UNIVERSITÀ DEGLI STUDI DI MILANO

DOCTORAL PROGRAM: PHILOSOPHY AND THE HUMAN SCIENCES

XXXVII cycle

DEPARTMENT OF PHILOSOPHY Piero Martinetti

PhD DISSERTATION

THE PERSISTENCE OF INDETERMINACY: CONTEMPORARY THEORIES AND KRIPKENSTEIN'S CHALLENGE TO REPRESENTATIONAL CONTENT

M-PHIL/05

SUPERVISOR: Prof. Elisa Paganini

PhD STUDENT: Sara Papic

Matr: R13201

https://orcid.org/0000-0002-2317-4180

PHD COORDINATOR: Prof. Niccolò Guicciardini Corsi Salviati

A.Y. 2024/2025

TABLE OF CONTENTS

ACKNOWLEDGMENTS

This thesis benefitted from the abundant and much needed help of colleagues, family, friends, and institutions – too many to list comprehensively.

I am first and foremost indebted to my supervisor, Elisa Paganini, who consistently supported my work throughout my PhD studies with her incisive feedback, which she never failed to deliver delicately. Elisa has always been generous with her time, expertise, and advice, some of which I will cherish forever ("a philosophy paper should not look like a mystery novel" comes to mind – her attempt at curbing my tendency towards suspense). I can only hope that some of her rigorous approach to philosophy has rubbed off on me.

I am also thankful to Anandi Hattiangadi, who was my mentor during the year I spent at Stockholm University with the Erasmus Traineeship programme. The breadth of her mentorship and the kindness she offered me is inspiring; it makes me hope I can one day pay it forward. Our conversations in her cosy SU office always led me to unexpected places and left a lasting mark on both the content and direction of this thesis.

I am grateful to all the audiences at conferences and seminars where I presented bits of what ended up constituting this thesis. These include, in chronological order: the audience at the 2022 Salzburg Conference for Young Analytic Philosophy; the audience at the 2022 San Raffaele Summer School in Philosophy; the participants in the Sign-Language-Reality Seminar in 2023; colleagues who frequented the Milan WIP seminars in 2023 and beyond; colleagues who attended the WIP seminar at Stockholm University in 2023; the audience at the 15th SIFA Conference; and the audience at the 7th PLM Conference.

A heartfelt thanks goes to all the friends I've made at the Milan philosophy department, who never failed to make the academic burdens we carried feel lighter. They are responsible for the lively and supportive environment we collectively relied on over the years. A particular thanks goes to my cycle cohort: Rosa Cinelli, Marina D'Amico, Rodrigo De Araujo, Yazan Freij, Matteo Gandolini, Alessandro Guglielmo, Salma Khan, Lucia Puppo, and Luigi Valletta – your friendship has been and continues to be an immense source of joy. I extend the same sentiment to the good friends I made during my time visiting at the philosophy department in Stockholm.

Many colleagues and friends in the philosophical world whom I've individually bothered with questions or with the contents of this thesis have been gracious enough to listen and respond in ways that led to it being much improved. These include Andrea Bianchi, Paul Boghossian, Max Deutsch, Paolo di Lucia, Andrea Guardo, and Alex Moran. I fear I have inevitably forgotten others who made valuable contributions here, especially if the conversations happened in person or over drinks.

Some sections of this thesis borrow heavily from previously published papers that have gone through the process of peer-review. I'd like to thank the anonymous reviewers for *Phenomenology and Mind* and for *Erkenntnis*, who helped me improve these sections with their precious comments.

I'm grateful for the support provided by the PhD scholarship awarded to me by the University of Milan, PRIN funds[1], and the additional funding granted by Erasmus+ during my

[1] Project PRIN 2022 PNRR (Project Code P20225A73K_003) – project title: Conceptual Negotiation for a Better Future: An Ethical and Conceptual Investigation.

second year.

Finally, I'd like to thank my husband, Andreja Tonev, whose unconditional and patient support shaped the last four years of my life into a joyful, productive time. His unwavering belief in my abilities has almost fully suppressed my tendency to doubt myself, and some of that attitude shines through in this work. Ultimately, my love for Andreja and our baby Nikola fuels everything I do.

