was involved in the historically relevant explanation of how L* was formed and reinforced in S. Again, we can explain why for a person to apply COW to a horse in the dark is incorrect since there is no behavior M of the person such that the law that horse-in-the-dark→COW was involved in the historically relevant explanation of why the law that COW→M was formed.

It is also easy to see that whether or not a law satisfies Dretske's content-determining condition is a contingent matter. To see this, consider a behavior M of mine in my 20s, i.e., my behavior of *stretching out my arm 20 cm ahead*. An object being 20 cm in front of me causes me to token a visual spatial experience N. In my 20s, I learn to do M when I token N if I want to get the object in front of me. Hence the law L1 that being 20 cm away→N is involved in the historical explanation of the law that N→M in me. I.e., L1 satisfies the content-determining condition in my 20s. Then when I am 40, it is the property of being 40 cm away that causes me to token N. Consider another behavior M′ of *stretching out my arm 40 cm ahead*. Again, I learn to do M′ when I token N if I want to get the object. So, the law L2 that being 40 cm away→N

is involved in the historical explanation of the law that N→M′ in me. I.e., L2 satisfies the content-determining condition in my 40s. Therefore, at two different temporal points, different laws uniquely satisfy the content-determining condition under Dretske's framework, which implies that under Dretske's framework, causal contingency holds.

Strong representationalists may say that this argument relies on the wide understanding of behavior, but there is also the narrow understanding of behavior. M and M′ can be type identified as *moving as directed by N.* Call this behavioral type M″. Consider the law L3 that *an object being 20cm or 40cm away*→N. Then strong representationalists may argue that there is a justifiable sense in which L3 is involved in the historical explanation of the law that N→M″, and hence L3 satisfies the content-determining condition in my 20s and 40s. Causal contingency may thus be refuted.

However, this strategy brings us back to the disjunction problem. If the strategy by narrowly understanding a behavior is allowed, then obviously many causal connections between mental representation and behavior would have multiple historical explanations by different laws between environmental factors and mental representations, and these representations have disjunctive contents. E.g., the strategy to refute causal contingency by identifying M and M′ as M″ implies that the content of my experience N is that *the object is 20 cm or 40 cm away*, which is disjunctive. As we can see, understanding behavior narrowly thus weakens the power of Dretske's account to discriminate which law determines the content and which doesn't, and hence makes it incapable of solving the disjunction problem.

§3.2.3 The Content-Determining Condition: Teleosemantics

Teleosemantics (Millikan 1984, 1989, 2009, Neander 2004/2012, 2017) is the theory that a subject's mental state X means the property P iff it is a *proper function* of X to carry the information of P iff *the law that P→X was formed and reinforced in the evolutionary history of the species to which the subject belongs*. (We can take the italicized clause as characterizing the content-determining condition.) For example, according to the standard teleosemantic

account, the reason why Oscar's WATER concept refers to H_2O while twin Oscar's WATER refers to XYZ is that in the evolutionary history of humans, the law that H_2O→WATER was formed but on the twin earth, in the evolutionary history of twin humans, the law that XYZ→WATER was formed. That explains why after Oscar (an Earthling) arrives at the twin earth, and applies his concept of WATER to XYZ, his application is incorrect since the law that XYZ→WATER was not formed in the evolutionary history of humans.

To see that causal contingency holds for the content-determining condition under the teleosemantic framework, consider a case discussed in Fodor's work (1990). As he says, frogs or toads (*Bufo Bufo*) dart on any black moving little dot at a certain velocity along a certain direction. In its natural habitat, black moving dots are mostly flies which provide nutrition to toads. If a toad is born in a laboratory in which all black moving dots are nutritious pellets, then teleosemantics implies that the toad would visually misrepresent these pellets as flies throughout its life. Furthermore, suppose scientists put a group of toads in the laboratory. Call them the first generation in the laboratory. Artiga (2013) argues that according to teleosemantics, several generations later, the content of the toad's visual experience when it sees a black moving dot is now a nutritious pellet, rather than a fly. The reason is that in the first generation, the law that a fly→the experience is formed and reinforced in the toads' natural habitat while in the later generation, it is the law that a nutritious pellet→the experience that is formed and reinforced in the toads' environment (i.e., the laboratory). Hence at different times, different laws satisfy the content-determining condition for the species, and causal contingency is true.[17] To generalize, teleosemantics allows the possibility that at the first generation, the neural state N is formed in virtue of the nomic connection L1 that P→N, but then several

[17] This case is given by Artiga (2013) according to his interpretation of Millikan's version of teleosemantics. As he points out, Neander's version denies the possibility of this case since according to Neander's theory, only sensible properties can figure in the content of experience. As we will see in §4.3, however, even Neander's theory would imply that the law satisfying the content-determining condition changes if we apply her theory to a revised version of inverted earth case. This means that even if strong representationalists incorporate Neander's teleosemantics rather than Millikan's, experiential content would still change since the law satisfying the content-determining condition changes.

generations later, N is preserved in virtue of the law L2 that Q→N. Hence causal contingency is true.

To deny causal contingency, strong representationalists have two strategies. First, (i) they can say that it is the law L3 that (P or Q) →N that satisfies the content-determining condition, and hence the law that satisfies the content-determining condition does not change within these generations. Second, (ii) they can argue that it is L1 that satisfies the condition since it is the very reason why the neural type N is formed. As we can now see, (i) only brings back the disjunction problem, and so I will not discuss it further.

The problem of (ii) is that there is no clear-cut environment which we could point to and justifiably claim that it is *this* environmental factor P in virtue of which N is adaptd and hence is to be identified with the content of N since natural selection in fact is a *gradual* process. Since to illustrate my point requires an evolutionary case involving a bigger time-scale, let's think of a different story.[18] Suppose one million years ago, a species of bacteria lived in the north pole which was their natural habitat, the bacteria had an internal sensory mechanism detecting the geomagnetic north, and oxygen was toxic to the bacteria. The survival value of the sensory mechanism is explained by the fact that surface water is oxygen-rich, and that as one approaches to the north pole, the bacteria's internal compass pointing to strict geomagnetic north must increasingly point down toward the ground or the seabed, away from the surface water. Hence the bacteria would successfully track the oxygen-free water by tracing the geomagnetic north. Suppose, in addition, the first token of bacteria's magnetic sensory state as a phenotype existed one million years ago in the north pole. Five hundred thousand years ago, they migrated to the south pole where there were iron sediments so that the sensory mechanism was fortified and stabilized. One hundred thousand years ago, they moved to places near the equator where there were non-iron magnetic ores, and the number of bacteria equipped with

[18] This case is borrowed from Dretske 1986.

the phenotype became predominant. In this case, any decision to pick one among the environments as *the* one where the phenotype is formed is arbitrary (cf., Neander 2004/2012). Hence it is difficult to say which environmental factor is the content of the phenotype, the geomagnetic field? The iron magnetic field? Or the non-iron magnetic field? In this same way, there is no non-arbitrary way to determine in which environment the neurotype N of a toad is adapted to and hence no non-arbitrary way to determined what content of the neurotype is selected for representing. Therefore, due to the arbitrary nature of strategy (ii), it is not helpful to strong representationalists.

To summarize, in §3.2, I argued that informational semantics is implausible if causal contingency is rejected. In §§3.2.1−3.2.3, I argued that all the three versions of informational semantics which are widely cited by strong representationalists are implausible if a law necessarily satisfies the content-determining condition.

§4 Informational Semanticists Meet the Inverted Earth

In response to the invert earth case, strong representationalists might accuse Block of presupposing the crude casual theory. If we apply a sophisticated version of informational semantics, can a counter-example of this sort still be constructed? In this section, I will evaluate some attempts to apply sophisticated developments of informational semantics to the inverted earth problem. I will show that either they fail to do what they are supposed to or even if they succeed, the inverted case can be adapted so that converse intentionalism is still violated.

§4.1 Fodor's Causal Asymmetric Dependence

Tye (1998) applies Fodor's CAD theory to argue that the content of Rose's ph-blue experience is yellowness no matter how many years she lives on the planet. To do so, one needs to show that the law that yellowness→ph-blue is asymmetrically dependent on the law that blueness→ph-blue. More specifically, one has to show that (1) at the nearest possible world at which the law that blueness→ph-blue does not hold, the law that yellowness→ph-blue does not

either; and that (2) at the nearest possible world at which the law that yellowness→ph-blue does not hold, the law that blueness→ph-blue still holds. To verify (1), consider the nearest possible world in which blueness does not cause ph-blue in Rose. Suppose blueness causes some other phenomenal character, say, ph-red in Rose. As Tye argues, since the function of the inverting lenses is to change every light wave to its complementary one, wearing the inverting lenses would make yellowness cause ph-red. So, the law that yellowness→ph-blue does not hold. To verify (2), think about the nearest world in which the law that yellowness→ph-blue does not hold in Rose. As Tye points out, this is exactly the actual world. But in the actual world, the law that blueness→ph-blue does hold in Rose. So the law that yellowness→ph-blue is asymmetrically dependent on the law that the law that blueness→ph-blue. Hence, the phenomenal character ph-blue does not mean the property yellow.

As it should be clear by now, Tye's application of CAD presupposes a non-synchronic notion of dependence (§3.2.1) since it involves two laws instantiated by the subject at different times. As I have argued, combining non-synchronic dependency with CAD leads to the implausible consequence that only the first law that is asymmetrically depended on by other laws in one's developmental trajectory satisfies the content-determining condition and hence determines the content of one's experience. On the other hand, if we combine the synchronic notion of dependence with CAD, then the law that blueness→ph-blue is asymmetrically depended on by other laws when Rose is on Earth while the law that yellowness→ph-blue is so when she is on Inverted Earth. Therefore, her ph-blue experience has different contents on different planets. Strong representationalism is thus falsified by the synchronic CAD. We can put the dialectic into a dilemma: asymmetrical dependence is either synchronic or not. If the former, then CAD implies that Rose's ph-blue experience has different contents on different planets, and if the latter, then CAD implies the implausible consequence that Lisa's experience always misrepresents the distances and the sizes of objects after she is 40 (§3.2.1). Overall, CAD is not helpful to strong representationalism.

§4.2 Dretske's Indicator Semantics

Adams and Dietrich (2004) suggest that Dretske' indicator semantics can be used to argue against Block's claim that the content of Rose's ph-blue experiences is yellowness no matter how long she lives on the planet (though this theory, as they point out, has hard a time solving the swampman problem). Since I am only concerned with the tension between causal contingency and converse intentionalim, it deserves a subsection to see whether Dretske's indicator semantics can resolve the tension or not.

As we have seen, Dretske's starting point is that a mental state carries information about many different features in the environment. The task of semantics is to provide a principle to pick one feature among the many as the content of the mental state. According to his indicator semantics, this principle is that a subject S's mental state X means F iff there is a behavior (type) M of S such that the law that F→X was involved in the historical explanation of why X causes M in S. According to this theory, it is indeed plausible that Rose's ph-blue experiences do not change content. To see that, the crux is the stipulation in Block's experiment that it is *functionally indistinguishable* to Rose whether she is on one planet or the other. If we could justify the assumption that functional indistinguishability implies behavioral sameness, then this functional indistinguishability entails that Rose's behaviors do not change, which in turn means that the causal relations between her color experiences and her behaviors never change before and after her traveling. This further implies that the historically relevant explanations of these causal relations between her color experiences and behaviors are the same, i.e., the same conditioning histories. So, the laws between the environmental features and the neural states, if they are involved in the historical relevant explanations of Rose's causations between experiences and behaviors on Earth, continues to be so on Inverted Earth. Therefore, the content of Rose's ph-blue experiences does not change.

This application of Dretske's theory to addressing the inverted earth problem is successful only if the claim that functional indistinguishability implies behavioral sameness is

true. However, the truth of the claim depends on whether the notion of behavior is understood narrowly or widely. When Rose's ph-red experience causes her to stretch out her arm, and to grasp the fruit in front of her, there are two ways to individuate her behavior. Individuated narrowly, her behaviors on Earth and Inverted Earth are of the same type which could be characterized as *grasping the object in front of her*. Individuated widely, her behaviors on the planets are type distinct. The one on Earth is characterized as *grasping the apple in front of her* while the other behavior on Inverted Earth is identified as *grasping the inverted apple in front of her*. Obviously, then, functional indistinguishability implies behavioral sameness only if behaviors are narrowly understood.

Unfortunately, as I have shown in §3.2.2, Dretske's indicator semantics becomes incapable of solving the disjunction problem if the narrow understanding of behavior is assumed. We can construct a dilemma here for strong representationalists. Behavior should be understood either narrowly or widely. If the former, then we have the disjunction problem, and if the latter, the application of indicator semantics to the inverted earth case is not successful. In sum, strong representationalism is still falsified by the inverted earth case if it is combined with Dretske's indicator semantics.

§4.3 Teleosemantics[19]

Ren (2016) argues that teleosemantics can successfully preserve converse intentionalism.[20] The reason is that according to teleosemantics, the content of a experience type of a subject is the information the carrying of which the experience type was selected for in the evolutionary history of the species to which the subject belongs. Rose is a human so the

[19] My argument in this section is inspired by Mendelovici's arguments (2013, 2016) with a thematic change. She argues that under the framework of informational semantics, especially teleosemantics, experiential content changes with time, but as she says teleosemanticists should allow for the possibility that content does not change. In contrast, my aim is to expose an internal integration problem for teleosemantic strong representationalists who are committed to endorsing informational semantics and denying causal contingency.

[20] Although what Ren has in mind is Dretske's teleosemantic theory (1995), the crux of his reasoning is still the idea that content is determined by the evolutionary history of the whole species, which historically comes from Millikan's and Neander's work in the 1980s. So I classify Dretske's 1995 work as a version of "the Millikan-Neander" teleosemantics.

content of her ph-blue experience is the information the carrying of which the experience type was selected for in the evolutionary history of humans, i.e., the color blue. Since no matter how many years Rose lives on Inverted Earth, she is still a human, and the content of her ph-blue experience is still the information the carrying of which the experience type was selected for *on Earth*. Hence the content of her ph-blue experiences is still blueness. Converse intentionalism is thereby preserved.

I don't think teleosemantics helps strong representationalists. To illustrate my point, let's change Block's story a little. Suppose there are no inverting lenses. However, on Inverted Earth, the atmosphere contains a special gas such that when an Earthling inhales it, it would invert the phenomenal characters the Earthling has when she lives on Earth. So, what the gas does is functionally the same as the inverting lenses. We may also assume that this effect is reversible: when an Earthling goes back from Inverted Earth to Earth, the inverting effects disappear quickly. Suppose a lot of Earthlings travel to Inverted Earth unknowingly. Eventually, they proliferate over many generations (assuming they do not marry any inverted Earthling). When one descendant, i.e., Rose, sees the sky on Inverted Earth, and tokens a ph-blue experience, what is the content? There are three laws that might determine the content of Rose's ph-blue experience, L1 that blueness→ph-blue, L2 that yellowness→ph-blue, and L3 that (blueness or yellowness)→ph-blue. As we have seen in §3.2.3, if L3, then the disjunction problem comes back. If L2, then strong representationalism is falsified since an Earthling and Rose have experiences with the same phenomenal characters, realized by the same neural state, but with different contents. Finally, L1 seems arbitrary since although Rose is a distant descendant of Earthlings, a lot of ancestors *closer* to her lived on Inverted Earth throughout their lives. Their neurophysiology (including Rose's) is actually preserved and reinforced in the environment on Inverted Earth. Thus it is more plausible that Rose's ph-blue experience is selected for representing yellowness (cf., Mendelovici 2013, 2016, Artiga 2013). If so, then by replacing the content change in Rose within a single lifetime with the content change in her

species over an evolutionary time, we can get an intergenerational version of the inverted earth problem, which is still formidable to teleosemantic strong representationalists.

§5 Conclusion

In this chapter, I argue that naturalistic informational semantics is in tension with converse intentionalism. Since strong representationalism is the conjunction of informational semantics and strong intentionalism, my argument shows that there seems to be an integration challenge to SR. In §1, I first present my main argument that phenomenal internalism, causal contingency, converse intentionalism, and informational semantics together imply a contradiction. In §2, I then clarify the nature of strong representationalism and explain why strong representationalism is the most promising implementation of the Naturalization Project, and why rejecting phenomenal internalism is unattractive to strong representationalism. In §3.1, I provide an overview of informational semantics. In §3.2, I argue that informational semantics is implausible if causal contingency is rejected. Finally, in §4, I re-examine strong representationalists' applications of informational semantics in the literature to solve the inverted earth argument. I argue that their applications are unsuccessful.

As I have said, my aim in this chapter is to raise an integration challenge to SR. If my argument is correct, SR needs new foundational principles to fulfill the Naturalization Project.

Chapter 3

Phenomenal Intentionality and Psychological Role

Abstract

This chapter presents a challenge to an influential type of argument for the phenomenal intentionality theory ("PIT" hereafter) in the philosophy of experience. One of the central theses of PIT is that experiential content is determined by the phenomenal character of experience. And the type of argument reviewed in this chapter involves the principle that no two experience tokens having the same phenomenal character play different psychological roles, which will be called "phenomenological/psychological supervenience" ("PPS" hereafter). In this chapter, I show that there exist counter-examples to PPS, which undermine those arguments for PIT that appeal to it. I first analyze several classic arguments for PIT in the literature, including Horgan and Tienson 2002, Pautz 2006a, Mendelovici 2018, and argue that these arguments all involve PPS. I then argue for the possibility of two experiences with the same phenomenal character that have different psychological roles. Lastly, I discuss and reply to possible objections.

§1 Introduction

The phenomenal intentionality theory ("PIT" hereafter) is a semantic theory of mental states, which considered one of the most promising theories of narrow content. The unifying theme of PIT is, as Pautz (2013) puts it, "phenomenology first", in the sense that other things derive from phenomenology, most especially, intentionality. Its central thesis is that

Weak PIT: There is *phenomenal intentionality* (Bourget and Mendelovici 2019).[1]

Its central notion—*phenomenal intentionality*—refers to "the kind of intentionality that arises

[1] Cf., Kriegel: "Phenomenal intentionality is a basic kind of intentionality and functions as a source of all intentionality" (Kriegel 2013b, p. 5); Mendelovici and Bourget: "All intentionality is phenomenal intentionality" (Mendelovici and Bourget 2014, p. 330); Mendelovici: "All intentional states arise from phenomenal consciousness" (Mendelovici 2018, p. 86).

from consciousness" (Mendelovici and Bourget 2014, pp. 329).[2]

This notion of "arising from" is in fact diverse, with different theorists offering different accounts of the relation between intentionality and phenomenology, e.g., identity, constitution, or grounding. I will use "determined by" to cover all of these. This is simply for stylistic reasons. I do not make any substantive assumption about the relation, except that it implies supervenience.[3] Furthermore, the notion of phenomenal intentionality as applied to experiential content entails that

> *Converse intentionalism*: Content supervenes on phenomenology, i.e., every two experience tokens having the same phenomenal characters have the same content.[4]

As we will see later, converse intentionalism plays a crucial role in some arguments for the claim that conscious experience is an instance of phenomenal intentionality, i.e., for

> *Experiential PIT*: Conscious experience is an instance of phenomenal intentionality, i.e., the content of an experience is determined by the phenomenal character of the experience.

Notice that Experiential PIT entails Weak PIT. If PIT theorists' argument for Experiential PIT fail, then their grounds for claiming weak PIT are undermined.[5]

In this chapter, I will examine a kind of argument for Experiential PIT/converse intentionalism regarding experiential content. In particular, the kind of arguments I am

---

[2] Cf., Kriegel: Phenomenal intentionality "is a kind of intentionality…that is grounded in phenomenal character" (Kriegel 2013b, p. 5); Mendelovici: "Intentionality that arises from phenomenal consciousness is…called phenomenal intentionality" (Mendelovici 2018, p. 85); Bourget and Mendelovici: "Phenomenal intentionality is intentionality that is constituted by phenomenal consciousness" (Bourget and Mendelovici 2016/2019).

[3] The notion of determination could be understood as a disjunction of those intended relations, i.e., X determines Y iff X is identical to Y or X constitutes Y or Y is grounded by X…etc.

[4] Cf., Horgan and Tienson: "There is a kind of intentional content,…,such that any two possible phenomenal duplicates have exactly similar intentional states vis-à-vis such content" (Horgan and Tienson 2002, p. 524); Chalmers' view that: "there is plausibly an entailment from experiential phenomenal properties to pure representational properties" (Chalmers 2004, p.158); "If two experiences share the same phenomenological content, then necessarily they share (a kind of) representational content" (Montague 2016, p.86). I call this principle "converse intentionalism" because in the literature of the representational theory of consciousness, the principle that same intentionality implies same phenomenology is called "weak intentionalism" and the principle that experiences have the same content iff they have the same phenomenology is called "strong intentionalism". There is so far no standard name for the converse of weak intentionalism. See Lycan 2019.

[5] In fact, the classic works argue for weak PIT by first establishing Experiential PIT and then using existential generalization. See Horgan and Tienson 2002, Loar 2003, Montague 2014, Mendelovici 2018.

targeting involves the following principle:

> *Phenomenology/psychology supervenience*: Every two experience tokens having the same phenomenal character play the same psychological role.

I will call this principle "PPS" for short. Accordingly, I will call the kind of arguments for Experiential PIT/converse intentionalism that involves PPS "phenomenal-psychological supervenience arguments" or "PPAs". I will argue that PPS is false, and thereby show that PPAs are unsuccessful.

Here is the plan. In §2, I will provide the background for discussion. In §3, I will discuss various versions of PPAs. In §4, I will present a banal counter-example to PPS. It shows the basic problem, but there are ways to answer it. In §5, I present another counter-example which does not inhere the problem that tarnishes the banal example.

Convention: Terms with the prefix "ph-" denote phenomenal characters, e.g., ph-red is the phenomenal character one has when one sees a red object under normal conditions.

§2 Phenomenal Intentionality and Its Opponents

By "phenomenal intentionality theories", I mean theories that aim to account for mental intentionality by appeal to phenomenology. Examples include Horgan and Tienson 2002, Loar 2003, Kriegel 2003, 2007, 2013a, 2013b, Horgan, Tienson and Graham 2004, Chalmers 2004, 2006, Pautz 2006a, 2006b, 2008, Mendelovici 2010, 2018, Montague 2010, 2016, Horgan and Graham 2012, Mendelovici and Bourget 2014, Bourget and Mendelovici 2016/2019.[6]

One motivation for PIT theories is based on a type of thought experiment in which an entity—e.g. a Cartesian ego or a brain in a vat (BIV)—is causally disconnected from the external environment but has experience with the same phenomenal charater as normal

[6] According to Kriegel (2013b), the first papers using "phenomenal intentionality" are Horgan and Tienson 2002, and Loar 2003 though Loar's had been circulating before Horgan and Tienson's. A notable precursor before the notion of phenomenal intentionality came about was Charles Siewert 1998.

humans. It is then argued that its experiences must also have the same intentional content (Siewert 1998, Horgan and Tienson 2002, Loar 2003, Kriegel 2013b). This type of thought experiment is designed to show that content follows phenomenology; that as long as an experiential state has the same phenomenology as a person's, it has the same mental content.

As I already mentioned, PIT is a *narrow content theory*. It entails that experiential intentional contents of mental states supervene on the internal states of the subjects who token them. This is because all PIT theorists agree on phenomenal internalism, the principle that phenomenology supervenes on internal states. Given phenomenal internalism, converse intentionalism, and the transitivity of supervenience, intentional content supervenes on internal states.

Narrow content theories contrast with wide content theories, which entails that contents do not solely supervene on internal states. A major species of wide content theories is *informational semantics* of experiential content. This approach understands intentionality in terms of the causal relation of experience to the external environment, and is endorsed by Dretske (1995), Lycan (1996), Tye (2000, 2009), Byrne and Hilbert (2003), Byrne and Tye (2006).

As I have said, my targets are PPAs (phenomenology-psychology supervenience arguments). Recall that by "PPAs", I mean those arguments that involve PPS (phenomenology/psychology supervenience). Since this principle plays a pivotal role in the later discussions, let me clarify it. It says that two experience tokens having the same phenomenal character play the same psychological role, i.e., they are governed by the same psychological functions. By a "psychological function", I mean a mapping primarily from psychological state types to other types. For example, inferences are psychological functions. They map a number of beliefs to another belief. We can say that a state token *e* is governed by a psychological function if and only if the type of *e* is in the domain or the range of the function. Finally, if we agree with the intuitive idea that experience tokens are type-identified by their

phenomenal character, then for any two tokens having the same phenomenal character, a function governing them would map the two tokens to state tokens of the same type. For example, consider the function of color inference. If you and I token ph-red experiences, this inference function maps both experience tokens to belief tokens of the type with the content that the relevant object is red.

Since there are many PPAs, it is useful to have a taxonomy of them. In the first type of PPA, PPS (i.e., phenomenology/psychology supervenience) is, or is entailed by, the conclusion of the argument. A standard PPA of this type motivates PIT by comparing two subjects with different spectra. I will call this "the argument from shifted spectrum", and illustrate it with the particular version in Pautz 2006a (cf., Montague 2016). In the second type of PPAs, PPS appears as a *premise*. This type usually proceeds by considering a pair of phenomenal duplicates, in which case I call it "the argument from phenomenal duplicates". I will focus on the version in Horgan and Tienson 2002, (cf., Loar 2003). In the third type of PPAs, PPS is significantly weakened to the effect that it does not require experiences having the same phenomenology to play exactly the same psychological role. Rather it only requires they play psychological roles *of the same kind*. I will discuss Mendelovici's (2010, 2018) argument as an example for this type.

Notice that as I mentioned, PPS is entailed by Pautz's argument. It is because his argument relies on the following principle:

> *Psychological supervenience*: Every two experience tokens having the same content play the same psychological role.

The difference between PPS and psychological supervenience is that the former says that psychological role supervenes on phenomenal character whereas the latter claims that it supervenes on *content*.

Since Pautz's argument relies on psychological supervenience (same content, same psychology), and since its conclusion entails converse intentionalism (same phenomenal

character, same content), the position as a whole entails PPS (same phenomenal character, same psychology). That's why I claim that PPS is entailed by his argument. Mendelovici's argument, meanwhile, relies on a significantly weakened version of psychological supervenience, which she calls "psychological involvement", and hence the conclusion of her argument entails a correspondingly weakened version of PPS. To summarize visually, please see the following diagram.

Type 1 (Pautz)	Type 2 (Horgan and Tienson)	Type 3 (Mendelovici)
1 …..	1 …..	1 …..
…..	…..	…..
n Psychological supervenience	n PPS	n Weakened Psychological supervenience
…..	…..	…..
…..	…..	…..
∴ converse intentionalism	∴ converse intentionalism	∴ converse intentionalism
∴ PPS		∴ Weakened PPS

All in all, if my work succeeds, then PPS is false and the advocates of PPAs either are not entitled to their theory (because their arguments for the theory entail a falsehood) or are not entitled to their arguments (because their arguments are premised on or presuppose a falsehood). While I do not deny that there may be other kinds of arguments for Experiential PIT/converse intentionalism that do not involve PPS, given that PPAs are wide-spread and common in the literature, I believe my work in this chapter poses a serious challenge to Experiential PIT.

§3 Phenomenal-Psychological Supervenience Arguments

In this section, I analyze the various PPA arguments. The argument from shifted spectrum is discussed in §3.1, the argument from phenomenal duplicates is in §3.2, and Mendelovici's argument that methodologically presupposes PPS is in §3.3.

Before we discuss those PPAs, I would like to revisit Block's (1990) famous "inverted earth thought experiment". Although its original conclusion is not about PIT, it

convincingly shows that there is a strong tension between informational semantics and converse intentionalism.

Block's inverted earth case has been considered one of the decisive arguments that reversed the trend toward informational semantics in the semantics of experiential content. In fact, many PIT theorists rely on it to reject informational semantics (Loar 2003, Chalmers 2004, 2006, Montague 2016). Here is how it goes. Suppose on the far side of the galaxy, there is a planet, called "Inverted Earth", on which everything is exactly the same as it is on Earth except that the colors there are the inverted ones of the colors on Earth. Hence, *e.g.,* the sky on Inverted Earth is yellow.[7] Suppose Rose, an Earthling, is transported to the inverted earth without knowing. During transit, a pair of inverting lenses is inserted which change every ph-color to its inverted ph-color, e.g., exchange ph-green and ph-red, ph-yellow and ph-blue. Hence when she arrives there, everything she sees looks the same as it does on the earth. As Block points out, when Rose looks up into the sky on both planets, her ph-blue experience tokens are caused by light of different colors, one being blue light, the other being yellow light. So according to informational semantics, her experience tokens have the correspondingly different contents, despite having the same phenomenal character. This case shows that informational semantics violates converse intentionalism.

After showing the tension between informational semantics and converse intentionalism, we now can proceed to see how PPAs work.

§3.1 Argument from Shifted Spectrum

This type of PPA leads to a conclusion that is, or entails, converse intentionalism. It is premised on *psychological supervenience* which in conjunction with converse intentionalism entails PPS (see §2).

I will illustrate this type of PPA with the argument in Pautz 2006a (also see Montague

[7] Let us assume whatever theory of color is needed to make such claims.

2016[8]). Pautz's argument assumes that tokening an experience with the phenomenal character ph-red, for example, is identified with standing in a *representational relation* to red, which he calls "sensory awareness". Given this, we have two mutually exclusive and jointly exhaustive hypotheses that (a) sensory awareness is wide (informational semantics) or (b) it is narrow. He has reason to think that (b) implies Experiential PIT, (and thereby converse intentionalism). So to argue for Experiential PIT/converse intentionalism, all he needs is to argue against informational semantics.

Here is his thought experiment. Pautz argues that there is a possible world in which the earth is the same as it is in the actual world, but the evolutionary history of the human species is different such that the neural structure of human brains (specifically, the structure undergirding opponent-process of color vision) is wired to the environment differently. In particular, where Max in the actual world sees an orange object and has a ph-orange experience *e* his counterpart Twin Max in the possible world sees an orange object but has a ph-*red* experience *e**. By informational semantics, the property contents of *e* and *e** are whatever cause the experiences, and thus *e* and *e** are both experiences *of orange.* However, according to Pautz, they would cause different behaviors in Max and Twin Max *even though they have the same content*. This is counter-intuitive, and shows that informational semantics is objectionable.

To see why Max and Twin Max would behave differently, it is helpful to know the hue circle. Color scientists use the hue circle to represent the structure among colors. *Crudely* put, if we fix the color at the position of 12:00 on a clock to be red, then colors at 3:00, 6:00, 9:00 are yellow, green, and blue respectively. The colors at intervals between 12:00 to 3:00,

[8] Montague's (2016) argument takes converse intentionalism as a principle that needs to be respected. She invokes a different version of psychological supervenience: sameness of the contents of two color experiences implies sameness of which color is attributed by the subjects, i.e., which color is believed by the subjects to be instantiated. Given informational semantics, subjects with shifted spectra have different beliefs about which color is instantiated, and thus their experiences have the different contents even if their experiences have the same phenomenology. Since it violates converse intentionalism, Montagues concludes that informational semantics is false.

3:00-6:00, 6:00-9:00, 9:00-12:00 are, respectively, orange, chartreuse-green, cyan, and purple. Importantly, colors at 12, 3, 6, and 9 o'clock are *unitary*, i.e., they do not look like mixtures of other colors whereas all the other colors are *binary*, i.e., they look like mixtures. E.g., orange, which appears between 12:00 and 3:00, looks like a mixture of red and yellow. Hence the effect of the spectrum shift conceived of by Pautz is that Max sees a unitary color iff Twin Max sees a binary,[9] and this leads to a difference in their classificatory behaviors. For instance, Max would classify orange (ph-orange) as binary, but Twin Max, experiencing ph-red, would not.

As we have seen, informational semantics implies that the experiences of Max and Twin Max in Pautz's case have the same content. Pautz's argument works insofar as this is the wrong result. But this can only be shown to be the wrong result if Pautz has an *independent* reason to show that their experiences have different contents. This is where Pautz offers a weak form of psychological supervenience (§2)[10]:

> If two actual or possible individuals have qualitatively identical color experiences, then they have the same color-related behavioral dispositions (2006a, p. 219).[11]

As we saw, Max and Twin Max do not have the same color-related behavioral dispositions since they have different classificatory dispositions. Pautz uses this to argue that Max's and Twin Max's experiences, though both caused by the same object, have different contents. This is how he refutes informational semantics and thereby establishes Experiential PIT and converse intentionalism.

Now we can have a summary of Pautz's shifted spectrum argument in the premise-conclusion form:[12]

---

[9] In reality, green is not diametrical opposite to red. The idealized model I describe here is attributed to Pautz by Byrne and Tye (2006). What really matters is that that humans could have evolved to have such a unitary/binary color-exchange. My argument later does not rely on the plausibility of the idealized model.

[10] Pautz's version is a weak form of psychological supervenience because it only says that *color-related* behaviors are the same if the contents are the same. It does not say *all* psychological roles are the same if the contents are the same, which explains why psychological supervenience is *strong*.

[11] Pautz identifies the qualitative color of an experience with its color content. Thus by "qualitatively identical color experiences", he means color experiences that must have the same content.

[12] This reconstruction of Pautz's argument is adapted from Byrne and Tye 2006.

1. Either (a) sensory awareness is identified with a wide relation, or (b) with a narrow relation.
2. If (b) is true, then Experiential PIT/converse intentionalism is true.
3. If (a) is true, then Max's and Twin Max's experiences have the same content.
4. Max's and Twin Max's experiences have different contents. (by Pautz's version of psychological supervenience)
5. (a) is false. (by 3, 4)
6. (b) is true. (by 1, 5)
7. Therefore, Experiential PIT/converse intentionalism is true. (by 2, 6)

In this argument, *psychological supervenience* justifies premise 4, and *converse intentionalism* is the conclusion. Since PPS is implied by the two, it is an implication of the argument.[13]

§3.2 Argument from Phenomenal Duplicates

I will illustrate the second type of PPA I wish to discuss using (Horgan and Tienson 2002. Cf., Loar 2003). This type generally starts off by comparing two phenomenal duplicates, uses some form of PPS to argue that the experiences of the twins have the same content, and thereby establishes converse intentionalism. Then, by arguing that only Experiential PIT (as opposed to informational semantics) can explain converse intentionalism, Experiential PIT is justified by *inference to the only available explanation* ("IOE" hereafter).

Here is Horgan and Tienson's argument. They first ask you (the reader) to conceive an arbitrary phenomenal twin *A* of yourself. *A* might be a disembodied Cartesian ego, or a BIV disconnected from its external environment. Suppose you token an experience *e* of seeing a picture hanging crooked, and *A* tokens an experience *e** which is phenomenally the same as *e*. Then they claim that since *e* and *e** are phenomenally the same, they would give rise to the same confirmation-procedures, *i.e.*, they would cause the same subsequent experiences in you and *A* of seeming to oneself to try to verify *e*/*e** respectively. Then, as the authors say:

The point is that differences in sensory-phenomenal content normally are reflected by

[13] Obviously, since Pautz assumes a weak version of psychological supervenience —it only concerns *color-related behaviors*—the version of PPS implied by Pautz's argument is weaker than the one I present in §2. I will address this issue in §§4−5.

differences in confirmation/disconfirmation procedures. Thus, sameness of confirmation/disconfirmation procedures provides *strong evidence* for sameness of content…(2002, p. 531, n. 17, emphasis original).

Thus, you and *A* having the same confirmation-procedures with respect to *e*/*e** constitutes strong evidence that *e* and *e** have the same content.

Then, as Horgan and Tienson continue, since the above reasoning makes no assumption about the physical constitution of *A*, it is plausible for us to generalize the same claim to *all* phenomenal duplicates, and hence converse intentionalism is established. In addition, since informational semantics cannot explain converse intentionalism, and Experiential PIT is the only available explanation, by IOE, Experiential PIT is true.

Specifically, in the first part of their argument, they assume the following which appears to be a weak form of PPS:

If two experience tokens have the same phenomenal characters, then they give rise to the same confirmation-procedures.

Strictly speaking, however, this principle is a weak form of PPS only if their notion of confirmation-procedures is a psychological notion, i.e., only if it refers to a certain sort of psychological procedure. But might it be argued that the confirmation procedure is not a psychological notion, but an epistemic one? To me, it seems to be both. Just as a purely physical procedure (e.g., in a physics lab) can also be a confirmatory procedure, the same can be true of a procedure that is purely psychological. Such procedures may have epistemic *ends,* but they involve purely physical/psychological *means.*

It is clear, moreover, that the authors do have purely psychological procedures in mind:

That these phenomenally identical experiences all have the same truth conditions is reflected in the fact that each of the experiences is subject in the same way to investigation as to whether it is accurate. For example, you and your phenomenal duplicate each might have the experience of *seeming to oneself* to be testing one's perceptual experience for

> accuracy by making measurements or using a level….[E]ach might have *the subsequent experience of seeming to oneself* to discover that the picture merely appears to be crooked because of irregularities of the wall… Or, you and your [twin] might….*have the experience of seeming* to discover that there actually is no picture—say, by *seeming to oneself* to discover that one has been looking at a clever holographic image….(2002, p. 524, emphasis added).

Note how the authors go out of their way to couch the procedures in psychological terms. In this way, their idea of confirmation procedures is not just an epistemological notion, but *also* a psychological one. It refers to the psychological consequences of the agent's attempt to check the veridicality of her experience.[14]

Now we can summarize their argument in premise-conclusion form:

1. Every two experience tokens having the same phenomenology have the same confirmation-procedures. (a weak form of PPS)
2. Suppose *A*'s token and mine have the same phenomenology. (Assumption for conditional proof)
3. Then *A*'s token and mine (would) have the same confirmation-procedures. (by 1, 2)
4. And so *A*'s token and mine (would) have the same content. (by 3 and the reason that sameness of confirmation-procedures is strong evidence for sameness of content)
5. All in all, if *A*'s token and mine have the same phenomenology, then they have the same content. (2-4 by conditional proof)
6. Since *A* was arbitrary, every two experience tokens having the same phenomenology have the same content. (1-5, generalizing from an arbitrary case)
7. Therefore, converse intentionalism is true.

Once converse intentionalism is established, informational semantics can be ruled out by saying that it violates converse intentionalism (citing Block 1990), and Experiential PIT can be justified by IOE.

In sum, the argument from phenomenal duplicates is premised on some form of PPS. It starts from sameness of phenomenology. To get to sameness of content, it is equipped with

[14] It is not an accident that the authors use such psychological terms. It is because they want to allow for the possibility that your twin is a BIV who does not have a flesh-blood body that could not really measure the crooked picture, whose confirmation procedures are systematically hallucinatory.

bridges from phenomenology to psychology, and from psychology to content. The former bridge is PPS. So if it is falsified, then the argument is undermined.

§3.3 Mendelovici's Argument

For this type of PPA, I will examine (Mendelovici 2018). The feature of this type is that it does not assume psychological supervenience in the original form, but only entails a significantly weakened version of PPS. Hence her version of PPA is stronger than the previous two (see §2).

Since the context of Mendelovici's argument involves the metaphysics of color, it is important to know the relevant views and how they are related to experiential content. According to informational semantics, color is the physical surface property of an ordinary object that causes our color experience. At first, it may seem that this surface property should be characterized in terms of its intrinsic features, but since it turns out that color scientists have a hard time providing a unified account of physical red in terms of such intrinsic features, informational semanticists often take physical red to be *the dispositional property that normally causes in normal humans the tokening of the neural state that realizes the ph-red experience* (Dretske 1995, Lycan 1996, Tye 2000, Byrne and Tye 2006). In contrast, PIT theorists' views are motivated by the BIV experiment. Since as PIT theorists say, the BIV and I have the same phenomenal character when we see red objects, they insist that our common experience is about redness. But because the BIV does not have a causal connection to the environment, redness cannot be any property that causes its (and therefore my) experience. Instead, PIT theorists conclude that redness is a qualitative, non-dispositional property. To avoid confusion, I will use "dispositional color" to refer to the content of our color experience as understood by informational semanticists, and "qualitative color" for the content intended by PIT theorists (Pautz 2006b).

Mendelovici's argument follows the pattern of IOE. She first lists two competing semantics of experiential content, informational semantics and PIT, and then adduces some

psychological data, arguing that the former makes false predictions about the data whereas the latter can correctly predict it. Hence, by IOE, PIT is true.

Let's see her argument in the premise-conclusion form:

1. If informational semantics is true, the psychological roles of our color experiences would suggest that color is a dispositional property.
2. However, the psychological roles of our color experiences suggest that color is intrinsic, *i.e.,* not dispositional or relational.
3. If Experiential PIT is true, the psychological roles of our color experiences suggest that its content is intrinsic.
4. Therefore, informational semantics is false and by IOE, Experiential PIT is true.

In order to independently justify premise 2, she appeals to a methodological requirement which she calls *psychological involvement*. As she says, "our intentional states generally play various psychological roles, and these psychological roles are *appropriate* to which content they represent." (p. 27, emphasis added) The idea is that any adequate semantics should make correct predictions about our relevant psychology, actions, and other intentional behaviors.

To illustrate her principle of psychological involvement, Mendelovici distinguishes between heaviness and weightiness (Mendelovici 2010, 2013). In her view, weightiness is a *relation* but heaviness is an *intrinsic property* because the former depends on the spatial distance between objects while the latter doesn't. As she argues, our experience of objects can only concern their heaviness since its psychological roles suggest that the relevant content is an intrinsic property. For example, we typically believe that a heavy object is heavy whether or not it is at the top of a mountain. We also naively put the same amount of effort to lift an object even at different locations. Finally, if we experience that something is hard to lift, we sometimes infer that it is still hard to lift on the moon. In contrast, if the psychological roles suggest that the relevant content is a relation, we should behave and infer quite differently. For instance, before inferring to the conclusion that an object is heavy, we would want to know the distance between the object and Earth. Or we would not infer that a heavy object at a place is

equally heavy at another place if we know that the locations are different.

Then, as Mendelovici points out,

> [c]onsiderations of psychological involvement also suggest that perceptual color representations represent…[qualitative] colors. Intentional states involving perceptual color representations are inferentially related to beliefs about…[qualitative] colors…For example, absent countervailing theoretical beliefs, a perceptual intentional state representing that an object is sky-blue might lead to the judgment that the object is some shade of…[qualitative] blue, i.e., that it has a…primitive, non-dispositional, non-relational, and non-mental property of blueness. (pp. 42-43)

As Mendelovici argues, however, informational semanticists would wrongly predict that people's behaviors and inferences related to colors suggest that they see a dispositional property. In contrast, the PIT theorist would make correct predictions since, according to her version of PIT, the content of color experience is qualitative color, not reflectance properties. Hence, we should reject informational semantics and endorse PIT.

Strictly speaking, Mendelovici's requirement of psychological involvement is not a version of psychological supervenience (§2). In fact, it is significantly weaker than psychological supervenience. Notice that in order for her argument to work, she does not need to claim that experiences having the same content play exactly the same psychological role. For example, it may be that you like redness but I am afraid of it, and so your experience of redness would cause you to go toward a red object while mine would cause me to avoid it.[15] Rather, she only needs to claim that your color experience and mine play psychological roles *of the same particular kind*. Which kind? Answer: in our current context, both experiences play the kind of psychological roles that suggest that their contents are intrinsic properties. Accordingly, what she denies is that two experiences can represent the same property (or the same relation), the role of one suggests that the represented feature is intrinsic while the other suggests that it is relational. In sum, Mendelovici's principle of psychological involvement says

[15] I greatly appreciate Prof. Mendelovici's clarifications on this point.

at least that experiences having the same content play psychological roles of the same kind with respect to whether the role suggests the relevant content is intrinsic or not.