INTRODUCTION

This thesis is concerned with the foundations of representational content. There is a familiar feature of words, sentences, beliefs, and perceptions:[2] they are about, or mean, or *represent* something. The sentence "cats are cute" is about cats being cute, and my belief that cats are cute is also about cats being cute. Most things in the world are not about something else – they just are. In virtue of what, then, do words, sentences, beliefs, and perceptions represent the things they represent? In other words, in virtue of what do they have the contents that they do?

In *Wittgenstein on Rules and Private Language* (1982/1995), Saul Kripke argued that there is no fact of the matter about what we mean by our words. Though Kripke intended to target theories of meaning in the linguistic sense, his arguments ultimately challenge theories of mental content as well. The effects of Kripke's argument have been devastating: Theories of content have struggled to avoid his sceptical paradox ever since. The aim of this work is to investigate and determine whether several contemporary approaches to representational content meet the sceptical challenge Kripke set up nearly fifty years ago. The result of my investigation is, unfortunately, negative. None of the contemporary approaches examined in this work successfully isolate what makes representations have the contents that they do.

In the second half of the 20th century, several philosophical debates related to indeterminacy became prominent. W. V. O. Quine's ideas on the indeterminacy of translation, first

---

[2] I limited myself to four examples here, but there are many other kinds of things that have representational properties.

covered in *Word and Object* (1960, pp. 26–73), suggested that it may be indeterminate what our terms refer to. Kripke argued that there are no facts, physical or mental, that could constitute anyone's meaning anything (an "incredible conclusion," 1982, p. 22). The "qua problem," first named as such by Kim Sterelny (1983), revealed that causal theories of reference struggle to assign determinate reference to natural kind terms. Similar problems related to indeterminacy manifested themselves in various forms, such as the disjunction problem (Fodor, 1984) and the distality problem (Dretske, 1981/1982, pp. 155–164) for causal and teleosemantic theories of mental content, or Nelson Goodman's new riddle of induction (Goodman, 1954/1983, pp. 59–81).[3] These problems highlight the occurrence of a vicious indeterminacy, and much ink has been spilled trying to curb it.

Despite the superficial similarity between these problems, they emerged within different branches of philosophy, which means that their existence as problems relates to different concerns. Regardless, the similar form in which the problems appear has been noticed many times.[4] I group them together not to suggest that they could all be solved in a unified manner, but to confront the reader with the severity and magnitude of the problem of indeterminacy as it appeared in the

3 Some of these problems (those relevant to the goals of this thesis) will be described and discussed in more detail within this work. The qua problem is described in Chapter 3, while I cover the disjunction and distality problems in Chapter 4.

4 The connections between these problems have been noted and analysed many times in the literature. In his introduction of Kripkenstein's paradox, Kripke links the private language argument with Quine's indeterminacy of translation (1982, p. 14) and Goodman's new riddle of induction (1982, p. 20); Maddy (1984) and Douglas (2018) link Kripkenstein's paradox with the qua problem; Deutsch (2021) even calls the disjunction problem "a variety of the qua problem" (p. 1816); Jared Warren (2024) notes that Lewisian reference magnetism has been used to "answer the challenges of grue and quus, Quine's indeterminacy of translation argument, and Putnam's model-theoretic argument against realism." These examples are by no means exhaustive.

literature in the past six decades. Not only is the problem big, but it has also proved itself to be very resilient; it often reappears despite refinements made in the afflicted theories.

This work will focus on indeterminacy as it affects theories of intentional or representational content. Intentionality has been defined in many ways. As I have alluded to at the beginning of this introduction, I will be using the term to mean aboutness, i.e., the peculiar capacity of some things to be about other things, to represent them. The notion I am interested in includes all kinds of representation, both mental and linguistic.[5] I will use the terms "representational content" and "intentional content" interchangeably throughout this work, and I will assume they are synonymous.[6]

The word "content" as I am using it here is a theoretical term that has been given the job of explaining the way in which beliefs, perceptions, words, and sentences represent (or are about) specific things. The assumption is that my belief that it is sunny today has a content, and that it is in virtue of that content that my belief is about (or represents) a certain state of affairs: it being sunny today. Other people can have beliefs with the same content; and it is possible for one subject to have different attitudes towards the same content. For example, I may *wish* that it is sunny today,