What role is of the kind that it suggests that the represented content is intrinsic (or relational)? What are the examples? Although Mendelovici does not present a general sketch of what the example she has in mind, it is instructive to reflect on her heaviness/weight case. As she points out, since heaviness is intrinsic, we would believe that a heavy thing on Earth is still heavy on the moon. Let's call this belief "M". Suppose I experience that some object is heavy. My experience causes me to believe that it is heavy. Then by M, I infer that it is heavy on the moon. On the other hand, if weight is the represented content of my experience, then we would not believe that a weighty thing on Earth is still weighty on the moon. If we know that the gravitational field is smaller, or the distance becomes longer, then we would believe that the object becomes less weighty if the environment is different. Hence, my experience would cause me to infer that the object is less weighty on the moon. Therefore, if one's experience with the content that something is P causes one to infer that the same thing is still P at different environments, then the psychological role of the experience suggests that the represented content P is intrinsic.[16]

We have seen three types of PPAs. In the first type (shifted spectrum), the argument is premised on psychological supervenience, which together with converse intentionalism, entails PPS. Much the same goes in the third type, except that a weakened version of psychological supervenience, i.e., the principle of psychological involvement, is presupposed in the methodology. In the second type (phenomenal duplicates), PPS appears as one of the

[16] This conclusion needs a qualification. As Mendelovici points out, an object does not retain its intrinsic properties under variations of its relational properties *if an additional causal influence is applied to it when varying its relational properties* (2010). For a red object is still red whether it is on the table or on the ground. However, if it is put under the bright sunlight, its color fades. In contrast, sometimes we can change an object's relational properties without exerting causal influences. E.g., I can simply make an object close (to me) by walking toward it. Hence, the conclusion could be revised as what follows: if one's experience with the content that something is P causes one to infer that the same thing is still P at *different environments without any additional causal influence*, then the psychological role of the experience suggests that the represented content P is intrinsic.

bridging premises from sameness of phenomenology to sameness of content. Let me reiterate that while I do not deny that there may be arguments for (Experiential) PIT that do not involve PPS, my targets in this chapter are PPAs.[17] Given that PPAs take center stages in many foundational works for PIT, the foundation of PIT may not be as stable as it appears.

In the next two sections (§§4−5), I will introduce two counter-examples to PPS. I will call the former "the banal counter-example", and the latter "the 45-degree shifted spectrum". And I will show how they undermine those PPAs we have seen.

§4 The Banal Counter-Example and Objection

In this section, I will first discuss a banal counter-example to undermine PPAs (§4.1), and then I will discuss and defuse a potential objection (§4.2).

§4.1 The Banal Counter-Example

Consider two normal persons, John and James. John correctly believes that red paints or pigments cannot be produced by mixing paints of other colors. He also correctly believes that orange paints can be produced by mixing red and yellow paints. James, in contrast, has the opposite beliefs. Due to some banal reason, e.g., he is misinformed by his friends, he wrongly believes red paints can be produced by orange and purple paints, and he also believes that orange paints cannot be produced by mixing paints. In fact, just like Twin Max, James believes that all paints of unitary colors are producible by mixing paints and all binary ones are not.

Now I would like to point out that John's and James' ph-red experiences are *prima facie* counter-examples to PPS. Their ph-red experiences have the same phenomenal character, but have different psychological roles. For example, suppose they both token ph-red experiences, and they both want to reproduce it. Then James would want to reproduce it by mixing other paints whereas John would not. Hence PPS appears to be falsified.

[17] In Horgan, Tienson and Graham 2004, the authors take sameness of the content of phenomenal duplicates (corresponding to premise 5 in the above argument) as an intuition without arguing for it. Loar (2003) also suggests a similar idea. Referring to the above argument, Pautz (2006a) also appeals to what he calls "the C-Dependence principle" to justify premise 4. In her work (2013), Mendelovici presents a defense for PIT from the possibility of reliable misrepresentation. None of these arguments involve PPS.

Notice that, however, the version of PPS in §2 says that two experience tokens having the same phenomenal character are governed by *exactly the same psychological functions*. This is too strong. Horgan and Tienson's version only require that the tokens are governed by the same empirical confirmation-procedures while Pautz's version only requires they are governed by functions relevant to color-related behaviors. Finally, Mendelovici's principle of psychological involvement requires that the tokens' roles suggest that the represented contents are intrinsic (or relational). So people might say that even if the banal case is a counter-example to PPS, it does not mean that PPAs are undermined.

Nevertheless, it is not difficult to adapt the John/James case to their principles. Let's start with Horgan and Tienson's argument. Notice that the difference between beliefs about the reproducibility of colors by mixing also brings about a difference in confirmation procedures. For example, suppose John and James each encounter a picture they experience as ph-red, but in conditions where each has a suspicion that the lighting condition is abnormal. In order to verify their color experiences, John and James would act differently. Presumably, John would compare the color of the picture with the paint he believes to be red while James would try to reproduce the picture's color by mixing paints. So we have a difference in confirmation-procedures. Next, consider Pautz's argument. Recall that his psychological supervenience in conjunction with converse intentionalism implies that two tokens having the same phenomenal character cause the same color-related behaviors. But it is clear that John's ph-red experience and James' cause different color-related behaviors. As I said, if they both want to reproduce their ph-red experience, James would want to proceed by mixing orange and purple paints while John wouldn't.

Finally, the banal example can also be revised to undermine Mendelovici's argument. Suppose instead of being misinformed of the mixability of colors, James is convinced by a British empiricist philosopher into believing that, contrary to its appearance, color is a relational property. For instance, he believes that when put in a dark room, no apple is red

anymore. So when both John and James see a red apple, John's ph-red experience causes him to infer that it is still red in the room whereas James' doesn't. The psychological roles of John's ph-red experience suggest that the represented content is intrinsic but the roles of James' experience suggest that it is relational.

§4.2 Ceteris Paribus?

The banal counter-example at its face appears to be doubly surprising. Though it would be surprising if the leading theories of phenomenal intentionality were felled by such a humdrum counter-example, it turns out to be surprisingly difficult for the PPA advocates to defuse it.

Since it involves misleading information, perhaps we can get around it by revising PPS as follows:

> Strengthened Phenomenology/psychology supervenience: *Ceteris paribus*, every two experience tokens having the same phenomenal character play the same psychological role.

This is called "strengthened" because the antecedent is strengthened. We can do a similar trick to psychological supervenience and Mendelovici's principle of psychological involvement to get around the banal counter-example. Since John and James have different background beliefs, something is not equal, and thus the antecedent of the strengthened PPS is not satisfied. Now the banal example is no longer a counter-example.

Let's now evaluate how successful this strategy is. Notice that when putting the ceteris-paribus condition into the antecedent, PPA advocates cannot literally mean that *every other factor remains equal* since clearly, in Horgan and Tienson's case, my twin could be a BIV, who does not have a flesh-and-blood body, and in Block's case, Rose is installed with lenses on the inverted earth. Hence, a reasonable reading is that

> If two experience tokens have the same phenomenal characters, and *other relevant factors are equal*, then they play the same psychological role.

Now if the PPA advocates say that the John/James case is not a real counter-example because some relevant factor is not equal, then they need to specify what the other *relevant* factors are. For example, they may say that one relevant factor is background belief, and so the antecedent of PPS would become "if the experiences have same phenomenology *and the subjects have the same background beliefs*…" Then, as the PPA advocate might continue, since John and James have different background beliefs the antecedent is not satisfied.

However, as we will see immediately, any strengthening strategy would logically weaken the principle, and hence the relevant PPA argument cannot get off the ground. In general, the strategy of declaring some variable X, in addition to phenomenology, to be relevant will face a dilemma. X could be occurrent or unconscious beliefs, history, past experience, or whatever variable you like. Either (a) X supervenes on phenomenology or (b) it doesn't. If (a), then the strengthened PPS or psychological supervenience reduces to the original PPS and psychological supervenience. If (b), it becomes questionable that experiential intentionality is *solely* determined by phenomenology, i.e., that phenomenology is the *unique determinant* of experiential intentionality.

For example, suppose Horgan and Tienson strengthens their PPS. Corresponding to the formalized argument (§3.2), the use of PPS is to license the step from premise 2 ("the experiences of my twin and me have the same phenomenology") to premise 3 ("the experiences have the same confirmation-procedures"). If the principle is strengthened by requiring some relevant factor X to be the same, then they need to justify an additional premise that my twin and I have the same X. It is not an attractive strategy for Horgan and Tienson to strengthen their principle because *they intend their thought experiment to be about phenomenology only*. As the authors say to their reader,

> Consider any creature who is a complete phenomenal duplicate of yourself.… Assume nothing *else* about this creature. The thought experiment thus builds in an epistemic "veil of ignorance" about this creature, in order to *filter out any factors other than*

> *phenomenology itself*. So for all you know about this arbitrary phenomenal duplicate of yourself, its sensory-perceptual experience…might be very largely illusory and hallucinatory concerning the real nature of itself and its surroundings…(2002, p. 524, emphasis added).

Apparently, the authors intend your twin and you have only one thing in common, i.e., phenomenology. Hence given this setting, they cannot justifiably claim that you and your twin have some other factor X in common.

On the other hand, suppose the authors want to strengthen the antecedent of PPS by adding a new relevant factor X, and reset their experiment in such ways that they can justify that you and your twin have X in common. Then they can justify that the experiences of you and your twin have the same content. However, as we see in the formalized argument (§3.2), premise 6 generalizes sameness of content to *all* phenomenal duplicates. So in the strengthened case, the claim of sameness of content can only be generalized to pairs of phenomenal duplicates *with the same X*. This is significantly weaker than converse intentionalism. For example, if they add background belief as a relevant factor, then all they can establish is that *the experiences of all phenomenal duplicates with the same background beliefs have the same content* (call this italicized claim "B"). However, recall that they eventually justify PIT by using IOE, i.e., by showing that only PIT can explain converse intentionalism. Accordingly, in this strengthened case, the authors would justify PIT by showing that only it can explain the claim B. PIT after all predicts that all phenomenal duplicates have the content, *a fortiori* that all phenomenal duplicates *with the same background beliefs* have the same content. But it is not the only explanation of B since instead of saying that phenomenology determines content to explain B, we can explain B more narrowly by saying that the content of a state is determined by its phenomenal character together with its causal relations to other background beliefs.

Similarly, consider Pautz's psychological supervenience. If he strengthens the antecedent by adding one relevant factor X, then the full statement of his psychological supervenience would be:

> If two actual or possible individuals have qualitatively identical color experiences, *and the subjects have the same X*, then the subjects would do the same color-related behaviors.

This weakened principle, together with converse intentionalism, implies a weak version of PPS, to which the John/James may not constitute a counter-example. However, referring back to the formalized argument (§3.1), Pautz needs a principle to justify premise 4, i.e., the experiences of Max and Twin Max have different contents. His reasoning is that assuming his original version of psychological supervenience, since Max and Twin Max have different color-related behaviors, by *modus tollens*, we can get premise 4. But now that psychological supervenience is weakened, he can no longer deduce premise 4 simply by appealing to the fact that Max and Twin Max have different color-related behaviors. The most he can infer is that premise 4 holds when Max and Twin Max have the same X.

But what X do Max and Twin Max have in common? Notice that Pautz is arguing against informational semantics so X had better be an internal factor. Otherwise, informational semanticists could easily get away by saying that the environmental conditions in the thought experiment is underspecified so there is no reason to claim that they have the same X. But what kind of internal factor could X be? Same neural state? No. It is because Pautz already stipulates that their neural structures are wired to the environments differently so that when Max and Twin Max see orange objects, they have different phenomenology. Same memory? Same belief? Same sub-personal information-process? Not obviously. Given that Pautz says that Twin Max's species has a different evolutionary history from *Homo sapiens*, who knows what kinds of processing algorithms are realized in Twin Max, and how they are realized? Thus, it is questionable that Pautz could justify that Max and Twin Max have some common internal

factor relevant to behaviors.[18] [19]

Finally, we can see that Mendelovici suffers from a similar problem. As we have seen, her argument methodologically presupposes the principle of psychological involvement. The question for her is whether she can presuppose a strengthened principle of psychological involvement and still run the argument.

Recall that according to her requirement of psychological involvement, semantics should correctly predict psychology. In the experiential case, Mendelovici requires that the semantics of experiential content should correctly predict the psychological role of the experience. It is because, by these lights, informational semantics makes the wrong predictions, that she is able to prove it inferior to PIT.

However, suppose Mendelovici endorses a strengthened principle of psychological involvement, e.g., by saying that the same experiential content *together with the same background belief* give rise to same kind of psychological roles. Then this makes Mendelovici's argument harder to get moving. Now that a new variable has been added, it is harder for Mendelovici to blame the erroneous predictions on informational semantics, rather than other factors. By the same token, the informational semanticist can attribute the error to the background beliefs, and thus save the theory.[20] This result should make the strategy

[18] Of course, Max and Twin Max have the same chemical constitution, e.g., both brains are constituted by carbon and water. In that case, the version of psychological supervenience would be that same content and same chemical constitution implies same behavior. But this principle is plausible only because it is entailed by the original psychological supervenience since according to the standard view in cognitive psychology, behaviors are identified functionally, and the chemical constitution shouldn't matter to the identification of behaviors. To illustrate my point, suppose Max gets a brain damage at the region for motor control, and has a replacement of the damaged region with an electronic tissue. In that case, when Max and Twin Max token color experiences, and have different color-related behaviors, intuitively we would still conclude that their experiences have different contents. It seems to me that this scenario shows that chemical constitution is not relevant to the identification of behavior. Hence if psychological supervenience strengthened by sameness of chemical constitution is plausible, it is so only because the original version is true. But I already argue that the original version, when coupled with converse intentionalism, entails PPS, which is falsified by my counter-example. So if the original version is rejected, the strengthened version by chemical constitution loses its plausibility.

[19] Perhaps Pautz could revise his case to facilitate this kind of answer. But it is hard to know what revisions would work. Moreover, as Byrne and Tye (2006) point out, any such revision is likely to undermine the plausibility of the claim that the duplicates could differ in their phenomenology after all.

[20] Furthermore, since sometimes it is difficult to detect one's own background belief, adding background belief as a relevant factor makes the conversation between informational semanticists and PIT theorists unverifiable/unfalsifiable, and unfruitful.

unattractive to Mendelovici.

§5 The 45-Degree Shifted Spectrum and Objection

Let's set aside whether or not the strengthening strategy succeeds in defusing the banal counter-example. There is a persisting intuition that the John/James case is unsatisfying precisely because it involves different (or even *defective*) background beliefs. Isn't it, as people might think, natural and intuitive to expect that John's and James' ph-red experiences would largely cause them to do the same thing if everything goes back to normal? In this section, I will present another case to undermine PPAs which does not involve any difference in background beliefs (§5.1), and then anticipate an objection to it (§5.2).

§5.1 Pautz's Thought Experiment Revisited and Mental Simulation

Recall Pautz's thought experiment where he exploits the unitary/binary shift of spectra. As he argues, there is a possible world where everything is the same as in the actual world except that the evolutionary history is different such that Twin Max's spectrum is shifted from that of his actual-world counterpart, Max (§3.1). Putting in terms of the hue circle, the effect of evolution in Pautz's case is to shift Max's spectrum 45-degrees around the hue circle, counter-clockwise.

I would like to begin by arguing that there is a peculiar aspect about Twin Max's ph-red experience, which goes unnoticed by Pautz himself in his original paper. To see that, consider an RGB monitor. Suppose Max sees a completely red monitor, and tokens a ph-red experience. Now if the picture is drawn nearer and nearer to Max, i.e., if Max zooms in closer and closer to the monitor, he would consistently token the same ph-red experiences. On the other hand, suppose Twin Max sees an RGB monitor, and also tokens a ph-red experience. What would happen to his experience if he zooms in closer to the monitor? Since Twin Max's spectrum is 45-dgree shifted from Max', the monitor he sees is actually purple which means that the monitor is illuminated with red light and blue light. Since red light and blue light look orange and purple to Twin Max, presumably, if he keeps zooming in, eventually he would see

that this monitor consists of many pixels, and that each of them looks either purple or orange. I.e., unlike Max's experience, if zooming in, Twin Max would token an experience with a very different phenomenal character.

Now let's think of the zooming-in process as a function, and we can immediately see the outline of a counter-example to PPS. The reason is that the same function, i.e., the zooming-in process, would map experiences with the same phenomenal character to experiences with different phenomenal characters in Max and Twin Max. For example, the zooming-in process would consistently map Max's ph-red experience to another ph-red experience, just like the identity function. In contrast, for Twin Max, the same process would map his ph-red experience to an experience with the character that is partly ph-purple and partly ph-orange Hence, experiences with the same phenomenal character do not have the same psychological role.

The same point can be illustrated if we do it backward. If Max sees a monitor consisting of pixels which are either blue or red, and then he zooms out, eventually the picture looks purple to him. Notice that both blue and purple are bluish. So for Max, the zooming-out process is a function that maps a ph-bluish experience (because the pixels look either blue or red to him) to another ph-bluish experience (because the picture looks purple to him). In contrast, if Twin Max starts by seeing a monitor consisting of pixels which are either purple or orange, and zooms out, then eventually the picture looks purely red to him. So for Twin Max, the zooming-out process maps a ph-bluish experience (because the pixels look either purple or orange) to a non-ph-bluish experience (because the picture looks purely red to him now). Thus, for Max and Twin Max, ph-bluish experiences do not evolve to experiences with the same phenomenal character even though both undergo the zooming-out process. More importantly, for Twin Max, whether an experience is ph-bluish or not is a function of his distance to the picture while it is not the case for Max.

Notice the crucial difference between the banal counter-example (John/James case) vs. the Max/Twin Max case. In the former case, John and James have different background

beliefs while in the Max/Twin Max case, there is no such a difference. We have seen that the banal example might be challenged on the basis of such a difference. Here in the Max/Twin Max case, no such difference is presupposed.

So far I have argued that the zooming process would map Max's a ph-red experience to another ph-red experience, and it wouldn't for Twin Max. However, we haven't had a clean counter-example to PPS yet. Since as someone might notice, during the zooming process, both agents continuously receive light from the external environments, and since they receive light of different colors, the fact that they have phenomenally different experiences at the end can be explained by the different inputs they receive from the environments, and hence this is not a clean case of narrow psychological difference between their ph-red experiences.[21]

Nevertheless, I think we can get a counter-example by exploiting the zooming differences between Max and Twin Max. One way to show their narrow psychological difference is to consider both agents when they do imaginations or simulations in their minds. For instance, suppose both token ph-red experiences, and are going to imagine in their minds what they would see on the basis of their initial experiences *if they were to zoom in*. Now, *if their simulations are veridical*, then they would have different (imagined) consequences: Max would have a ph-red mental imagery while Twin would have a partly ph-orange and partly ph-purple mental imagery. Notice that during the simulation process, they don't receive inputs from the external environments. Thus, the difference between their simulation processes must be *narrow*.

Still, some people might complain that the idea of a *veridical* mental state or mental process cannot be defined narrowly, and hence my argument in the above paragraph does not work. Nevertheless, instead of saying that if their simulations are veridical, then they would have different mental images eventually, nothing significant is changed if we say that their

[21] I thank Prof. Mendelovici for pointing out this to me.

results of simulations would be different when the simulations are done properly or *normally*. I.e., assuming that Max's and Twin Max's simulation processes are normal, the processes, during which no external input is received, would map ph-red experiences to different mental images. This, I claim, is a genuine counter-example to PPS.

One may ask: Why not say that Twin Max's normal simulation maps ph-red experience to ph-red imagery? Why not say that his simulation is *mistaken* when Twin Max tokens a partly ph-orange and partly ph-purple mental imagery at the end of the simulation? Do we have any justification to claim that Twin Max's normal simulation maps ph-red experience to partly ph-orange and partly ph-purple mental imagery? I think we do. Suppose, for *reductio*, Twin Max's normal simulation of zooming-in maps ph-red experience to ph-red mental imagery. Then presumably, on the basis of his normal simulation, he would have a certain kind of expectation about what experience he would token if he sees something that looks red to him and starts to zoom in. For example, suppose Twin Max sees a monitor that looks red. Before he starts to move toward it, he would have the belief that "since that monitor looks reddish, but neither bluish nor yellowish, then it continues to look the same if I move toward it". Then when he moves toward it, Twin Max would be surprised since his expectation is not met. Contrary to what he expects, the monitor looks partly orange and partly purple to him. Therefore, if Twin Max's normal simulation of zooming-in is the same as Max's, then his predictions are systematically false and insofar as his strategies for meeting them rely on those predictions, his desires fail to be met generally. Consider another example: assume that Twin Max is installed with a pair of lenses with a high resolution so that no mid-size object looks red to him if it is 1 meter (or less) away from him. Suppose he has the desire to avoid anything that looks bluish. And on the basis of his simulation, he would expect that no apple would look bluish however close it is. When he sees an apple that looks red to him, he moves toward the apple to get it. And when the apple is within the 1-meter range of Twin Max, it looks partly purple and hence bluish, and thus Twin Max would stop moving and feel frightened. Hence,

Twin Max's desire of eating non-bluish apple generally is not satisfied. Again, this is implausible. Since as Pautz says, Twin Max is equipped with the 45-degree shifted spectrum by natural selection, it is highly improbable that a Pautzian agent like Twin Max could survive if his predictions are systematically false and his relevant desires are not satisfied. Hence the normal simulations of Max and Twin Max are different.

Let me bring back the present case to PPAs by starting with Horgan and Tienson's argument. Suppose Max and Twin Max each encounter a picture they experience as ph-red, but in conditions where each has a suspicion that the lighting condition is abnormal. In order to verify their color experiences, Max and Twin Max would have different confirmation-procedures even if they both token ph-red experiences. For example, on the basis of their simulations, they would have different predictions about the resultant experiences once they go near to inspect it. In particular, to confirm their initial experiences, they need different resultant experiences: Max needs a ph-red one while Twin needs a partly ph-orange and partly ph-purple one. Hence the confirmation procedures through which they confirm their initial experience are different.

Now consider Pautz's principle. Recall that his psychological supervenience together with converse intentionalism, implies that same phenomenology of color experiences causes the same *color-related* behaviors. However, it is clear that in the present case, the subjects do have different color-related behaviors. For example, when an object far away causes Max to toke a ph-red experience who desires to have it, Max would move towards it until he succeeds. But Twin Max in the same situation would stop moving when the object is close to him.

Finally, let's think about Mendelovici's principle of psychological involvement. Recall her notion of psychological roles: "These roles might be roles in the inferences we are disposed to make, the behaviors we are disposed to engage in, or the higher-order thoughts or introspective judgments about our intentional states we are able or likely to have." (2018, pp. 27-28) Now, unlike the traditional inverted spectrum, people with the 45-degree shifted

spectrum have plenty of different psychological dispositions from normal people. They also have different inferential dispositions. For instance, on the basis of their respective simulations, Max is disposed to infer that colors are intrinsic to objects while Twin Max is likely to infer that they are something that depends on spatial distance. Consequently, they are likely to have different higher-order thoughts about their color experiences. Max is likely to infer that his color experience represents an intrinsic property of an object but Twin Max isn't.

More importantly, according to her principle of psychological involvement, Mendelovici insists that it is impossible that two experiences represent the same intrinsic property (or the same relation) but at the same time, one's role suggests that its content is intrinsic and the other's suggests that its content is relational (§3.3). Since by converse intentionalism, experiences having the same phenomenology have the same content, Mendelovici's claim implies that *any subject's color experience having the same phenomenology as a normal human's also plays a psychological role of the same kind with respect to whether the role would suggest the represented content is intrinsic or relational.* However, as I have argued, this italicized sentence can be put into question given the present example since Twin Max's ph-bluish experience plays the role that may suggest differently from Max's.

In this subsection, I have presented the Max/Twin Max example without any difference in background beliefs, and have argued that it can adjusted to fitting the relevant psychological roles specified in the versions of PPS examined in this essay. Horgan and Tienson's PPS concerns confirmation-procedures, Pautz's concerns color-related behaviors and Mendelovici's weak version is about whether the role suggests the content is an intrinsic property or not. Whatever notion of relevant psychological roles they have in mind, however, I see no in-principle obstacle to constructing a counter-example of the same general sort. As of right now, I do not think PPA advocates can dismiss this example simply by saying that it is irrelevant to the versions of PPS involved in their arguments.

§5.2 Background Beliefs and Possibility

Does mental simulation require beliefs? Is it mediated by inference? Some people might want to challenge my counter-example by arguing that mental simulation requires background beliefs, and since presumably different simulations require different background beliefs, the Max/Twin Max case is not better than the banal example.

There might be evidence from empirical sciences that human mental simulation is mediated by beliefs. Nevertheless, to refute PPS, all I need is *a possible situation* in which the mental simulation is not mediated by beliefs. In this situation, you token a series of mental images and you don't do any inference. It could be conceived as a series of mental states, the first of which is an experience, and the rest of which are all mental imageries. And when engaging in this, Twin Max does not do any inference (which is perhaps like what a normal person does when she "sees" things in her dreams). (I want to point out that the condition that Twin Max's normal mental simulation is not mediated by beliefs is not part of, or entailed by, Pautz's original story. So properly speaking, the character I call "Twin Max" in this section should be separated from Pautz's. To avoid confusions, from now on, I will call my character "Clone", and his counterpart in the actual world "Jones")

PPA advocates might argue that Clone is impossible. My reply is that in the literature, almost all PPA advocates use more *recherché* examples against informational semantics, e.g., disembodied Cartesian egos or BIVs (Horgan, Tienson 2002, Loar 2003, Chalmers 2004, Kriegel 2013, Mendelovici and Bourget 2014). This is no accident. For in order to show that other factors (specifically, causation) are irrelevant to semantics, they need to imagine situations in which those factors are abstracted away (specifically, minds that are causally disconnected from the environment). I struggle to believe that my thought experiment about Clone is more scientifically outlandish than a brain removed from its skull and suspended in liquid, let alone a disembodied Cartesian ego.

Moreover, my story of Clone seems to me to be essentially similar—and therefore

every bit as possible—as Pautz's Twin Max. For they differ only by whether or not mental simulations are mediated by beliefs, i.e., in Pautz's scenario of Twin Max, the process of mental simulation is not discussed whereas in my thought experiment, it does not involve beliefs. This functional difference, in my view, should not matter with respect to the fundamental laws. If the one's overall conditions are compatible with the fundamental (metaphysical/physical) laws, the other's should be too. Given that PPA advocates use the Cartesian ego/BIV cases and Pautz's cases to argue against informational semantics, it is hard for them to deny my experiment's possibility while accepting theirs.

§6 Conclusion

In this chapter, I argue that a particular kind of arguments for the phenomenal intentionality theory, called "the phenomenal-psychological supervenience arguments" or "PPAs", is not successful. I first provide the background in §2. In §3, I discuss three types of PPAs: the argument from shifted spectrum which entails PPS (§3.1), the argument from phenomenal duplicates which is premised on the principle (§3.2), and the argument methodologically presupposing a weakened version of the principle (§3.3). In §4, I presented a banal counter-example to undermine those PPAs, and argued that it is not easy for PPA advocates to get around the banal example. In §5, I presented another counter-example, i.e., the Jones/Clone case, without the flaw that is claimed to baffle the banal example. I finally argued that it undermines PPAs more satisfactorily.

I do not intend my work to rebut PIT generally. Nevertheless, given that PPA appears in many classic works for PIT, if my argument works, then the foundations of PIT as laid out in those classic works should be put into doubt.

Chapter 4

Psychointentionalism

Abstract

Both the standard semantic theories of experiential content, strong representationalism and the phenomenal intentionality theory ("SR" and "PIT" hereafter) endorse *converse intentionalism*, the thesis that experiential intentionality supervenes on phenomenology. However, it can be shown that in conjunction with the semantic theories presupposed by SR and PIT respectively, converse intentionalism suffers from various counter-examples which involve color inversion or shifted spectrum (e.g., Block 1990, 2003). In this chapter I propose a replacement, called "(converse) psychointentionalism", which holds that experiential content supervenes on phenomenology *and psychological role*. This is a psychological solution that does not presuppose any particular semantic theory. I will first provide expositions of the fundamental concerns of both SR and PIT and why they endorse converse intentionalism. Then I will discuss how those counter-examples constitute great challenges to SR and PIT, and how converse intentionalism gives rise to those counter-examples. Finally, I will show that psychointentionalism avoids the counter-examples while addressing the fundamental concerns of SR and PIT. This provides strong justification for my replacement.

§1 Introduction

Both the standard semantic theories of experiential content, strong representationalism and the phenomenal intentionality theory ("SR" and "PIT" hereafter) endorse converse intentionalism, that experiential intentionality supervenes on phenomenology. SR and PIT suffer from troublesome counter-examples. E.g., Block's inverted earth case (1990, 2003) has been taken as one of the strongest cases against SR. In a similar fashion, a case can be constructed against PIT (see chapter 3). In this chapter, I will argue that these counter-examples arise from converse intentionalism, and will propose a principle replacing it. My proposal, called "(converse) psychointentionalism", holds that the content of an experience

supervenes on the conjunction of its phenomenal character and psychological role. I will argue that it not only can address those counter-examples, but also is palatable to both SR and PIT, which provides strong reason for replacing the old intentionalism with psychointentionalism.

Here is the plan. In §2, I discuss the fundamental concerns of SR and PIT, why they all endorse converse intentionalism, and why this matters to understanding how both theories suffer from certain kind of counter-examples. In §3, I present my theory. In §4, I show how the new theory can avoid those fatal counter-examples. Finally, in §5, I provide renderings of psychointentionalism under which it naturally fits into SR's and PIT's programs.

Conventions: Terms with the prefix "ph-" denote phenomenal characters, e.g., ph-red is the phenomenal character one has when she sees a red object under normal conditions.

§2 Semantics and Converse Intentionalism

§2.1 Semantic of Experiential Content

Most standard semantic theories of experiential content can be divided into two groups: informational semantic theories, and phenomenal intentionality theories ("PIT" hereafter) (Mendelovici and Bourget 2014). The idea of informational semantics is that an experience has as part of its content a property P iff the normal cause of the experience is P (Rupert 2008), whereas the central thesis of PIT is that the content of an experience is determined by its phenomenal character (Kriegel 2013b). The debate between the two has dominated discussions on foundational issues in semantics of experiential content.

Since the central thesis of PIT implies that intentional contents supervene on phenomenology, given phenomenal internalism and the transitivity of supervenience, it also implies that intentional content supervenes on internal states[1] [2] In contrast, informational semantics understands intentionality in terms of the relation of experience to the external

[1] Phenomenal internalism is the claim that phenomenology supervenes on internal states.

[2] Main advocates of PIT are Horgan and Tienson (2002), Loar (2003), Kriegel (2003, 2007, 2013a, 2013b), Horgan, Tienson and Graham (2004), Chalmers (2004, 2006), Pautz (2006a, 2006b, 2008), Horgan and Graham (2012), Mendelovici (2018), Montague (2010, 2016), Bourget and Mendelovici (2016/2019).

environment.[3] Thus the conflict between informational semantics and PIT about experiential content is related to the dispute between semantic externalism vs. the narrow content theory in philosophy of mind and language at large.

There have been many discussions on the dispute between externalism and the narrow theory in the literature.[4] As Kriegel (2013b) summarizes, the two camps are driven by two different motivations. According to him, the primary goal of externalists is to (ontologically) naturalize mental properties whereas the concern of PIT theorists is to have an account of intentionality that accommodates our psychological intuitions (see Mendelovici and Bourget 2014 for similar observations).

According to Mendelovici and Bourget (2014), the motivational connection between externalism and naturalism can be explained by the history of philosophy of mind and language in the late 1970s. At that time, various classic arguments for the externalist-causal theory of reference were proposed (Putnam 1975, Burge 1979, Kripke 1980, Dretske 1981). According to such theories, content is determined by some appropriate causal relation between the mental state and the environment. Within this framework, intentional properties appear to be naturalizable since it reduces contents to worldly objects and properties, and the relation between an intentional state and its content to a causal relation that can be implemented physically. Building on the externalist-causal theory, the remaining question for naturalists is simply to figure out how the implementation works. As Mendelovici and Bourget point out, many informational semanticists (Millikan 1986, Fodor 1987, p.98; 1990) acknowledge that externalist arguments motivate their approaches to naturalizing intentionality.

The narrow theory, on the other hand, is motivated by the intuition that the content

[3] Predominant informational semanticists are e.g., Millikan (1986, 1989), Dretske (1986, 1988, 1995), Lycan (1996, 2000/2019), Tye (2000, 2009), Byrne and Hilbert (2003), Byrne and Tye (2006), Fodor (2010), Neander (2017).

[4] For insightful discussions on the fundamental disagreement between the narrow content theory and semantic externalism, see Kriegel 2011, 2013b, Mendelovici and Bourget 2014, Yli-Vakkuri and Hawthorne 2018, Mendelovici 2018.

of a mental state should not (greatly) diverge from its psychological role (Lewis 1979, Loar 1988a, b, Mendelovici and Bourget 2014). One thought experiment elucidating this intuition appeals to an isolated brain in a vat ("BIV" for short). The BIV is neurophysiologically the same as yours, and tokens the same experiential state as you. As Kriegel says, we have the intuition that the BIV's experiential tokens have the same content as yours and play the same psychological role. However, since the connection to the environment has been lost, the externalist must claim that the BIV's experience either has a content different from yours or is contentless. Since the narrow theorist takes this to be implausible, they insist that content has to be *narrow*, i.e., that content supervenes on internal states.

In sum, the narrow theory's advantage is in accommodating the intuition that content does not diverge from internal psychology while externalism's strength is its potential to naturalize content.[5] Against this background, we will see in §2.2 how informational semantics and PIT as instances of externalism and the narrow theory respectively work in the experiential domain.

§2.2 Converse Intentionalism

Despite their fundamental disagreement, many informational semanticists and PIT theorists agree upon the following principle[6]:

> *Converse intentionalism*: Experiential content supervenes on phenomenology, i.e., every two experience tokens having the same phenomenal character have the same content.[7]

---

[5] This is not to say that externalists do not care about the psychological intuitions or narrow theorists do not care about the naturalizability of content. The idea is that *when it is impossible to achieve both*, externalists would prioritize the naturalizability of content over the psychological intuitions while narrow theorists would prioritize the opposite.

[6] I call this principle "converse intentionalism" because in the literature of the representational theory of consciousness, the principle that same intentionality implies same phenomenology is called "weak intentionalism". There is so far no standard name for the converse of weak intentionalism. See Lycan 2000/2019.

[7] Among informational semanticists, Dretske (1995), Lycan (1996), Tye (2000), Byrne and Tye (2006) endorse converse intentionalism. Millikan (1989, 2009) and Fodor (2010) remain silent whereas Neander (2017) feels suspicious about it. All PIT theorists agree with converse intentionalism. E.g., Horgan and Tienson: "There is a kind of intentional content,…, such that any two possible phenomenal duplicates have exactly similar intentional states vis-à-vis such content" (Horgan and Tienson 2002, p. 524); Chalmers' view that: "there is plausibly an entailment from perceptual phenomenal properties to pure representational properties" (Chalmers 2004, p.158); "If two experiences share the same phenomenological content, then necessarily they share (a kind of)

In the literature on the representational theory of consciousness, the identity thesis that all phenomenal properties are identical to experiential intentional properties (or at least that they co-supervene) is called "strong intentionalism". This is not to be confused with what I will call "strong *representationalism*" ("SR" hereafter) which is *the conjunction* of informational semantics *with* strong intentionalism. [8] Notice that the identity thesis entails that phenomenology and intentionality co-supervene, which further entails converse intentionalism. So SR also entails converse intentionalism.

Why do SR theorists and PIT theorists agree upon converse intentionalism? Recall that SR theorists are externalists whose primary concern is the naturalizability of mental properties. Since the mid-twentieth century, naturalists have made important progress in naturalizing intentionality by identifying intentional properties of the mind with its causal relations to physical objects and properties. With the identity thesis between experiential intentionality and phenomenology, which entails converse intentionalism, SR theorists could have a viable strategy for naturalizing phenomenal properties. Thus, without reducing phenomenology to any physical state directly, SR theorists identify phenomenology with experiential intentionality, and the latter with some appropriate kind of causal relation. That way, they bridge the phenomenological gap between the physical and the phenomenal, which has motivated the most troublesome problems for classical physicalism in philosophy of mind.[9] I shall call SR theorists' aim to naturalize phenomenology "the naturalist concern".

Why do PIT theorists also subscribe to converse intentionalism? Recall that their central thesis is that phenomenology determines experiential content. To understand their motivations for endorsing converse intentionalism, two questions are relevant. First, what is

representational content" (Montague 2016, p.86).

[8] In fact, in the literature of the representational theory of consciousness, "strong representationalism" and "strong intentionalism" are used interchangeably (see note 6). In this chapter, I follow Montague's (2014) distinction between different versions of representationalism. What I call "strong intentionalism" corresponds to her *general representationalism*, and what I call "strong representationalism" is roughly equivalent to her *standard representationalism*.

[9] For the idea of the phenomenological gap, please see Chapter 1, §2.2, and Chapter 2, §2.1.

their notion of *determination*? Second, why phenomenology (as opposed to other mental properties)? To answer the first question, by "determination", PIT theorists mean a relation whose logical strength is no weaker than that of a *hyperintensional* relation, e.g., grounding, constitution or identity. Since hyperintensionality entails metaphysical necessitation, the central thesis of PIT entails converse intentionalism.

Second, why phenomenology? It is because, as we have seen, for content to match the psychology of a BIV, what PIT theorists need to account for the nature of content is something which is *intimately connected* to the internal states of the BIV, and their answer is that content is metaphysically determined by or identified with phenomenal character. Thus, for PIT theorists, converse intentionalism must be true since, of all the mental states capable of characterizing the BIV's psychology, there seems nothing more intimate to a BIV than its phenomenology. I shall call PIT theorists' aim to find a semantics that matches the content of a BIV's experience to its psychological role "the narrowness concern".

Paradoxically, however, converse intentionalism leads to problems for both theories. Block (1990, 2003) proposes a famous argument, called "the inverted earth argument", which I take to show that there is a counter-example to the conjunction of informational semantics and converse intentionalism, and hence to SR. Suppose on the far side of the galaxy, there is a planet, called "Inverted Earth", on which everything is exactly the same as it is on Earth except that the colors there are the inverted ones of the colors on Earth. Hence, e.g., the sky on Inverted Earth is yellow.[10] Suppose Rose, an Earthling, is transported to Inverted Earth without knowing. During transit, a pair of inverting lenses is inserted which change every ph-color to its inverted ph-color, e.g., exchange ph-green and ph-red, ph-yellow and ph-blue. Hence when she arrives there, everything she sees looks the same as it does on the earth. As Block points out, when Rose looks up into the sky on the two planets respectively, her ph-blue experience

[10] Let us assume whatever theory of color is needed to make such claims.

tokens are caused by light of different colors, one being blue light, the other being yellow light. So according to informational semantics, her experience tokens have the correspondingly different contents, despite having the same phenomenal character. However, in that case, we have a violation of converse intentionalism.

On the other hand, to see that converse intentionalism creates a problem for PIT, we should remind ourselves that PIT is a narrow theory, aimed at accommodating the intuition that content matches psychology. In order for any theory of content to achieve this aim, it has to satisfy something like the following:

> *Psychological supervenience*: For any two experiential state tokens, if their contents are the same, then *ceteris paribus* they play the same psychological role.[11]

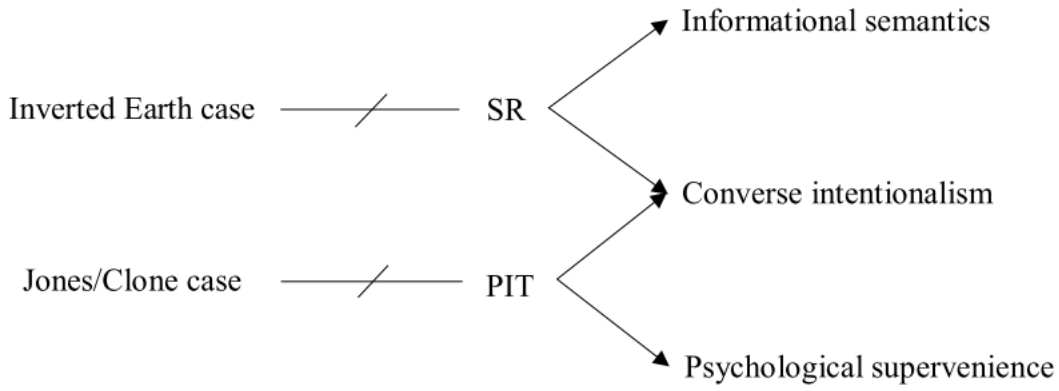
By saying two states "play the same psychological role", I mean that they are governed by the same psychological functions (see Chapter 3, §2). In fact, PIT theorists accept different versions of this principle (Loar 1988a, b, Horgan and Tienson 2002, Pautz 2006a, Mendelovici 2018).

Inspired by Block's case, I constructed in Chapter 3 a counter-example to the conjunction of psychological supervenience (some version of which is part and parcel of PIT) and converse intentionalism (which is entailed by it). Notice that converse intentionalism and psychological supervenience together entail that the *psychological roles of an experience supervene on its phenomenal character*. My case is a counter-example to the italicized claim. Let's begin by observing that some colors look like mixtures of two different colors. E.g., orange looks like the mixture of red and yellow, and purple looks like the mixture of red and blue. Some colors, on the other hand, don't look like mixtures in the slightest, e.g., red, green, yellow, and blue. We call the former kind "binary colors", and the latter kind "unitary colors".

[11] Of course, this principle by itself is unreasonably strong. The versions endorsed by PIT theorists are suitably weaker than psychological supervenience formulated here. My aim here is simply to show that there is a counter-example to the conjunction of converse intentionalism and psychological supervenience, and hence the foundation of PIT's program is not as firm as it appears. See Chapter 3 discussing this issue in detail.

Now consider a possible world where everything is the same except that the evolutionary history of human is different so that the phenomenal character of a unitary color and the phenomenal character of a binary color are switched E.g., yellow looks ph-orange and orange looks ph-red. In effect, the evolutionary history turns the human spectrum around the hue circle by 45 degrees counter-clockwise. Let Clone be a counterpart in that world of an actual person Jones. I argue that the ph-red experiences of Jones and of Clone have different narrow psychological consequences, and hence we have a counter-example to the conjunction of converse intentionalism and psychological supervenience. For example, imagine Jones sees a monitor that looks red to him. Since he is normal, that means that the monitor is illuminated with red pixels. On the other hand, Clone also sees a monitor that looks red to him. But since his spectrum is 45-degree shifted from Jones', the monitor he sees is purple. It follows that the monitor is illuminated with red and blue pixels. Presumably, when both zoom in, Jones will continues tokening ph-red experiences while Clone will token an experience which is partly ph-orange and partly ph-purple since according to his spectrum, red looks orange and blue looks purple to Clone. Accordingly, if they both have ph-red experiences, then if both imagine or simulate in their minds what would happen when they were to zoom in, one of them would token mental imageries with the same phenomenal character while the other one wouldn't (Chapter 3, §5). Hence they would have experiences with the same phenomenal character but different psychological consequences.

We can have a diagrammatic representation of the logical relations among the key principles and cases:



Arrows mean entailing. Crossed lines mean contradicting.