---

[5] Historically, the original meaning of "intentionality" only included the aboutness of mental phenomena and was even postulated by Franz Brentano – who introduced the concept – to be the exclusive and differentiating property of the mental (1874/2009, p. 68). This influential idea is now colloquially known as intentionality being "the mark of the mental." However, some philosophers, mostly due to a difference in methodological approach caused by the rise of behaviourism and the linguistic turn, argued that some properties of language at the very least resemble the aboutness exhibited by mental phenomena (see e.g. Dennett 1969/1996, pp. 19–32). Some have argued that language exhibits intentionality derivatively (e.g. Searle, 1983/2004, pp. vii–viii). To remain neutral in this respect I will use the term as inclusively as possible and will not differentiate between the aboutness exhibited by language and the aboutness exhibited by mental states. The goal of the thesis is to investigate whether content determinacy can be achieved, by whatever means necessary.

[6] Though I am in good and plentiful company, not everyone agrees on the conflation between representation and intentionality. See Laura Gow (2024) for a recent argument in favour of the separation of these two notions.

just as I may imagine it. I can also utter the expression "it is sunny today," and that expression has something in common with my belief that it is sunny today. The thing they have in common is, by assumption, their content. However intuitive these initial remarks may be, there is widespread disagreement over the *nature* of content as I've defined it. In particular, even assuming that there is some real phenomenon corresponding to the theoretical notion of content, it is unclear what (if anything) this notion picks out. Different theories of meaning posit that contents have radically different natures. For example, teleosemantics sees mental contents as instantiated by certain physical representational vehicles; proponents of the phenomenal intentionality programme view contents as identical to the phenomenal properties of a subject's experience. Since I am investigating whether it is possible for *any* theory to successfully account for content, I am not going to assume (almost) anything about its underlying nature.

To clarify what has been said until now, the main goal of this thesis is to assess whether there have been successful advancements in answering the content determination question:

CDQ:[7] What makes representations, linguistic or mental, have the content that they do?

For example, what makes it the case that I believe *that it is Monday*, and not that the Earth is a cube, or anything else? In virtue of which facts does the word "gold" refer to the chemical element with the atomic number 79, and does not refer to tables, or mosquitos? In Chapters 3, 4, and 5, I consider the answers provided by three kinds of theories that belong to very different camps, but that nonetheless have a shared aim—to give a foundational account of representational content,

[7] It cannot be ignored that this question is, as it stands, no less loaded than "have you stopped beating your wife?" It contains both the assumption that there are representations and that they have their contents in virtue of something else. In other words, the question excludes both eliminativist and primitivist approaches to content.

whether linguistic or mental. The approaches I consider are direct causal theories of reference, teleosemantics, and phenomenal intentionality theories. The first two are approaches that have naturalisation as their aim, while the last does not. The last two are theories of mental content, while the first is a theory of linguistic content.

I contrast the content determination question (CDQ) with a more general question about representation, which raises a different set of issues that will not be the focus of this thesis: What makes it the case that something is a representation at all? This work will not include discussion of why there are any representations as opposed to there being none, or what makes representing entities different from non-representing entities. Though these questions are interesting in their own right, and despite the fact that they are often related to those I aim to answer, they are simply beyond the scope of this work.

Theories that aim to answer the CDQ struggle with indeterminacy. In and of itself, this is an interesting fact. In practice, we have no doubts whatsoever that our words, sentences, and thoughts have determinate meanings and that they do so with a certain constancy across time, speakers, and contexts. Given these practical realities about language and our own mental lives, the difficulty in accounting for determinate content is perplexing. Kripke (1982) tackled the issue of content determination in the broadest manner possible and arrived at a sceptical answer: There are no facts in virtue of which anyone means anything by a sign. Due to the generality and influence of Kripke's critique, I will use it as a framing device within this thesis. Precisely because of their generality, Kripke's insights into why the theories he considered fail to give a positive answer to CDQ will be useful in the initial assessment of the success of new theories, and they will provide additional unity to the sometimes unpleasantly broad scope of this work.