The inverted earth case contradicts the conjunction of informational semantics and converse intentionalism, and the Jones/Clone case contradicts the conjunction of converse intentionalism and psychological supervenience. Since SR and PIT suffer from similar counter-examples, and converse intentionalism is a common commitment of both theories, it is reasonable to suspect that converse intentionalism may be the culprit.[12] The issue this chapter intends to address is thus to find an adequate replacement of converse intentionalism for SR and PIT. Let's use "SR*" and "PIT*" to refer to the theories resulting from replacing converse intentionalism with my proposal in SR and PIT respectively (§3). In §5, I will argue that SR* and PIT* can adequately address the fundamental concerns of SR and PIT while avoiding the counter-examples.

Let's be clear about what I mean by "replacing converse intentionalism". My aim is to find a new principle that can do at least four things: (a) it can help SR theorists address their

[12] Here is one way to justify my conjecture that converse intentionalism is the culprit. Both informational semantics and PIT take the causal roles of intentional states supervene on the contents. To informational semantics, the relevant causal roles are relations to the external environments while to PIT, they are internal psychological relations. On the other hand, phenomenology presumably is an intrinsic property while the causal role of a mental state is relational. The inverted earth case and the Jones/Clone case in fact exploit our fundamental intuition that *relations do not supervene on intrinsic properties*. When converse intentionalism enters into the equation, however, there would be a necessary connection between phenomenal characters and causal roles, which violates our fundamental intuition. As we can clearly see now, the source of all the problems we have so far is *not* about whether content is external or internal. It is rather that *all* semantic theories—SR or PIT, externalism or the narrow content theory—take content to be (related to something) relational while converse intentionalism binds the phenomenal and the intentional, and thereby the intrinsic and the relational. Hence, if we want to keep our fundamental intuition, then we should reject or at least revise, converse intentionalism. Thank Prof. Mendelovici for pushing me to justify my conjecture.

naturalist concern, i.e., it can help to naturalize phenomenology, (b) it can help PIT theorist address the narrowness concern. E.g., it should at least explain the narrowness intuition that the contents of a person's and of a BIV's experiences are the same if they are phenomenal twins. (c) The new principle should avoid the qualia-swapping counter-examples that motivate the replacement. Finally, (d) it should not suffer from new counter-examples that arise merely from the replacement.

Notice that it is not my intention in this chapter to argue that converse intentionalism is objectionable in itself. Nor is it my aim to show that my proposal is true. Nor do I claim that in order to help SR and PIT, there is no other alternative to replacing converse intentionalism. My work in this chapter is principally *exploratory*. I.e., without exhausting all other alternative solutions, I suggest that we can solve the problems of SR and of PIT by replacing converse intentionalism with a weaker principle. It is not my intention to show that the weaker principle is independently true, but to justify its adequacy by exploring its potential of performing the explanatory jobs that were assigned to converse intentionalism by SR and PIT.

§3 Converse Psychointentionalism

To start with, I propose the following as my replacement of converse intentionalism:

> *Converse psychointentionalism*: Every two experience tokens having the same phenomenal characters *and psychological roles* have the same content.

Logically, converse psychointentionalism is weaker than the original version.

I want to emphasize it is not my intention in this chapter to compare the virtues and vices of SR and PIT. Instead, I will argue how, in conjunction with the principles in the diagram of §2.2, converse psychointentionalism can avoid the troublesome color swapping cases (§4). *In none of these discussions do I presuppose either semantic theory*. Of course, that does not mean that my replacement does not have any semantic intimations. On the contrary, as we will see in §5, I will explore some possibilities about how to further develop a semantic theory in accordance with converse psychointentionalism. As I will show, my account can fit into SR's

and PIT's foundational programs respectively.

In the next section (§4), I will show that in any case of color swapping, the relevant capacities of the subjects are characterized by different functions and hence their color experiences do not play the same psychological role. Then since they do not have the same psychological role, the antecedent of converse psychointentionalism is not satisfied, and thus these cases are not counter-examples to converse psychointentionalism.

§4 Defusing Potential Counter-Examples

In this section, I will first argue that converse psychointentionalism is not subject to counter-examples involving color swapping (§4.1). Nor is it likely we will find a counter-example to converse psychointentionalism by shifting around the hue circle. I will discuss the inverted earth case in §4.2.

That the ph-red experiences of Jones and Clone have different psychological roles can be seen immediately because as I have said, Clone's spectrum is shifted from Jones' such that their unitary/binary phenomenal characters exchange. As I have already argued, the ph-red experiences of Jones and Clone have the different psychological roles even though they have the same phenomenal character. So, this is a case where the antecedent of converse psychointentionalism is falsified, and hence not a counter-example.[13]

§4.1 Psychologically Equivalent Color Inversion is Impossible

Let me begin my argument with the following stipulation:

> *Stipulation*: For any distinct agents A and B, if B is a *color invert* of A, then (i) if, when receiving red light, A tokens a ph-red experience, then B tokens a ph-green experience, and (ii) if, when receiving green light, A tokens a ph-green experience, then B tokens a ph-red experience.

In this subsection, I will argue that if B is a color invert of A, then either B's color experience

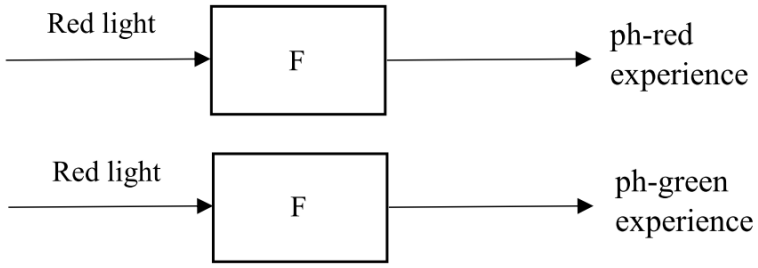
[13] I.e., no counter-example to the principle implied by converse psychointentionalism and psychological supervenience, which has the antecedent of the former and the consequent of the latter. Cf., §2.2.

is not psychologically equivalent to A's or they have different phenomenology. Hence any case involving color inversion falsifies the antecedent of converse psychointentionalism, and is therefore not a counter-example.

Consider an arbitrary agent Rose* whose ph-red experience has the same psychological role as Rose's ph-red experience but has an inverted spectrum from Rose's, i.e., Rose* is a color invert of Rose.

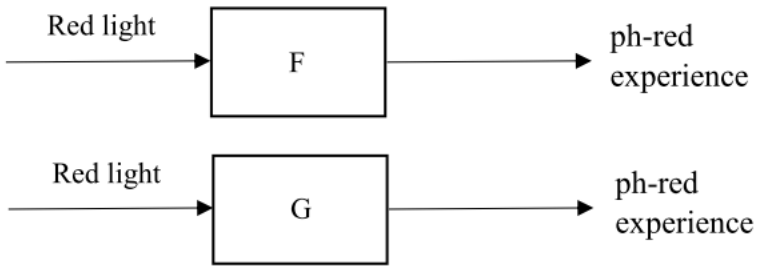
Now suppose Rose and Rose* receive the same red light. The light causes experiences with different phenomenology. We can draw a diagram for this case (call it "A"):

Case A



("F" denotes the functionality between the light and the experience. The arrows and the box at the top represent the tokening of Rose's color experience, and the ones at the bottom represent the tokening of Rose*'s color experience.) Notice that Case A is not a counter-example to converse psychointentionalism since to be a counter-example, the phenomenology of Rose and Rose* should be the same, but in Case A, their experience tokens have different color phenomenology. Similarly, the case where Rose's functionality and Rose*'s are different but their experience tokens have the same phenomenology (call this case "B") is not a counter-example either:

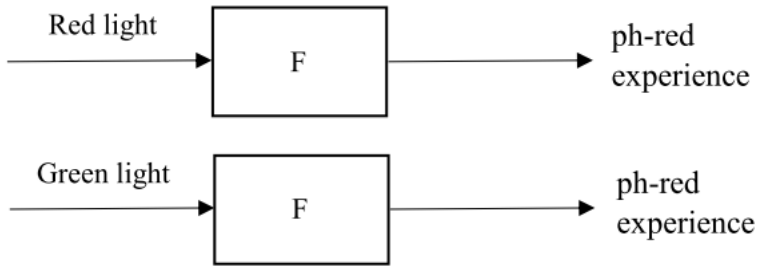
Case B



Case B is not a counter-example to converse psychointentionalism because their functionalities are different.

To get a real counter-example, my opponent needs a case in which Rose and Rose* have the same functionality and their experience tokens have the same phenomenology (call this case "C"). Because they are inverts, they can only have the same phenomenology if they receive light of opposite colors. Hence, C must look like this:

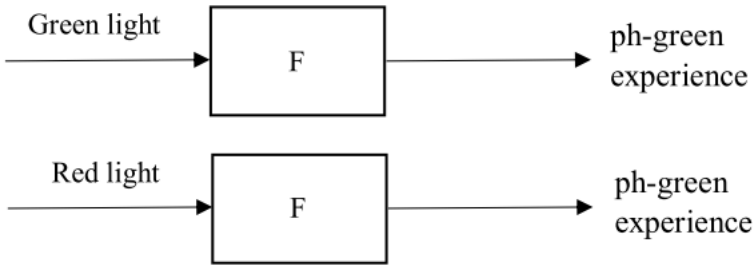
Case C



That is to say, the functionality F has a many-to-one mapping, in particular, F(red light)=F(green light)=ph-red experience.

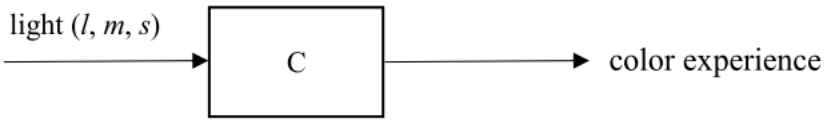
But now, given this one-to-many mapping, what would happen if Rose receives green light and Rose* receives red? Since each of them receives the opposite light to what they receive in C, they should each token experiences with the opposite phenomenology to what they tokened in C. So we should get Case D as in the diagram below:

Case D



But Case D entails that the functionality F has a different mapping. In particular, whre in Case C F(red light)=F(green light)=ph-red experience, in Case D F(red light)=F(green light)=*ph-green* experience. Since it is necessary that no ph-red experience token is a ph-green experience and *vice versa*, Case C and Case D make contradictory demands on the functionality of F.

Let me illustrate my argument with the contemporary cognitive science of color experience. Let's use variables *l*, *m*, and *s* to represent the quantity of light entering the eye the long, medium, and short wavelengths respectively. See the diagram below:



Then according to the opponent process theory in color science[14], the nervous system computes the function C such that:

C(*l*, *m*, *s*)=ph-red experience iff $l-m>0$ and $l+m-s=0$

C(*l*, *m*, *s*)=ph-green experience iff $l-m<0$ and $l+m-s=0$.[15]

If Rose and Rose* are psychologically equivalent (i.e., they have the same functionality C), and if both of them receive red light, C should return the same phenomenal experience for both. That means that they are not inverts. If on the other hand, they are inverts and they both receive the same red light, then they cannot be psychologically equivalent. Rose* experience must be mediated by a function C* such that C*(*l*, *m*, *s*)=C($-l$, $-m$, $-s$). Either

[14] Here I use the simplified model provided by Hardin (1988).

[15] Rose would token a ph-yellow experience iff her $l-m=0$ and $l+m-s>0$, and she would token a ph-blue experience iff $l-m=0$ and $l+m-s<0$.

way, it is not a counter-example to converse psychointentionalism involving color inversion since it violates the antecedent of the principle. Similar considerations can be used to show, *mutatis mutandis*, there is no counter-example involving other kinds of color shifts.

A final remark: notice though I have illustrated this argument using the opponent process theory, it does not rely on it, nor on any other any psychophysical theory of color vision. The role of these empirical theories is merely to flesh out the details of the relevant functionality, and thus provide a concrete example. As we have seen, my argument is independent from these details.

§4.2 Inverted Earth

Let's move onto the inverted earth case. A key worry relevant to my later argument is: where are the lenses inserted? If they are put immediately in front of Rose's eyes, then obviously the light, reflected from the sky, inverted by the lenses, coming into her eyes, is in fact blue light, not yellow. So arguably, her ph-blue experience is still normally caused by blue light after all. Hence, by some simple version of informational semantics, the content of Rose's ph-blue experience on the inverted earth is still blueness, not yellowness.

For Block's argument to work, the inverter thus needs to be placed "inside" the perceiving system. Anticipating this worry, Block replies,

> You give me an independently motivated conception of the boundary between the inside and the outside of the system, and I will tailor an inverted earth example to suit….You will recall that I mentioned an alternative to the inverting lenses,…, namely a neural inverter behind the retina….I don't know any reason why it shouldn't be in principle possible for a miniaturized silicon chip to register…impulses [in the optic nerve] and substitute impulses for them (1990, pp. 71-72).

It is clear that Block's argument relies on the premise that the inverter is put inside the system, and changes the internal state of the system.

As Bartlett (2008), and Ren (2016) point out, for Block's argument to succeed, the inverter has to be pushed inward such as to partly constitute the neural state that realizes the

experiential state token. Suppose it doesn't. Let "Yellowbrain" be the brain state of a normal person that processes the information of yellow light, and that causes the ph-yellow experience, and "Bluebrain" be the corresponding one for blue light, causing the ph-blue experience. If the inverter does not (partly) constitute the neural state that realizes Rose's ph-blue experience, then it must change Yellowbrain into Bluebrain before her ph-blue experience is tokened. Hence her ph-blue experience is still caused by Bluebrain, not by Yellowbrain. This implies that her ph-blue experience is caused by the brain state that normally processes the information about blue light, and thus arguably means blue, rather than yellow. In contrast, if the inverter (partly) constitutes the neural state that realizes her ph-blue experience, then her ph-blue experience starts to exist no later than the inverter is triggered. Hence the ph-blue experience can be said to be caused by Yellowbrain, and can be said to mean yellow.

Let's use the term "Rose_e" to refer to Rose on the earth, and "Rose_i" to mean Rose with the inverter on Inverted Earth. I want to argue that the ph-blue experiences of Rose_e and Rose_i have different psychological roles, and therefore, they are psychologically inequivalent. As we have seen, in Rose_e, her ph-blue experience is caused by Bluebrain, and does not involve the inverter while in Rose_i, her ph-blue experience is caused by Yellowbrain and involves the inverter. Now, consider this question: are Bluebrain in Rose_e and Yellowbrain in Rose_i functionally identical? Either they are or they are not. If they are functionally identical, then they normally process the information about the same light. There are two possible results: either (i) Bluebrain in Rose_e normally processes the information about yellow light as Yellowbrain in Rose_i does or (ii) Yellowbrain in Rose_i normally processes the information about blue light as Bluebrain in Rose_e does. Either way, if informational semantics (which is entailed by SR) is true, then their ph-blue experiences mean the same, verifying the consequent of converse psychointentionalism. Consequently, Block's argument fails to establish his intended conclusion. On the other hand, if Bluebrain in Rose_e and Yellowbrain in Rose_i are not functionally identical, then the ph-blue experiences of Rose_e and Rose_i have different

psychological roles since the experiences have different (functionally individuated) psychological antecedents. Consequently, their ph-blue experiences are psychologically inequivalent. Thus, the conjunction of converse psychointentionalism and informational semantics is immune from the Inverted earth case.

Some people might object to my claim of a psychological-role difference by saying that in Block's setting, it is stipulated that everything looks indistinguishable to Rose before and after landing on Inverted Earth. If there is a psychological difference, then that difference is inaccessible. But (it might be claimed) every psychological difference *is* accessible. Therefore, my claim of a psychological difference between Rose_e and Rose_i is absurd. However, the principle that every psychological difference is accessible is questionable as shown by arguments from empirical science. For example, in a series of experiments involving priming, it is shown that subjects' behavioral dispositions are changed when they are presented with pictures of some objects. The effects are robust even if the subjects are presented with the pictures within very short time periods—so short that the subjects are not aware of them (Bermúdez 2014). Hence whether a person is primed or not brings about a psychological difference, but that difference is not accessible to the subjects. Hence, even though everything looks indistinguishable to Rose before and after landing on Inverted Earth, it is not implausible to say that there is a psychological difference between her ph-blue experiences on either planet, and hence it is not a counter-example to converse psychointentionalism.

§5 Renderings

As I have said in §2, converse psychointentionalism should achieve four goals: (a) address the naturalist concern, (b) address the narrowness concern, (c) should not suffer from qualia-swapping counter-examples discussed in my dissertation, and (d) should not generate more problems resulted from the replacement. I have verified (c) in §4. In this section, I want to show that converse psychointentionalism can indeed achieve (a) and (b). I will not discuss (d) until chapter 5.

Let's start with (a). Consider strong representationalism. Extending my account, we can have the following theses:

> *Weak psychointentionalism*: phenomenology and psychological role supervenes on content, i.e., every two experience tokens having the same content have the same phenomenology and psychological role.
>
> *Strong psychointentionalism*: weak psychointentionalism and converse psychointentionalism are true.[16]

In effect, logically what strong psychointentionalism says is that for any two experience tokens having the same psychological role, the content and the phenomenology mutually supervene on each other. The next question is: how is it to be understood? As I have said, the primary concern for SR theorists is to naturalize phenomenal consciousness (§2.2), and they did so by endorsing the identity thesis that phenomenology is identical to intentionality. Following the same spirit, one way of rendering psychointentionalism to fit SR's program and to address the naturalist concern is thus to take it as saying that intentionality is identical to the conjunctive property of phenomenology and psychological role. Since as we have seen, informational semanticists argue that an intentional property is a physical property, it would follow that the conjunctive property of phenomenology and psychological role is too. And since it is plausible that if a conjunctive property is physical, so are its conjuncts, that would imply that phenomenal properties are physical. Hence there is no hindrance to naturalizing phenomenal properties if we adopt this conjunctive identity thesis.[17]

---

[16] This extension mimics other two theses in the literature on SR: weak intentionalism (that phenomenology supervenes on content) and strong intentionalism (that weak and converse intentionalism are true) (see note 6 and Lycan 2000/2019).

[17] One might be wonder how this conjunctive identity thesis interacts with *the transparency data* adduced by SR theorists to argue for the original identity thesis (Tye 2000, Siewert 2004, Molyneux 2009, Speaks 2009, 2015). As far as I know, no formulation of the transparency data is incompatible with the conjunctive identity thesis I propose here. For example, consider the formulation of the transparency data in Molyneux 2009: "We experience the properties of which we are directly aware as properties of external surfaces and objects and not as properties of experience itself." Taken literally, this datum only suggests that phenomenal properties are properties of external entities (i.e., they are part of the intentional content), but it does not very specifically say *which* properties of the external entities the phenomenal properties are *identical to*. Nor does it make an identity claim of any sort. I.e., although according to the transparency data, when I introspect the phenomenal character of my color experience, the only property I am directly aware of is the color of the external object, not the shape, it does not

How does converse psychointentionalism fit in PIT's program and how does it address the narrowness concern? Recall that in their program, semantics should be able to accommodate the narrowness intuition. For this reason, they argue that intentionality is determined by phenomenology (§ 2.2). Accordingly, I recommend that converse psychointentionalism is rendered as saying that experiential intentionality is determined by psychological role *and* phenomenology. Since phenomenology and psychological roles are both narrow, converse psychointentionalism is still a narrow theory and should be palatable to most PIT theorists.

Notice that unless psychological roles are determined by phenomenology, converse psychointentionalism implies that experiential content is not wholly determined by phenomenology, and hence is not phenomenal intentionality. This means that some strong version of PIT is false. To be more articulate, Mendelovici and Bourget (2014) distinguish three versions of PIT:

> *Weak PIT*: There exists phenomenal intentionality.
>
> *Moderate PIT*: Weak PIT is true and all intentionality is either phenomenal intentionality or derived from phenomenal intentionality.
>
> *Strong PIT*: Weak PIT is true and all intentionality is phenomenal intentionality.

Now *if* psychological roles are not determined by phenomenology, converse psychointentionalism cannot satisfy *strong* PIT. Some PIT theorists may thus complain that my replacement is disappointing, and insist on returning to converse intentionalism. However, as we have seen in §2.2, the Jones/Clone case already falsifies the idea that experiential content is wholly determined by phenomenology. So we have an independent reason to reject strong PIT.[18] Therefore, although only converse intentionalism satisfies strong PIT, it doesn't matter

say *which particular color* I am directly aware of. It is thus not incoherent to endorse both the transparency data and the conjunctive identity thesis.

[18] Relatedly, though their works are not about experiential content, Chalmers (2018), Yli-Vakkuri and Hawthorne (2018) both use Twin-Earth style cases to point out that there exists at least one *thought state* whose content is not exclusively determined by its qualitative character, which constitutes another reason against strong PIT.

much since strong PIT is false.

I would like to emphasize that (converse) psychointentionlism is only a theory about *experiential* content. It is silent about other kinds of intentionality, e.g., it says nothing about the contents of beliefs or desires. However, very crucially, converse psychointentionalism does not imply that there are no phenomenal intentional states. All it says is that experiential content is not one of them. Hence, it is compatible with both weak PIT and moderate PIT.

Still, people might be worried that if experiential intentionality is not wholly determined by phenomenology, how could we have a theory of intentionality that accommodates the narrowness intuition about experiential content? To answer this question, I want to distinguish two cases. In the literature, PIT theorists typically use two kinds of thought experiments to pump the narrowness intuitions: one appeals to a disembodied Cartesian ego, and the other involves a BIV. Let's start with the BIV case. Consider a BIV which is a molecule-by-molecule duplicate of your brain. As many PIT theorists point out, if the BIV tokens an experience which is indistinguishable from yours, then it is intuitive that the content of its experience is the same as yours (Loar 2003, Horgan, Tienson and Graham 2004, Kriegel 2013b, Mendelovici and Bourget 2014). It seems in making this argument, we are appealing to converse intentionalism or intuition to that effect. However, I want to point out that we can explain the intuition by converse psychointentionalism in conjunction with the assumption (call it "A") that *the experience tokens of you and your BIV duplicate have the same narrow psychological role*. If A is true, and since the experience tokens of you and your BIV duplicate have the same phenomenal character, then by converse psychointentionalism, your experience tokens thus have the same content. Notice that assumption A is reasonable because the BIV is a molecule-by-molecule duplicate of your brain, and you two should share exactly the same narrow brain properties.

What about the Cartesian ego case (Horgan and Tienson 2002)? Consider a normal person and her disembodied phenomenal duplicate. Do their experiential tokens have the same

content? Notice that in this case, we do not have reason to assume that their experiential tokens have the same psychological role since the subjects have entirely different ontological constitutions. Nevertheless, in this case, I do not think the intuition that their experiential tokens have the same content is as compelling as in the BIV version. Recall the Jones/Clone example. Suppose now Clone is a Cartesian ego. At a moment *t*, Jones and Clone both token ph-red experiences, and at the next moment *t*+1, the tokens cause them to have different mental states. It does not seem implausible to say that the experiential tokens at *t* have different contents. Indeed, without considering a disembodied ego, if some person like Clone is capable of tokening ph-red experiences, but consistently attributes the content of the ph-red experience to purple grapes and not to apples, it is plausible to say that the content of her experience is different from the one we have when we both token ph-red experiences. Therefore, even though converse psychointentionalism cannot accommodate the narrowness claim about experiential content in the Cartesian case, it does not need to since the narrowness claim in this case is not sufficiently plausible.

Some may say that the explanation by converse intentionalism is better than the explanation by converse psychointentionalism plus the same-psychological-role assumption since the former requires one principle but the latter needs two. However, simplicity is only a theoretical tie-breaker when we have two equally adequate theories. Of two theories that do the job, we should choose the simpler. Converse intentionalism, however, is so strong that in conjunction with psychological supervenience, it contradicts the Jones/Clone case. Converse psychointentionalism on other hand strikes the balance: it is not too strong so that in conjunction with psychological supervenience, it is compatible with the Jones/Clone case, and is not too weak so that with appropriate assumptions, it can accommodate our narrowness intuition. Overall, converse psychointentionalism can provide better explanations, and should be recommended to PIT.

§6 Conclusion

In this chapter, I defend a principle, called "converse psychological intentionalism", which is intended to be a replacement for converse intentionalism, endorsed by the two standard semantic theories of experiential intentionality, strong representationalism (SR) and the phenomenal intentionality theory (PIT). In §2, I provide the background for discussions. I explain why both SR and PIT need converse intentionalism, the counter-examples to it, i.e., the inverted earth case, and the Jones/Clone case. In §3, I present converse psychointentionalism. In §4, I show how converse psychointentionalism can satisfactorily avoid the counter-examples to the original converse intentionalism. In §5, I provide some renderings of psychointentionalism, and show how under these renderings, it can fit in SR's and PIT's programs. If my arguments are correct, then semanticists should abandon converse intentionalism and stop being worried about the inverted/shifted spectrum, and they would do more fruitful work with psychointentionalism.

Chapter 5

Conclusion: Objections and Replies

§1 Introduction

The aim of this chapter is to address some worries and objections which do not belong to any of the previous chapters. In §§2−3, I address the worries that since in Chapter 4, I claim that converse intentionalism is the culprit for the problems for SR and PIT, some people may say that I should exhaust other possible semantic theories while keeping converse intentionalism. In these two sections, I will discuss two specific semantic theories, functional role semantics and the phenomenal appearance theory. In §§4−5, I clarify some possible misunderstandings about my argument for psychointentionalism and relatedly show how psychointentionalism could be incorporated with informational semantics. In §6−8, I address some lingering worries from previous discussions. Finally, I conclude this chapter with a short remark on external-world skepticism.

Convention: Terms with the prefix "ph-" denote phenomenal characters, e.g., ph-red is the phenomenal character she has when she sees red objects.

§2 Functional Role Semantics

In Chapter 4, I argued that psychointentionalism can satisfactorily address SR's and PIT's primary concerns. Some people might complain that I did not explore other theoretical possibilities. What if, as they might say, there is another theory of experiential content that can better address SR's and PIT's concerns? If so, does it presuppose converse intentionalism? Or does it abandon converse intentionalism and converse psychointentionalism altogether? If, as in the former case, there is a theory which presupposes converse intentionalism yet successfully addresses both concerns, then my claim in Chapter 4 that converse intentionalism is the culprit for the problems of SR and PIT is undermined. If, as in the latter case, there is theory that does not presuppose psychointentionalism, then it seems that psychointentionalism is dispensable after all. In this and the next sections, I will take up these challenges. In particular, for the

former case, I will consider *the phenomenal appearance theory* ("PAT") which presupposes strong intentionalism. For the latter case, I will discuss Harman's account of experience which is the conjunction of functional role semantic ("FRS") and weak intentionalism. I will raise some *prima facie* challenges to both theories if they are applied to addressing the concerns of SR and PIT. Note that I don't intend to give knock-down arguments against FRS and PAT. I only want to show why neither of them looks promising, at least *prima facie*, in addressing those concerns.

Recall that the primary concern of SR is the Naturalization Project, in particular, to fill in the phenomenological gap (Chapter 2, §2.1), and the concern of PIT is to give an account of the content of a mental state that matches the psychological role of that state. In particular, its verdict in the Frege case is that the relevant states have different contents and in the BIV case, its verdict is that they have the same content. I will call SR's concern "the naturalist concern", and PIT's concern "the narrowness concern". I will evaluate FRS and PAT with respect to these two concerns.

Let's begin with FRS. FRS is the theory that the content of a mental state supervenes on its functional role (Harman 1987, Loar 1988a, b, Block 1986, 1998). Conventionally, FRS theories are classified into two versions. The first one is the so-called "long-arm FRS", and the second one is "the short-arm FRS". The difference is whether the entity to which the mental state stands in a functional relation is inside the skin/skull or not. Long-arm FRS allows that entity to exist outside the skin whereas short-arm FRS insists that it can only be inside the skin. I will evaluate both theories with respect to the narrowness concern first, and then the naturalist concern.

To begin with, notice that it is in fact unclear how a long-arm theorist would make sense of the narrowness concern. For the very reason why a long-arm theorist takes functional relations to entities outside the skin to be semantically relevant is because according to long-arm FRS, the distinction between the internal and the external is *arbitrary* (Harman 1987,

1990). Hence even if long-arm FRS could address the narrowness concern, it is not clear the solution it provides is really satisfactory to narrow theorists. In contrast, short-arm theorists do concede the distinction between the internal and the external. Predominant short-arm theorists, e.g., Block (1986, 1988), Loar (1988a, b), typically hold a *two-factor* view of semantics. The first factor is the causal relation between the mental state and an external entity, which determines *reference*, and the second factor is the short-arm functional roles of the mental state which determines its *content, properly called*. Short-arm theorists thus insist that content is narrow, and hence it seems that they are capable of addressing the narrowness concern.

To test this, let's think about how they would argue that the experiences of me and of a BIV which is my phenomenal/physical duplicate have the same content. Since they hold that content supervenes on the short-arm role of experience, they have to somehow argue that the experiences of me and the BIV have the same functional role. One way is to assume *phenomenology/psychology supervenience* ("PPS", see Chapter 3, §2), but in Chapter 3, I already argued that PPS is false. Another way is to motivate the assumption that the experiences of me and the BIV have the same functional role by arguing form the fact that the BIV and me are physical duplicates.

However, the main problem of short-arm FRS is that functional role by itself is too weak to determine experiential content. This is called "the problem of content indeterminacy" in the literature (Horgan and Graham 2012, Mendelovici 2018, pp. 72−76; Mendelovici and Bourget 2020, pp. 568−570). We can fully characterize the functional roles of an experiential state without assigning it a determinate content. Indeed, it is the problem of content indeterminacy that convinces many narrow theorists to reject FRS and embrace PIT. Notice that I am not saying that short-arm FRS is *fundamentally* flawed given the problem of content indeterminacy. What I am saying is that functional role is too weak to determine the kind of content that we typically expect experience to have, i.e., a determinate property. After all, some

short-arm FRS theorists, e.g., Block, suggest the content determined by functional role is not a determinate property but a function from contexts to properties.

Let's move on to the naturalist concern. To naturalize mentality, an FRS theorist could have two routes. The first is to naturalize phenomenology, and then intentionality, and the other is to naturalize intentionality and then phenomenology. We already see the reason why the first route cannot work, i.e., the phenomenological gap. As Levine (1983) argues, the functional role of a mental state does not determine its phenomenal character. In addition, we intuitively resist the identification between functional role and phenomenology. As long as functional role is the only resource we have, the phenomenological gap is still open. Hence that leaves us only the second route.

The second route is to naturalize content first and then use weak intentionalism (Chapter 2, §2.2) to naturalize phenomenology. Unfortunately, we already see that short-arm functional role is not sufficient to determine content. So the second route is not available to short-arm FRS either. How about long-arm FRS? As I have said, long-arm theorists obliterate the distinction between inside and outside. So even if they can naturalize phenomenology by the second route, it is not clear their notion of phenomenal character is what we intuitively have. For example, given our intuition of phenomenology, a BIV which is a molecule-by-molecule of my brain should, quite intuitively, have the same phenomenal character as mine wherever it is located. However, since long-arm theorists insist that environments play crucial roles in determining phenomenology, they would say that in a thought experiment like this, without specifying the long-arm functional roles, any verdict about what phenomenal character the BIV has is arbitrary (Harman 1987). To reiterate, I am not saying that long-arm FRS is wrong about phenomenology. They need more work to flesh out this view. However, I am not optimistic about a view that does such a violence to our ordinary intuitions.

§3 The Phenomenal Appearance Theory

PAT says that every sensible property of an external object is associated with what is called a "phenomenal appearance", "phenomenal look", or "phenomenal seeming". Suppose I go to a store to buy a blue shirt. However, due to its misleading lighting condition, I buy a white one. When I walk out of the store, and you see I bought a white one, you say: "I thought you wanted a blue shirt." I reply: "I did, but this shirt *looked blue* to me."[3] In this context, intuitively, the sentence "this shirt looks blue" is true. If so, then plausibly, it cannot be because of the shirt's color since it is white. It follows that the content of the sentence is not dependent on the color, but something else. This extra ingredient is what PAT theorists call the phenomenal look or the appearance of the shirt.

PAT theorists of experience argue that the same thing applies to experiential content. When I see the white shirt in the store, in a sense my visual experience is veridical since part of the content of my experience is the blue-appearance of the shirt (Shoemaker 2006, Glüer 2009, 2012, 2014, Brogaard 2014a, b). Some PAT theorists, e.g., Shoemaker (2000, 2006), hold that a visual experience has two different contents. One is about the color of the relevant object, and the other is about the phenomenal appearance of the object. Hence, in our example, my experience in the store is both veridical and non-veridical. With respect to color, it is not veridical, but with respect to appearance, it is. Some other PAT theorists, e.g., Glüer (2016, 2018), think that there is only one content, the appearance content, and hence experience is always veridical. Later, I will only focus on Shoemaker's account, and evaluate his PAT in accordance with the narrowness concern and the naturalist concern.

Why does Shoemaker endorse PAT? As he says (2006), there are three requirements. The first one is transparency according to which whenever we introspect our phenomenal character, we always experience it *as a property of external objects*. Transparency convinces

[3] This vignette is adapted from Glüer 2016.

him that appearance is a property of physical objects. Second, this identity thesis convinces him of strong intentionalism, i.e., content and phenomenology co-supervene on each other. Third, Shoemaker believes that *color inversion without misrepresentation* is possible. Specifically, it is possible that when two people, A and B see the same object under bright lighting conditions, they token experiences with inverted color phenomenology, and both their experiences are veridical. Putting it all together, Shoemaker argues that a coherent way to satisfy all three requirements is to hold the view that color-appearance is part of the content of experience.

To elaborate, according to Shoemaker, (i) appearance is a property of the external object, and (ii) the phenomenal character is identical to the appearance content. If the relevant content is appearance, then (i) can satisfy transparency. Furthermore, in the above case of color inversion, weak intentionalism implies that A's and B's experiences have different contents, and Shoemaker holds that both experiences are veridical. (i) and (ii) can explain why this is the case. Given (i), a green object has two color appearances, green-appearance and red-appearance. One of the observers, e.g., A, tokens a ph-green experience, and the other, B, tokens a ph-red experience. Shoemaker's theory holds that given (ii), the content of A's experience is green-appearance while the content of B's experience is red-appearance. Hence, both experiences are veridical. Similarly, consider a case of color inversion in which the observers token experience with the same phenomenology, e.g., ph-red, but different referents. (i) and (ii) can explain why both experiences are veridical by arguing that the objects, though having different colors, have the same color-appearance.

What is color appearance? In particular, what is red-appearance? Shoemaker's (2000) account, as reconstructed in Speaks 2015, is that:

> <u>*Red-Appearance Disposition*</u>: Red-appearance is identical to the disposition to produce ph-red experience in any agent of type S in circumstance C (cf., Speaks 2015, chapter

22).[4]

Now we can see why it seems not promising to address the narrowness concern with Shoemaker's PAT even if it presupposes converse intentionalism. To make our discussion concrete, let's stipulate C to be "under the bright Sun". The next question is what S could be? As Speaks points out, S cannot be too inclusive. If we type-individuate my color invert and me as of the same type, then no object can instantiate red-appearance. For by *Red-Appearance Disposition*, that object has to produce in me and my invert ph-red experiences under the bright Sun, but by definition, my invert would token a ph-green experience. If no object can ever instantiate red-appearance, then neither of our ph-red experiences could be veridical whenever we token them. Then it follows immediately that color red representation without misrepresentation is impossible. *A fortiori*, color inversion without misrepresentation is impossible, contrary to what Shoemaker's PAT is expected to do. If, on the other hand, S is exclusive—S only refers to normal Earthlings, say—then green objects cannot instantiate red-appearance because they cannot produce ph-red experience in normal Earthlings under the bright Sun. Now consider a case of color inversion in which I see a red object and my invert sees a green one, and we both token ph-red experiences. In that case, there are two possibilities. First, our experiences have the same content, i.e., red-appearance, but my invert's is not veridical. But in that case, it seems puzzling how color inversion without misrepresentation could *ever* be possible. Second, to make my invert's experience veridical, we have to say that her experience has a different content. E.g., the green object she sees has the disposition to produce ph-red experience in *Inverted Earthlings* (not normal Earthlings) under the bright Sun, with a corresponding difference in content. However, this contradicts converse intentionalism since the phenomenal character of our experiences is the same but the contents are different.[5]

[4] Shoemaker (2000, p. 466): an object's having a color appearance is "its looking a certain way to *certain* perceivers in virtue of having a certain color", and appearance properties are dispositions that "things have in virtue of producing (in *certain* circumstances) experiences of certain sorts…" (emphases added).

[5] My argument in this paragraph is an elaboration on Speaks' (2015, chapter 22).

If Shoemaker's PAT can successfully address the narrowness concern of PIT, then, in this particular case, it should at least be able to confer the verdict that the contents of my invert's and my personal experiences are the same, as is embraced by PIT. Since as we have seen, it isn't, it is thus hard to see how Shoemaker's PAT can provide solutions to PIT's narrowness concern more generally.[6]

Let's consider how PAT fares with respect to addressing the naturalist concern. I don't think we should be optimistic in this regard. This is due to a widely known problem for PAT, called "the circularity problem" (Shoemaker 2006, Glüer 2012, Speaks 2015). The problem is that for PAT theorists, there is no non-circular explanation of phenomenal character. To see this, recall that PAT theorists also endorse the identity thesis between phenomenal character and intentional content where the relevant content is appearance. From this we can infer that:

> *Equation 1*: The ph-red character of an experience E is identical to E's intentional property of *having red-appearance as part of its content*.

By Red-Appearance Disposition, we can derive the equation that

> *Equation 2*: E's property of *having red-appearance as part of its content* is identical to E's property of *having as part of its content the disposition to produce ph-red experience in any agent of type S in circumstances C.*

Enter equation 2 into equation 1 in the place of "ph-red" on the right-hand side, we get

> *Equation 3*: The ph-red character of an experience E is identical to E's property of *having as part of its content the disposition to produce ph-red experience in any agent of type S in circumstances C.*

As we can see, the notion of ph-red is on the right-hand side of equation 3. We can keep iterating

[6] People might think that the reason why Shoemaker's PAT is ill-equipped for the narrowness concern because of Shoemaker's personal insistence on the possibility of inversion without misrepresentation. If, as some might wonder, we drop this insistence, maybe some version of PAT can address the narrowness concern. As far as I know, Glüer's view is only one of the PAT theories on the flip side of Shoemaker's in this regard. She insists that inversion without misrepresentation is impossible (2012). However, to rule out the possibility, she proposes a long-arm functional account of phenomenology which implies phenomenal externalism. Hence, just like long-arm FRS, even if Glüer's PAT can "solve" the narrowness concern, it is not clear the notion of phenomenal character is relevant to our discussions.

this process and the notion of ph-red still remains. That means the notion of phenomenal character is not cancellable. If we want to use equation 1 to explain the notion of ph-red and use equation 2 to explain the notion of red-appearance, then we end up with a circular explanation of both.[7]

Given the circularity result, it is hard for me to see how PAT by itself can address the naturalist concern, i.e., to fill in the phenomenological gap. Notice that PAT and SR agree on the identity thesis between phenomenology and experiential intentionality, and consequently, they all endorse converse intentionalism. In addition, as we have seen, historically SR made important progress in filling the gap which has baffled traditional physicalists. In this respect, PAT seems to drag us back to where we were by holding experiential content which is identified with phenomenology to concern appearances but holding a view of appearances in which phenomenology is non-cancellable. Hence even if converse intentionalism is presupposed, PAT preserves the phenomenological gap, and the naturalist concern is still with us.

Two remarks before we move on: first, as I previously said, I do not intend to provide a knock-down argument that PAT cannot address the narrowness and the naturalist concerns. Rather, I only want to point out that these tasks are not as easy for PAT as it seems in the first instance. Second, to repeat, my goal in this section is to explore whether there is a way to address the narrowness and the naturalist concerns while keeping converse intentionalism. To this aim, I use PAT as my case study, and as I have argued, it doesn't seem promising that PAT would meet the requirements. Previously, I said that there was a legitimate worry that my claim of converse intentionalism as the culprit for our problems may be undermined if we have a

[7] Shoemaker (2006) doesn't think it is a problem. He fully embraces the circularity result. Glüer (2012) attempts to solve this problem, but the crucial technique in her work is to use linguistic characterizations for ph-red in equation 2. Roughly speaking, instead of saying the relevant disposition as *producing ph-red experience*, she characterizes it as producing a type R of experience in the agent where R is a type of experience such that the agent would normally call the object "red" iff R is tokened by the agent. As she admits herself, this is only a semantic account for the sentence "*x* appears red", not a metaphysical account of red-appearance. Since the naturalist concern is a metaphysical one, even if Glüer's account succeeds in solving the circularity problem, more work needs to be done to address the concern.

theory that solves the problems and keeps converse intentionalism. Now I conclude that this worry is at least alleviated to some degree.

§4 Psychological Role and Color Inversion

I have been presupposing the narrow notion of psychological role. However, some may object that my argument in Chapter 4, §4 presupposes the wide notion of psychological role since the relevant function is individuated in terms of external inputs, i.e., light rays. Hence, as my objector might say, my argument would collapse if the relevant function is individuated narrowly. In particular, my objector might say if the relevant function is narrow, then there may still be a case in which the agents' spectra are inverted but their ph-red experience tokens have the same psychological role.

To make our discussion more concrete, let's divide the whole process of producing color experience into two parts, processing done by the eyes, and that done by the brain. Then my argument in Chapter 4, §4 effectively shows that if the eyes and brains of Rose and Rose* are governed by the same functions, Rose and Rose* cannot be inverts to each other. However, as my objector says, since the function of the eyes is individuated in terms of the light, the states of the eyes should not be counted as internal states of Rose or Rose*. As in Block's original thought experiment, the inverter only changes the functioning of Rose's eyes without changing the functioning of her brain. So my objector may say we still have a case of psychologically equivalent color inversion according to my stipulation of color inversion in Chapter 4, §4. If every instance of psychologically equivalent color inversion is a counter-example to converse psychointentionalism, then we seem to have a case falsifying converse psychointentionalism.

Earlier, for simplifying our discussion, I did assume that every instance of psychologically equivalent color inversion is a counter-example to converse psychointentionalism. Now, I think it is time to remove the ladder, i.e., I don't think all instances of psychologically equivalent color inversion are counter-examples to converse

psychointentionalism. I.e., even if there is an instance of psychological equivalent color inversion, it does not follow that the experience tokens have different intentional contents whether the correct semantics is informational semantics or PIT. Suppose informational semantics is correct. As we have seen, Block already admitted that when the inverter is installed outside, e.g., in the eyes, informational semanticists can justifiably say that the experience tokens still have the same content. For as they may say, the experience tokens of Rose on Earth and on Inverted Earth, though having different actual causes, have the same normal cause. On the other hand, suppose PIT is correct. Then it just immediately follows that the ph-red tokens of Rose on the different planets have the same content since converse psychointentionalism is logically entailed by the central thesis of PIT.

To summarize, if the inverter is installed inside, then the relevant tokens have different narrow psychological roles, and hence they do not constitute a counter-example to converse psychointentionalism. If the inverter is installed outside, then it is not obvious that the relevant tokens have different contents, and hence converse psychointentionalism still holds. Wherever it is installed, there is no obvious threat to converse psychointentionalism (cf., Bartlett 2008, Ren 2016).

§5 Psychological Role and Informational Semantics

People might complain that my claim is too suggestive when I say that "if the inverter is installed outside, informational semanticists can justifiably say that the experience tokens still have the same content" in the last section. To address this worry, I will illustrate my claim using Fodor's informational semantics.

Now consider Rose with inverting lenses many years after she arrives at Inverted Earth. As Block argues, everything looks indistinguishable for Rose on both planets. He even assumes that the language in Inverted Earth is also inverted. Given her environment on Inverted Earth, Rose cannot detect that the inversion occurs.