A final introductory note concerns physicalism and its relationship with intentionality. One question that is often asked (implicitly or explicitly) along with CDQ is the following: Is the thing

that makes representations have the content that they do a physical or natural thing? Is it reducible to a physical or natural thing? This question captures a common worry that emerges when tackling intentionality: Are meaning and the mental part of the natural world?[8] Many philosophers who attempted to answer CDQ did so with this question firmly in mind. As Fred Dretske put it in *Knowledge and the Flow of Information*,

> The entire project can be viewed as an exercise in naturalism—or, if you prefer, materialistic metaphysics. Can you bake a mental cake using only physical yeast and flour? The argument is that you can […] we have all the ingredients necessary for understanding how purely physical systems could occupy states having a content (meaning) characteristic of knowledge and belief. (1981/1982, p. xi)

As can be gauged from this quote, the expectation was that content *could* be reduced to or fully explained by underlying natural phenomena. The idea that meaning and the mental are somehow separate from nature and cannot be captured by our best physical theories is not only unattractive to many philosophers but unacceptable, a non-possibility. Philosophy in the analytic tradition is characterised by strong naturalistic metaphysical and methodological assumptions. One of these assumptions is that there are no "queer", "spooky" entities – everything must, in one way or another, fall within the domain of physical reality. Because of this, many philosophers start from the assumption that meaning and the mental *must* ultimately be reducible to the physical; it is only a matter of figuring out the theoretical details of how such reduction works (or, to put it more vividly, it is a matter of finding the right recipe).

8 There is much to be said about what physicalism is, how it relates to naturalism, and what it means for everything to be physical within different physicalist programmes. However, I cannot devote too much space to this discussion in this work; what matters here is the common aversion to metaphysical dualism and the consequent attempts to seamlessly insert meaning and mental phenomena into the natural world. The methods employed to achieve this have historically taken different forms.

It is no surprise that so much energy has been spent on attempting to answer the CDQ: Its resolution, if given in physicalist terms, could mean the completion of the naturalistic puzzle. It is also no surprise, then, that the repeated failures in answering the CDQ have cast doubt over the viability of taking a physicalist approach to meaning and the mental. If all efforts to explain the determinacy of content by relying on physical facts fail, one starts suspecting that intentional content is simply not physical. However, this line of thinking is mistaken. As Kripke has argued, difficulties in determining content cannot be overcome by attempting to cite *any kind of fact* as the determiner of meaning, and that includes notoriously non-physicalist candidates such as a hypothetical introspectively accessible quale associated with meaning something (1982/1995, pp. 42–43, 69). The difficulty in answering CDQ, then, is not an indictment of the physicalist programme; the adoption of non-naturalistic metaphysical or methodological assumptions does not resolve the indeterminacy at hand (or at least, not immediately).[9] In other words, the problem of what determines representational content is not a problem exclusive to naturalistic theories: It is a problem for everyone.

9 One non-naturalistic way to immediately resolve CDQ is to adopt a crude form of semantic primitivism in which we take intentional content to be sui-generis and irreducible. Kripke characterises this strategy as being "desperate" and "completely mysterious" (Kripke, 1982/1995, p. 51). I am not sure whether Kripke is right about this, but I will not consider semantic primitivism in detail in this work. Similarly, I am not going to discuss the merits of semantic irrealism either, i.e. the position that fully embraces the conclusion of the sceptical paradox. This is primarily a choice made due to limitations in space (and time).

THESIS SUMMARY

Having introduced the main topic of this work, I will now provide a summary of the contents of every chapter section by section. The summary should serve as a roadmap to the reader in their navigation of the thesis. This work is structurally divided into five chapters, and each chapter contains several sections.

CHAPTER 1

Chapter One overviews and interprets the sceptical paradox introduced by Kripke in *Wittgenstein on Rules and Private Language* (WRPL), which leads to the conclusion that no one ever means anything by a word. I isolate the constraints Kripke's arguments place on theories of content. These constraints will be used in assessing the adequacy of contemporary theories of content in later chapters.

Section 1.1

This section straightforwardly presents the sceptical challenge as it appears in *WRPL*.

Section 1.2

This section straightforwardly presents some of the arguments Kripke provides against (at the time) extant theories of content.

Section 1.3

This section synthesises the constraints Kripke puts on theories of content through his arguments. The constraints can be understood in the form of three requirements: the extensionality, normativity, and non-circularity requirement. I finish the section by considering whether some of the responses to *WRPL* have taken the scope of the sceptical conclusion to be broader than it truly is. I conclude that while the move from scepticism about meaningful instances to scepticism about abstract meanings is warranted, it is not contained in WRPL and is not needed for the purposes of this thesis.