Nevertheless, this does not mean that there is no difference in the psychological role

of her ph-blue experience. Recall the idea of the identity of psychological role invoked throughout my dissertation:

> Every two experience tokens play the same psychological role iff they are governed by the same psychological functions.

To argue that there is a difference in the psychological role of her ph-blue experience, we only need to find two situations such that Rose is in the same background mental states, then she tokens a ph-blue experience in these situations, and consequently she tokens different experiences or beliefs…etc. Now suppose Rose is a scientist of optics, and she has the knowledge of how the spectral analyzer works. The spectral analyzer is a device to do spectral analysis on light, measuring the wavelengths of the light reflected by an object. So she believes the following proposition P: if I use a spectral analyzer to analyze the light reflected from any object I see as blue, then I would observe that the light has a high frequency. One day, she plans to do a spectral analysis on an object she sees as blue. Whether the experiment is done on Earth or Inverted Earth, Rose has the same other physical knowledge and background mental states, including her belief that P. Then she tokens a ph-blue experience, but she would token different mental states consequentially. For example, if she is on Earth, Rose would become more confident about the proposition P or her physical knowledge while if she is on Inverted Earth, Rose would either doubt the reliability of her color experience or wonder why her experiment produces wrong results. Hence there is a difference in the psychological role of her ph-blue experience on the two planets.

Some people might say that my argument relies on the assumptions that Rose is a scientist, and she has the knowledge of optics. If Rose has no knowledge of optics, or if she lives at a time when the spectral analyzer has not been invented, or if she lives at a time when the wave theory of light is not proposed, then it seems plausible, as the objector says, the ph-blue experience of Rose has the same psychological role on both planets. However, this reply is not successful. For as Jerry Fodor, a predominant informational semanticist, argues, the

notion of psychological role is evaluated *counterfactually*, i.e., to test whether or not a mental state has a psychological role, we should consider whether a counterfactual is true of that mental state which corresponds to the psychological role. E.g., John has a belief B with the content that Lucky is a dog only if the following counterfactual is true: if it *were* the case that John wants to have a dog as his pet, he has the belief B, and he has no counter-reason to do otherwise, then he *would* have Lucky as his pet. More importantly, in Fodor's view, the antecedent of the counterfactual does not have to be realizable given the agent's current environment. For example, suppose that H_2O and XYZ can be distinguished by some chemical experiment, a litmus-like test, say. Call it "the wetmus test". If the wetmus paper is dipped in H_2O, its color becomes darker, but if it is dipped in XYZ, its color becomes lighter. Then the following counterfactual is true of Oscar's water concept, but not of Twin Oscar's XYZ concept: if *x were* to dip a piece of wetmus paper in water, *x* would see that its color becomes darker (where *x* should be inserted with Oscar or Twin Oscar). According to Fodor, the fact that this counterfactual is true of Oscar's water concept but not of Twin Oscar's XYZ concept shows that their concepts have different psychological roles even if they do not have the technology to synthesize wetmus available in their lifetime. (This is actually Fodor's informational solution (1987, 1990) to Putnam's twin earth argument.)

Accordingly, even if Rose is not a scientist, or the wave theory of light has not been proposed or the spectral analyzer has not been invented, the following counterfactual is true of Rose's ph-blue experience on Earth, but not true of her ph-blue experience on Inverted Earth: if Rose *were to* use the spectral analyzer to analyze the light reflected from any object she sees as blue, then she *would* observe that the light has a high frequency. In addition, since this counterfactual has different truth values with regard to Rose on the planets whether or not the antecedent is realizable, the ph-blue experience of Rose on the planets has different psychological roles *whatever other assumptions are made about Rose*. The same conclusion still holds whether Rose is a soldier or a professor, whether she lives in 21st century or in

medieval times. It seems hard to conceive a thought experiment where the agent's spectrum is inverted, and her ph-blue experience has the same psychological role.

Hence, according to Fodor's argument in defense of informational semantics, if there is a change of content of a mental state, it is unlikely that there is no change in its psychological role. Referring back to my claim in the beginning of this section, if the inverter is installed outside, then supposedly there is no change in the psychological role of Rose's ph-blue experience. But then according to Fodor's view, it follows that the contents of Rose's ph-blue experience tokens on both planets are the same.

§6 Supervenience

According to Kim (1987), supervenience is divided into two kinds: the weak one and the strong one. Here are the formulations:

> <u>*Weak supervenience*</u>: *A*-properties weakly supervene on *B*-properties iff for any possible world *w* and any individuals *x* and *y* in *w*, if *x* and *y* are *A*-discernible in *w*, then they *B*-discernible in *w*.
>
> <u>*Strong supervenience*</u>: *A*-properties strongly supervene on *B*-properties iff for any possible worlds w_1 and w_2, and any individuals *x* in w_1 and *y* in w_2, if *x* in w_1 is *A*-discernible from *y* in w_2, then *x* in w_1 is *B*-discernible from *y* in w_2.

Saying that one property weakly supervenes on another property does not entail that it strongly supervenes on the second property. E.g., suppose for any individual in a world, if its *mass* is the same on two occasions in that world, then its *weight* is the same on those occasions in that world. However, there may be another world in which that individual still has the same mass but does not instantiate the same weight because the gravitational law is different. In this case, we have an instance of weak supervenience without strong supervenience. It should be noted that I have been presupposing weak supervenience throughout my dissertation.

Some may argue that psychointentionalism is false if it is formulated in terms of strong supervenience, especially if it is incorporated with informational semantics.

Furthermore, since, for the reason of naturalization, strong representationalists[8] identify experiential intentionality with phenomenology, the relevant notion of supervenience for SR has to be strong. This is because of the necessity of identity. If phenomenology is identical to intentionality, it cannot be the case that an individual instantiates the same phenomenal property at different worlds, but it only instantiates different intentional properties at worlds respectively. Similarly, as I suggested in Chapter 4, §5, psychointentionalism can be understood as saying that experiential intentional properties are identical to the conjunctive properties of phenomenal character and psychological role. Given this new identity thesis, psychointentionalism in the strong form is logically entailed.

Let's see how a counter-example could be constructed against the conjunction of informational semantics and psychointentionalism in the strong form. Imagine in a possible world, *w*, in which the physical laws are different from the ones in the actual world. An agent, called "Rosa", has the brain which is a molecule-by-molecule duplicate of Rose in the actual world. In this world, blue light would produce ph-yellow experience, and yellow light produces ph-blue ones in Rosa. Furthermore, contrary to the physical laws in the actual world, according to the laws in *w*, light with high frequencies is red, and light with low frequencies is blue. So the reading from the spectral analyzer on yellow light in *w* is the same as that from the analyzer on blue light in the actual world. If informational semantics is assumed, then it seems that the ph-blue experiences of Rose and Rosa would have the same phenomenal character, the same psychological role, but different content. Hence it seems that we would have a counter-example to informational semantics and psychointentionalism in the strong form.

This is a legitimate worry. But I don't think it is insurmountable. First, this case is a genuine counter-example only if some direct-referentialist form of semantics is presupposed for color names and color concepts. In that case, color names and color concepts are rigid, i.e.,

[8] Notice that, however, a weak version of SR only holds that intentional property and phenomenal property co-supervene on each other, but are not identical.

"blue light" refers to the same thing in both worlds. However, direct-referentialism is not the only option for strong representationalists. In fact, strong representationalists can hold the descriptive theory of color concepts (McLaughlin 2003) or the functionalist view of colors (Dretske 1995, Lycan 1996, Tye 2000, Byrne and Hilbert 2003, Byrne and Tye 2006). According to these views, the color blue is identified with the property that normally causes ph-blue in normal human brains. If so, then it is not implausible to say that the contents of the ph-blue experiences of Rose and Rosa are still the same, and hence do not constitute a counter-example to psychointentionalism in the strong sense. To wit, their experiences are about the same functional property although it is physically realized in different ways at the worlds respectively.

Secondly, this argument presupposes that there is a *world-independent* notion of color. To illustrate what I mean, consider the concept of a (physical) triangle. According to the physical geometry of the actual world, the sum of the interior angles of a triangle is not equal to 180 degrees since the physical geometry of the actual world is not flat. Imagine someone claims that in all metaphysically possible worlds, the sum of the interior angles of a triangle is not equal to 180 degrees. Another person comes in and says, "you are wrong because there is a metaphysically possible world whose physical geometry is flat, and in that world, the sum of the interior angles of a triangle equals 180 degrees!" In a sense, they are talking past each other. For the former person can say that her notion of a triangle is different from the latter's since whether something is a physical triangle or not depends on the physical geometry of the world. Similarly, it can be argued that whether some light is blue or yellow depends on the physical laws of the world. Indeed, it is curious why the light that produces a ph-blue experience in Rosa is called "yellow" even if Rose and Rosa have intrinsically the same brains, and that light has the same frequency and other physical characters as blue light in the actual world. It would be more plausible to say that yellow light in that world is different from yellow light in the actual world. Hence it is not clear that the ph-blue experiences of Rose and Rosa have different

contents. (In this context, it is instructive to compare Kripke's reply (1980) to the objection that temperature might not be identical to mean kinetic energy since it is perfectly conceivable. As he says, in that case, what is conceived of is not the property we refer to by the term "temperature" in the actual world, and so when someone claims that she could conceive otherwise, she is not really conceiving of temperature after all.)

Finally, recall that psychointentionalism is intended as a replacement of standard strong intentionalism. To be a satisfactory replacement, it suffices for my project that when incorporated with informational semantics, psychointentionalism (i) addressed the primary concern of the standard SR, (i.e., the naturalist concern) and (ii) does not suffer more counter-examples than SR does. Since psychointentionalism is logically entailed by standard strong intentionalism, there are no more counter-examples to it than to standard strong intentionalism, and hence (ii) is satisfied. If the Rose/Rosa case is a counter-example to psychointentionalism in the strong form, then it must also refute the standard SR. Whatever reason is invoked by the standard strong representationalists to neutralize this counter-example is also applicable here.[9] Hence, there shouldn't be more worries about psychointentionalism (with the strong version of supervenience) than there are about standard SR.

§7 The Phenomenological Gap

I argued that a supposed virtue of SR is that unlike traditional physicalism, it respects our phenomenological intuition (Chapter 2, §2.2). Recall that SR identifies phenomenology and intentionality. In my terms, this identity thesis leaves no phenomenological gap, i.e., there is no phenomenal feature that we would attribute to one but would resist attributing to the other.

[9] See Pautz's counter-example (2006a) to SR that involves a transworld comparison of color experiences, and Byrne & Tye's reply (2006). As Byrne and Tye say, "So, as Pautz would describe the case, there is no obvious reason to suppose it is metaphysically possible. Allegedly, Twin Maxwell [an agent with a different spectrum in another world] is a product of natural selection, someone operating under the same laws as Maxwell [a normal human in the actual world] with a similar kind of visual system, whose experiences represent the same range of colors as Maxwell's, and who not only has no abnormalities whatsoever in his visual system but also is subject to significant color illusions. Pautz simply stipulates that all these conditions can be met together. A defender of Tye's theory may reasonably deny it. Each condition is indeed metaphysically possible, but they are not all possible together." (Byrne and Tye 2006, p. 253)

Since in Chapter 4, §5, I proposed a *conjunctive* identity thesis—the thesis that the conjunctive property of phenomenology and psychological role is identical to intentionality—people might wonder if this conjunctive identity thesis brings back the problem of the phenomenological gap.

To address this worry, consider traditional physicalism that phenomenal states are identical to physical states. To make it visually salient (but somewhat misleading), let me express it with the following equation (P): The physical=the phenomenal. Similarly, I express the conjunctive identity thesis with the following equation (CI): Intentionality=(Phenomenology & Psychological Role). As I have said, there is a phenomenological gap between a phenomenal state and a physical state. I.e., there exists at least a phenomenal feature that we attribute to phenomenal states but resist attributing to physical states, e.g., the feature of *being qualitative*. For equation (P), we would attribute qualitativeness to the right-hand side but resist attributing it to the left-hand side. However, with respect to (CI), it is not clear to me that we would attribute qualitativeness to the right-hand side since it refers to a conjunctive property of phenomenology *and psychological role*. If being qualitative implies non-dispositional, then qualitativeness *cannot* be attributed to the right-hand side of (CI). Hence, it is not clear to me that (CI) gives rise to phenomenological gaps since we would not attribute phenomenal features to the conjunctive properties on the right-hand side either.

Perhaps, my opponent might say that we attribute the feature of *being partly qualitative* to the right-hand side of (CI) but resist attributing it to the left-hand side. In that case, I have to admit that I don't understand what it means to say a conjunctive property is *partly* qualitative. To illustrate my worry, consider the conjunctive property of being water and of being on Mars. Notice that the property of being water stands in the identity relation to the property of being H_2O, i.e., being water is identical to being H_2O. However, it does not make sense to say that the conjunctive property of being water and being on Mars is *partly* identical to the property of being H_2O. Does that imply that the conjunctive property is partly not

identical to the property of being H_2O? Does it make sense to say that? Similarly, even if phenomenal properties are qualitative, it does not make sense to me to say that the conjunctive property of phenomenology and psychological role is partly qualitative. Hence, *prima facie*, I see no reason that the conjunctive identity thesis would be baffled by the phenomenological gap challenge.

§8 Have I Reinvented the Wheel?

There are theories of experience in the literature that also invoke psychological factors to explain phenomenology. Without a standard name for these accounts, I shall call them "the psychological view". According to this view,

> The psychological view: Phenomenal property *F* = the property of being *R*-related to *p* (Speaks 2015, p. 176; Rey 1992, Chalmers 2004).

Here "F", "R", and "p" are all free variables to be supplemented with specific values. "F" ranges over phenomenal characters, e.g., ph-red. "R" ranges over representational relations, e.g., seeing, hearing. Finally, "p" ranges over contents. An appropriate instantiation of this formula is that:

> The phenomenal property *ph-red* = the property of standing in the *seeing* relation to *red* (Speaks 2015, p. 176).

Hence, according to the psychological view, psychological relations enter into the equation while in strong intentionalism, as formulated in my dissertation, they don't.

Some people might be worried that I reinvent the wheel when I propose psychointentionalism to replace converse intentionalism. The reason is that the psychological relations entering into the equation in the psychological view are typically type-identified functionally. If both the psychological view and psychointentionalism invoke psychological roles, then it appears that I have just reinvented the wheel.[10]

[10] I am grateful for Prof. Zoe Drayson for raising this question to me.

However, I don't think we should be misled by this *façade* of similarity. The difference between the psychological view and psychointentionalism becomes transparent if we see their verdicts on the inverted earth case—one of the central problems in my dissertation—from both theories. I have argued with quite a lot of effort that psychointentionalism is compatible with informational semantics even in the case of color inversion. In contrast, Chalmers and Speaks reject informational semantics. This is because if we keep the R-relation fixed, then phenomenal character and content covary according to the above equation. Since in the inverted earth case, we keep the R-relation to be visual experience, and since it is argued that Rose has the same phenomenal character, i.e., ph-blue, throughout, the relevant content cannot change. If informational semantics implies that content changes in this case, then it is clear why Chalmers and Speaks reject informational semantics. On the other hand, Rey's verdict is that although Rose's experiences of seeing the sky on both planets are introspectively indistinguishable to her, she is in phenomenally *different* states (which does not mean that Rey is a phenomenal externalist).[11] Contrary to Rey's verdict, psychointentionalism says that the two experiences have the same phenomenal character, but different psychological roles (without making any commitment on content).

More fundamentally, the deep schism between the psychological view and psychointentionalism lies in where the factor of psychological roles come into the equation. For the psychological theorists, it is on the same side of the equivalence as content whereas in psychointentionalism, it is on the same side as phenomenology. A full comparison between the two views goes beyond the scope of my dissertation. Nevertheless, let me briefly explain my rationale. As I have said in Chapter 4, note 12, phenomenology is intrinsic while psychological role and content are relational. If we put the latter two on the same side of the equation, then

[11] Rey endorses the language of thought hypothesis, and applies it to experience. So the psychological relations in the equation are type-identified in a more fine-grained way than Chalmers' and Speaks'. Hence in Rey's view, the two experiences of seeing the sky on both planets have different psychological roles (whether or not they have the same content), and hence have different phenomenal characters.

what we get is a claim that some intrinsic property is identical to a conjunctive property whose conjuncts are all relational. I think this result is counter-intuitive and susceptible to a lot of metaphysical counter-examples. For this reason, I think psychointentionalism is better than the psychological view.

A Concluding Remark: External-World Skepticism

In the preface of my dissertation, I pointed out that converse intentionalism plays a crucial role in the argument for external-world skepticism. Here is an instance of that species of arguments: Suppose I see a blue object in front of me. Call this experience "E". Then I form a perceptual belief B that there is a blue object in front of me with E as B's sole justification. Now, for all I know, I could be a disembodied Cartesian ego which tokens an experience phenomenally the same as E. In that case, my experience is not veridical. If so, then my belief B is unjustified, which implies that I do not know that there is a blue object in front of me. Let's see the reconstruction of this argument:

1. Necessarily, if I am a disembodied Cartesian ego living in a world where no color is instantiated, then my experience is not veridical.
2. Possibly, I am a disembodied Cartesian ego living in a world where no color is instantiated.
3. Possibly, my experience is not veridical. (1, 2)
4. If possibly my experience is not veridical, then B is unjustified.
5. B is unjustified. (3, 4)
6. Therefore, I don't know that there is a blue object in front of me.

A few remarks: first, every modal term in this argument quantifies over epistemic possible worlds. Second, premise 4 relies on the classical idea that justification provides certainty. I.e., if E justifies B, then given that the agent tokens E, she is certain of B. To elaborate on this, the following view of justification is presupposed in the argument:

> *The certainty principle*: if E justifies B, then (i) E is veridical/true in all epistemically possible worlds, and (ii) in all epistemically possible worlds, if E is veridical/true, then B is true. (Lewis 1986)

Third, various strategies have been provided against this argument. For example, Dretske (1970) argues that epistemic possibilities have to be relevant so that the Cartesian possibility poses no threat. Lewis argues that the range of epistemic possibilities can vary across contexts so that the modal terms in premises 3 and 4 may quantify over different ranges of epistemic possibilities, and hence the argument may not be valid. Finally, reliabilists, e.g., Goldman (1979), choose to reject the certainty principle altogether, arguing that for E to justify B, E does not need to guarantee B's truth. In the following discussion, I will set aside these strategies. My goal is to see how converse (psycho)intentionalism can shed new light on external-world skepticism.

Suppose premises 2 and 4 are true. What is our reason to accept premise 1? In my view, the intuitiveness of premise 1 derives from converse intentionalism. It is suggestive to see how this argument interacts with converse intentionalism in the contemporary context. In a contemporary version of this argument, the notion of a disembodied Cartesian ego is replaced by that of a BIV (Putnam 1981). For instance, in the BIV version, Putnam holds the causal theory of mental content and argues that if I am a BIV, then my experience or belief would mean something else, and thus my experience may still be veridical (1981). However, Putnam's reply seems inadequate if we go back to the Cartesian version. Since as Kriegel (2011) pointed out, converse intentionalism is intuitively compelling. And if I am a disembodied ego which tokens an experience phenomenally the same E, then by converse intentionalism, my experience has the same content as E's which implies that my experience is not veridical. That is, if converse intentionalism is true, then premise 1 seems justified.

After Chapters 2 and 3, it should be clear by now that what I'd like to reject is converse intentionalism. Whichever semantic theory of experience you like, you should not hold converse intentionalism. As is shown in Chapter 2, if you are an informational semanticist, then converse intentionalism brings back the disjunction problem which any informational semanticist is committed to solving. On the other hand, as we have witnessed in Chapter 3, if

you are a PIT theorist, then converse intentionalism is also problematic. For converse intentionalism in conjunction with psychological supervenience (a commitment of PIT) entails the principle that phenomenal sameness entails sameness of psychological role. However, as the Jones/Clone case in Chapter 3 shows, it is possible that two phenomenally indistinguishable experience tokens have different psychological roles, which implies a dilemma to PIT theorists. However intuitive it is, converse intentionalism is not tenable.

Furthermore, as I said at the end of Chapter 4, converse intentionalism is not as intuitive as it appears once the Jones/Clone case is taken into consideration. If at a moment *t*, Jones and Clone both token ph-red experiences, and at the next moment *t*+1, their ph-red tokens cause different mental states, then it is doubtful to say that their ph-red tokens have the same content. Similarly, suppose you know a person who is capable of tokening ph-red experiences, and you also know that she consistently attributes the content of her ph-red experiences to grass, not to maple leaves. In this case, it is actually quite intuitive to say that her ph-red experience and yours have different contents. In sum, both intuitively and theoretically, we have strong reason to abandon converse intentionalism. If so, then we don't have to accept premise 1.

Now since in Chapter 4 I advocate converse psychointentionalism, might we not have a Cartesian revenge by adapting the original argument in terms of converse psychointentionalism? For example, consider this argument:

1*. Necessarily, if I am a BIV and my experience has the same psychological role, then my experience is not veridical.
2*. Possibly, I am a BIV and my experience has the same psychological role.
3*. Possibly, my experience is not veridical. (1*, 2*)
4*. If possibly my experience is not veridical, then B is unjustified.
5*. B is unjustified. (3*, 4*)
6*. Therefore, I don't know that there is a blue object in front of me.

Premise 1* can be justified by converse psychointentionalism. For if it is true, then my experience would have the same content as E's which implies that my experience is not veridical.

However, if we revise the skeptic argument this way, then the weakest premise in my view is premise 2*. Let's think about what reason we have such that we have to accept premise 2*. Notice that premise 2 in the previous argument is different from premise 2* here. The former is an arm-chair speculative hypothesis while the latter is an empirical claim. The former claim can be justified by some form of conceivability arguments but the latter has to be justified by empirical science. Now I want to point out that given our current empirical science, it is already a far-fetched idea that a brain floats in a vat full of nutritious fluid, connected with electrodes, whose states could be deliberately manipulated by some agents, and *still functions well*! It would only be more improbable if we add to this far-fetched idea the requirement that it gives rise to the same psychological structure as a particular normal human's brain, i.e., mine. Furthermore, as far as we know in contemporary cognitive science, the development of the psychological role of a mental state heavily depends on some proper interactions between the agent and the environment. E.g., in the research on Molyneux's problem, scientists found data suggesting that people who previously were blind have difficulty connecting forms of information about the shape of an object through different modalities after they regain their visual capability (Degenaar and Lokhorst 2005/2017). I.e., they cannot form the correct grip aperture to hold an object simply on the basis of their visual representation of that object. This suggests that for an agent's visual experience to have the same psychological role, e.g., to cause the correct grip aperture, she needs some interaction to her environment. It is extremely unlikely that a BIV without any sensory organ can develop such a psychological connection on its own.

Some may ask: what about a swamp-BIV which is spontaneously created as a result of random events (Davidson 1987), and whose experience has the same psychological role as

mine? Surely, the existence of the swamp-BIV is compatible with our nomological laws. Isn't it sufficient to show that the swamp-BIV is possible? I fully concede that there is a *model* of our nomological laws, which might be reasonably construed as a possible world in which such a swamp-BIV exists.[12] But this only shows that the existence of the swamp-BIV does not violate our nomological laws, which is far from saying that our nomological laws *entail* that the existence of the swamp-BIV is possible. To put it another way, as I have emphasized, premise 2* is an empirical claim. What we need to justify it is a piece of *empirical evidence* that at least positively indicates the truth of premise 2* given our empirical science. Merely showing that the existence of the swamp-BIV does not violate the laws constitutes no more positive evidence for the truth of the possibility claim given our empirical science than the fact that the existence of phlogiston/ether/ghosts does not violate the nomological laws constitutes positive evidence for the claim that phlogiston/ether/ghosts are possible given our empirical science. Given that we still lack such evidence for premise 2*, we do not have to accept it.

In sum, the original argument for external-world skepticism presupposes converse intentionalism. It is my goal in my dissertation to show that converse intentionalism is objectionable. If my arguments are correct, then we should abandon it. If so, then the original argument is undermined. On the other hand, if the argument for skepticism is revised as presupposing converse psychointentionalism, then although there is no knock-down argument showing that it is unsound, its premise that *possibly I am a BIV and my experience has the same*

[12] One might justify premise 2* with the following argument:

1#. If it is impossible that I am a swamp-BIV and all our nomological laws are true, then there is no possible world in which all the nomological laws are true and I am a swamp-BIV.

2#. There is a possible world in which all the nomological laws are true and I am a swamp-BIV.

3#. Therefore, it is possible that I am a swamp-BIV and all our nomological laws are true. (1, 2, *modus tollens*)

This argument by itself is not sufficient to establish premise 2*. For we can reasonably ask the advocate of this argument what is her independent ground to claim premise 2#. Without such a ground, her "argument" only begs the question. If her ground is simply to list out all our vocabulary used in the languages of the nomological laws and of describing a swam-BIV, and then to prove that there is a maximally consistent set of sentences of those languages in which the nomological laws and "I-Sen is a swamp-BIV" are true, this set of sentences however is only a *formal model* for the claim *that I-Sen is a swamp-BIV and all the nomological laws are true*. Whether or not this model should be construed as corresponding to a genuine possible world requires further empirical justification.

psychological role not only lacks empirical support, but also is empirically implausible, and hence the strength of the revised argument is significantly weakened.

Bibliography

Adams, F. & Dietrich, L. (2004). Swampman's revenge: Squabbles among the representationalists. *Philosophical Psychology* 17 (3), 323-340. https://reurl.cc/7oAEDy.

Bartlett, G. (2008). On the correct treatment of inverted earth. *Pacific Philosophical Quarterly* 89 (3), 294-311. https://doi.org/10.1111/j.1468-0114.2008.00322.x.

Bermúdez, J. (2014). *Cognitive science: An introduction to the science of the mind*, 2nd ed. Cambridge: Cambridge University Press.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9, 261-325.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy* 10 (1), 615-678.http://dx.doi.org/10.1111/j.1475-4975.1987.tb00558.x.

Block, N. (1990). Inverted earth. *Philosophical Perspectives* 4, 53-79. https://doi.org/10.2307/2214187.

Block, N. (1998). Conceptual role semantics. In E. Craig (Ed.), *Routledge Encyclopedia of Philosophy* (pp. 242-256). London, UK: Routledge.

Block, N. (2003). Mental paint. In M. Hahn, & B. Ramberg (Ed.), *Reflections and replies: Essays on the philosophy of Tyler Burge* (pp. 165-200). Cambridge, Massachusetts: MIT Press.

Block, N. (2007). Wittgenstein and qualia. *Philosophical Perspectives,* 21 (1), 73-115. https://doi.org/10.1111/j.1520-8583.2007.00121.x.

Bourget, D. (2010). Consciousness is underived intentionality. *Noûs* 44 (1), 32-58.

Bourget, D., & Mendelovici, A. (2016/2019). Phenomenal intentionality. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Retrieved from https://reurl.cc/4mDl1v.

Brogaard, B. (2014a). The phenomenal use of "look" and perceptual representation. *Philosophy Compass* 9 (7), 455-468. https://doi.org/10.1111/phc3.12136.

Brogaard, B. (2014b). Introduction. In B. Brogaard (Ed.), *Does perception have content?* (pp. 1-53). Oxford, UK: Oxford University Press.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy* 4 (1), 73-122. http://dx.doi.org/10.1111/j.1475-4975.1979.tb00374.x

Byrne, A., & Hilber, D. (2003). Color realism and color science. *Behavioral and Brain Sciences* 26 (1), 3-21. https://10.1017/s0140525x03000013.

Byrne, A., & Tye, M. (2006). Qualia ain't in the head. *Noûs* 40 (2), 241-255. https://doi.org/10.1111/j.0029-4624.2006.00608.x.

Byrne, A. (2009). Experience and content. *Philosophical Quarterly* 59 (236), 429-451. https://doi.org/10.1111/j.1467-9213.2009.614.x.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2 (3), 200-219.

Chalmers, D. (1997). The conscious mind: In search of a fundamental theory, revised edition.

Oxford, UK: Oxford University Press.
Chalmers, D. (2004). The representational character of experience. In B. Leiter (Ed.), *The future for philosophy* (pp. 153-181). Oxford, UK: Oxford University Press.
Chalmers, D. (2006). Perception and the fall from Eden. In T. Gendler, & J. Hawthorne (Ed.), *Perceptual experience* (pp. 49-125). Oxford, UK: Oxford University.
Chalmers, D. (2017). Naturalistic dualism. In S. Schneider, & M. Velmans (Ed.), *The blackwell companion to consciousness*, 2nd edition (pp. 363-373). Oxford, UK: Wiley-Blackwell.
Chalmers, D. (2018). Book Review of Yli-Vakkuri and Hawthorne's *Narrow Content*, In *Notre Dame Philosophical Reviews*. Retrieved from https://reurl.cc/Q3l2Lq.
Davidson, D. (1986). A coherence theory of truth and knowledge. In E. Lepore, & B. McLaughlin (Ed.), *Truth and interpretation: Perspectives on the philosophy of Donald Davidson* (pp. 307-319). Oxford, UK: Blackwell.
Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association* 60 (3), 441-458. https://reurl.cc/9XAWOj.
Degenaar, M. & Lokhorst, G.-J. (2005/2017). Molyneux's problem. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Retrieved from https://reurl.cc/zzGQZ0.
Dretske, F. (1970). Epistemic operators. *Journal of Philosophy* 67 (24), 1007-1023. https://doi.org/10.2307/2024710.
Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, Massachusetts: MIT Press.
Dretske, F. (1986). Misrepresentation. In R. Bogdon (Ed.), *Belief* (pp. 17-36). Oxford, UK: Oxford University Press.
Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, Massachusetts: MIT Press.
Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, Massachusetts: MIT Press.
Egan, F. (1995). Computation and content. *Philosophical Review* 104 (2), 181-203. https://doi.org/10.2307/2185977
Farkas, K. (2013). Constructing a world for the senses. In U. Kriegel (Ed.), *Phenomenal intentionality* (pp. 99-115). Oxford, UK: Oxford University Press.
Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, Massachusetts: MIT Press.
Fodor, J. (1990). *A theory of content and other essays.* Cambridge, Massachusetts: MIT Press.
Fodor, J. (1994). *The elm and the expert: Mentalese and its semantics*. Cambridge, Massachusetts: MIT Press.
Fodor, J. (2010). *LOT2: The language of thought revisited*. Oxford, UK: Oxford University Press.
Garson, J. (2001). Natural semantics: Why natural deduction is intuitionistic. *Theoria* 67 (2), 114-139. https://doi.org/10.1111/j.1755-2567.2001.tb00200.x

Goldman, A. (1979). What is justified belief? In G. Pappas (Ed.), *Justification and knowledge* (pp. 1-25). Boston: D. Reidel.

Glüer, K. (2009). In defence of a doxastic account of experience. *Mind & Language* 24 (3), 297-327. https://doi.org/10.1111/j.1468-0017.2009.01364.x.

Glüer, K. (2012). Colors and the content of color experience. *Croatian Journal of Philosophy* 12 (36), 421-437.

Glüer, K. (2014). Looks, reasons, and experiences. In B. Brogaard (Ed.), *Does perception have content?* (pp. 1-53). Oxford, UK: Oxford University Press.

Glüer, K. (2016). Intentionalism, defeasibility, and justification. *Philosophical Studies* 173 (4), 1007-1030. https://doi.org/10.1007/s11098-015-0538-6.

Glüer, K. (2018). Defeating look. *Synthese* 195 (7), 2985-3012. https://reurl.cc/d5bpWD.

Hardin, C. (1988). *Color for philosophers*. Indianapolis: Hackett.

Harman, G. (1987). (Non-solipsistic) conceptual role semantics. In E. LePore (Ed.), *New directions in semantics* (pp. 55-81). London, UK: Academic Press.

Harman, G. (1990). The intrinsic quality of experience. *Philosophical Perspectives* 4, 31-52. https://doi.org/10.2307/2214186.

Hendry, R. (2010). Ontological reduction and molecular structure. *Studies in History and Philosophy of Modern Physics* 41 (2), 183-191. https://reurl.cc/8yY9kX.

Horgan, T. (1994). Naturalism and intentionality. *Philosophical Studies* 76 (2-3), 301-326. https://doi.org/10.1007/BF00989833.

Horgan, T., & Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In D. Chalmers (Ed.), *Philosophy of mind* (pp. 520-533). Oxford, UK: Oxford University Press.

Horgan, T., Tienson, J., & Graham, G. (2004). Phenomenal intentionality and the brain in a vat. In R. Schantz (Ed.), *The externalist challenge* (pp. 297-318). Berlin: Walter De Gruyter.

Horgan, T., & Graham, G. (2012). Phenomenal intentionality and content determinacy. In R. Schantz (Ed.), *Prospects for meaning* (pp. 321-344). Berlin: Walter De Gruyter.

Horgan, T. (2014). Phenomenal intentionality and secondary qualities: The quixotic case of color. In B. Brogaard (Ed.), *Does perception have content?* (pp. 329-350). Oxford, UK: Oxford University Press.

Jackson, F. (1977). *Perception: A representative theory.* Cambridge, UK: Cambridge University Press.

Jackson, F. (1982). Epiphenomenal qualia. In J. Heil (Ed.), *Philosophy of mind: A guide and anthology* (pp. 762-771, 2004). Oxford, UK: Oxford University Press.

Kim, J. (1987). "Strong" and "global" supervenience revisited. *Philosophy and Phenomenological Research* 48, 315-326. https://doi.org/10.2307/2107631.

Kim, J. (2010). *Philosophy of mind,* 3rd ed. London, UK: Routledge.

Kind, A. (2008). Qualia. In J. Fieser & B. Dowden (Ed.), *Internet Encyclopedia of Philosophy*.

Retrieved from https://iep.utm.edu/qualia/.

Kriegel, U. (2003). Is intentionality dependent on consciousness? *Philosophical Studies* 116 (3), 271-307. https://doi.org/10.1023/B:PHIL.0000007204.53683.d7.

Kriegel, U. (2007). Phenomenal inexistence and phenomenal intentionality. *Philosophical Perspectives* 21 (1), 307-340. https://doi.org/10.1111/j.1520-8583.2007.00129.x.

Kriegel, U. (2011). *The sources of intentionality*. Oxford, UK: Oxford University Press.

Kriegel, U. (Ed.). (2013a). *Phenomenal intentionality*. Oxford, UK: Oxford University Press.

Kriegel, U. (2013b). The phenomenal intentionality research program. In U. Kriegel (Ed.), *Phenomenal intentionality* (pp. 1-26). Oxford, UK: Oxford University Press.

Kriegel, U. (2017). Brentano's concept of mind: Underlying nature, reference-fixing, and the mark of the mental. In S. Lapointe, & C. Pincock (Ed.), *Innovations in the History of Analytical Philosophy* (pp. 197-228). New York: Palgrave Macmillan.

Kripke, S. (1980). *Naming and necessity*. Cambridge Massachusetts: Harvard University Press.

Lewis, D. (1979). Attitudes de dicto and de se. *Philosophical Review* 88 (4), 513-543. https://10.2307/2184843.

Lewis, D. (1986). Elusive knowledge. *Australasian Journal of Philosophy* 74 (4), 549-567. https://doi.org/10.1080/00048409612347521.

Levine, J. (1983). Materialism and qualia: The explanatory gap. In J. Heil (Ed.), *Philosophy of mind: A guide and anthology* (pp. 772-780, 2004). Oxford, UK: Oxford University Press.

Loar, B. (1988a). Social content and psychological content. In R. Grimm, & D. Merrill (Ed.), *Contents of thought: Proceedings of the 1985 Oberlin Colloquium in Philosophy* (pp. 99-110). Tucson: University of Arizona Press.

Loar, B. (1988b). Reply: A new kind of content. In R. Grimm, & D. Merrill (Ed.), *Contents of thought: Proceedings of the 1985 Oberlin Colloquium in Philosophy* (pp. 121-139). Tucson: University of Arizona Press.

Loar, B. (1997). Phenomenal states, 2nd ed. In O. Flanagan, N. Block & G. Güzeldere (Ed.), *The nature of consciousness: Philosophical debates* (pp. 597-616). Cambridge, Massachusetts: MIT Press.

Loar, B. (2003). Phenomenal intentionality as the basis of mental content. In M. Hahn & B. Ramberg (Ed.), *Reflections and replies: Essays on the philosophy of Tyler Burge* (pp. 229-258). Cambridge, Massachusetts: MIT Press.

Lycan, W. (1996). *Consciousness and experience*, Cambridge, Massachusetts: MIT Press.

Lycan, W. (2000/2019). Representational theories of consciousness. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.* Retrieved from https://reurl.cc/3LAkY9.

McLaughlin, B. (2003). Colour, consciousness, and colour consciousness. In S. Quentin & J. Aleksandar (Ed.), *Consciousness: New perspectives* (pp. 97-154). New York: Oxford University Press.

Melynek, A. (2002). Papineau on the intuition of distinctness. *SWIF Philosophy of Mind* 4.

Retrieved from https://reurl.cc/4mDl4j.

Mendelovici, A. (2010). *Mental Representation and Closely Conflated Topics*, Dissertation. Princeton University.

Mendelovici, A. (2013). Reliable misrepresentation and tracking theories of mental representation. *Philosophical Studies* 165 (2), 421-443. https://reurl.cc/Mdm2NK.

Mendelovici, A. & Bourget, D. (2014). Naturalizing intentionality: Tracking theories versus phenomenal intentionality theories. *Philosophy Compass* 9 (5), 325-347. https://doi.org/10.1111/phc3.12123.

Mendelovici, A. (2016). Why tracking theories should allow for clean cases of reliable misrepresentation. *Disputatio* 8 (42), 57-92. https://doi.org/10.2478/disp-2016-0003.

Mendelovici, A. (2018). *The phenomenal basis of intentionality*. Oxford, UK: Oxford University Press.

Mendelovici, A. & Bourget, D. (2020). Consciousness and intentionality. In U. Kriegel (Ed.), *Oxford handbook of the philosophy of consciousness* (pp. 560-585). Oxford, UK: Oxford University Press.

Millikan, R. (1986). *Language, thought, and other biological categories: New foundations for realism*. Cambridge, Massachusetts: MIT Press.

Millikan, R. (1989). Biosemantics. *Journal of Philosophy* 86 (6), 281-297. https://jphil198986652.

Millikan, R. (2009). Biosemantics. In A. Beckermann, B. McLaughlin, & B. Walter (Ed), *The oxford handbook of philosopher of mind* (pp. 394-406). Oxford: Oxford University Press.

Molyneux, B. (2009). Why experience told me nothing about transparency. *Noûs* 43 (1), 116-136. https://10.1111/j.1468-0068.2008.01698.x.

Molyneux, B. (2011). On the infinitely hard problem of consciousness. *Australasian Journal of Philosophy* 82 (2), 211-228. https://reurl.cc/VXZ2jQ.

Montague, M. (2010). Recent work on intentionality. *Analysis, 70*(4), 765-782. https://doi.org/10.1093/analys/anq090.

Montague, M. (2016). *The given: Experience and its content*. Oxford, UK: Oxford University Press.

Nagel, T. (1974). What is it like to be a bat? In J. Heil (Ed.), *Philosophy of mind: A guide and anthology* (pp. 528-538, 2004). Oxford, UK: Oxford University Press.

Neander, K. (2004/2012). Teleological theories of mental content. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.* Retrieved from https://reurl.cc/9XAW1O.

Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. Cambridge, Massachusetts: MIT Press.

Papineau, D. (2002). *Thinking about Consciousness.* Oxford: Oxford University Press.

Pautz, A. (2006a). Sensory awareness is not a wide physical relation: An empirical argument against external intentionalism. *Noûs* 40 (2), 205-240. https://doi.org/10.1111/j.0029-4624.2006.00607.x.

Pautz, A. (2006b). Can the physicalist explain colour structure in terms of colour experience? *Australasian Journal of Philosophy* 84 (4), 535-564. https://reurl.cc/6l4GLM.

Pautz, A. (2008). The interdependence of phenomenology and intentionality. *The Monist* 91 (2), 250-272. https://doi.org/10.5840/monist20089124.

Pautz, A. (2009). A simple view of consciousness. In R. Koons & G. Bealer (Ed.), *The waning of materialism* (pp. 25-66). Oxford, UK: Oxford University Press.

Pautz, A. (2013). Is phenomenology the ground of intentionality? In U. Kriegel (Ed.), *Phenomenal intentionality* (pp. 25-66). Oxford, UK: Oxford University Press.

Potrč, M. (2013). Phenomenology of Intentionality. In D. Fisette & G. Fréchette (Ed.), *Themes from Brentano* (pp. 165-188). Amsterdam, Netherlands: Rodopi.

Putnam, H. (1975). The meaning of "meaning". *Minnesota Studies in the Philosophy of Science* 7, 131-193. http://hdl.handle.net/11299/185225.

Putnam, H. (1981). *Reason, truth, and history*, Cambridge: Cambridge University Press.

Ren, H. (2016). Inverted earth revisited. *Erkenntnis* 81 (5), 1093-1107. https://link.springer.com/article/10.1007/s10670-015-9786-2.

Rey, George. (1992). Sensational switched. *Philosophical Studies* 68 (3), 289-319. https://doi.org/10.1007/BF00694849.

Rupert, R. (2008). Causal theories of mental content. *Philosophy Compass* 3 (2), 353-380. https://doi.org/10.1111/j.1747-9991.2008.00130.x.

Searle, J. (1986). Intentionality and its place in nature. *Synthese* 61 (3), 87-100. https://doi.org/10.1007/BF00485486.

Shoemaker, S. (2000). Phenomenal character revisited. *Philosophy and Phenomenological Research* 60 (2), 465-467. https://doi.org/10.2307/2653497.

Shoemaker, S. (2006). On the Ways Things Appear. In T. Gendler, & J. Hawthorne (Ed.), *Perceptual experience* (pp. 461-480). Oxford, UK: Oxford University.

Siegel, S. (2014). Affordance and the content of perception. In B. Brogaard (Ed.), *Does perception have content?* (pp. 51-75). Oxford, UK: Oxford University Press.

Siewert, C. (1998). *The significance of consciousness*. Princeton, New Jersey: Princeton University Press.

Siewert, C. (2004). Is experience transparent? *Philosophical Studies* 117 (1-2), 15-41. https://link.springer.com/article/10.1023%2FB%3APHIL.0000014523.89489.59.

Speaks, J. (2009). Transparency, intentionalism, the nature of perceptual content. *Philosophy and Phenomenological Research* 79 (3), 539-573. https://doi.org/10.1111/j.1933-1592.2009.00293.x.

Speaks, J. (2010/2019). Theories of meaning. In E. Zalta (Ed.), *Stanford Encyclopedia of*

Philosophy. Retrieved from https://plato.stanford.edu/entries/meaning /

Speaks, J. (2015). *The phenomenal and the representational*. Oxford: Oxford University Press.

Stanley, J. & Szabó, Z. (2000). On quantifier domain restriction. *Mind and Language* 15 (2-3), 219-261. https://doi.org/10.1111/1468-0017.00130.

Thompson, B. (2008). Representationalism and the argument from hallucination. *Pacific Philosophical Quarterly* 89(3), 384-412. https://doi.org/10.1111/j.1468-0114.2008.00327.x

Tye, M. (1998). Inverted Earth, swampman, and representationalism. *Philosophical Perspectives* 12, 459-478. https://doi.org/10.1111/0029-4624.32.s12.20.

Tye, M. (2000). *Consciousness, color, and content*. Cambridge, Massachusetts: MIT Press.

Tye, M. (2009). *Consciousness revisited: Materialism without phenomenal concepts*. Cambridge, Massachusetts: MIT Press.

Yli-Vakkuri, J. & Hawthorne, J. (2018). *Narrow content*. Oxford: Oxford University Press.