CHAPTER 2

This chapter's aim is to analyse and interpret the normativity requirement posed by Kripke's sceptical paradox. The normativity requirement has been interpreted in contrasting ways in the responses to WRPL. The question that must be answered to properly understand the normativity requirement is the following: Is meaning normative? If not, why do we feel guided in our usage of language? Much of the chapter relies on analogies with discussions in metaethics. I ultimately conclude that meaning is not normative and that the normativity requirement should be interpreted in the "simple" manner proposed by Paul Boghossian (1989). The upshot of this is that Kripke's argument is not a priori, i.e. it does not affect all possible theories. If the argument is not a priori, it does not preclude the possibility that a new theory could resolve the sceptical paradox.

Section 2.1

This section overviews some of the literature devoted to interpreting the normativity requirement. Two main ways of reading the requirement can be individuated: those that posit meaning is normative in the sense that it provides semantic correctness conditions, and those that believe meaning is normative in the sense that it is inherently motivating, i.e. that meaning is categorically normative. The two ways of understanding meaning normativity are exclusive because semantic correctness conditions are not inherently motivating. I take the two interpretations of meaning normativity to lead to two different readings of Kripke's argument, as Anandi Hattiangadi (2007) did: one in which it is a priori and one in which it is a posteriori.

Section 2.2

This section initiates the investigation into whether we have reason to believe meaning is categorically normative. If we did, that would favour an interpretation of Kripke's argument as a priori. If the argument is a priori and sound, the endeavour of this thesis would be misguided from the very beginning: the CDQ could not be answered in principle.

Section 2.3

Section three of this chapter has the goal of assessing whether linguistic correctness (as contrasted with semantic correctness) can be a source of categorical meaning normativity. I introduce the notion and flesh it out by defining it through adherence to use-conditions. I argue that linguistic correctness so defined is constitutive of conventional meaning.

Section 2.4

This section considers whether linguistic correctness, which has been shown to be constitutive of conventional meaning in the previous section, provides speakers with categorical

norms. I analyse David Enoch's (2006) argument against the possibility of deriving categorical *moral* norms from what is constitutive of agency and apply its reasoning to the case of meaning norms. I conclude that we cannot derive categorical meaning norms from what is constitutive of conventional meaning and so do not have reason to believe there are any categorical meaning norms. This leads me to adopt the "simple" reading of the normativity requirement, and consequently, an a posteriori reading of the sceptical argument. Kripke's argument only demonstrates that the CDQ has not been answered, not that it *cannot* be answered.

CHAPTER 3

This chapter initiates the part of the thesis dedicated to assessing recent theories that attempt to answer the CDQ. I focus here on unmediated (direct) causal theories of reference, which promise to deliver the meanings of some terms purely through causal contact between speaker and referent.

Section 3.1

This section provides a brief explanation of causal theories of reference, reference fixing through baptisms, and reference borrowing. Unmediated causal theories of reference suggest that reference could be explained in a fully naturalised manner, at least in some cases. This means that they aim to provide a reductive account of some aspects of meaning and are a promising avenue for answering the CDQ in a way that does not immediately run into the issues identified by Kripke in *WRPL,* unlike e.g. descriptive theories of reference.

Section 3.2

Unmediated causal theories of reference run into the qua problem. The qua problem highlights that by excluding descriptive elements, these theories lead to unacceptable referential indeterminacy. I overview the history and some early attempts at resolving the qua problem, which mostly seem to "pass the referential buck:" they explain the determinacy of reference via the determinacy of certain mental contents.

Section 3.3

This section presents a recent attempt at resolving the qua problem for unmediated causal theories of reference proposed by Max Deutsch (2023). Deutsch argues that causal relevance is sufficient for explaining how reference is determined. He believes a newly introduced kind term's reference is grounded in the property that is causally relevant to the act of baptism. He proceeds by analogy with other instances of causal relevance we find unproblematic, such as the idea that a liquid can break a glass beaker in virtue of its heat and not in virtue of any of its other properties.

Section 3.4

I argue that Deutsch's argument fails because causal relevance is not sufficient for securing determinate reference. In any single baptism, many properties are causally relevant; but given Deutsch's strategy, we need a single property to be causally relevant for baptisms to ground reference determinately. The analogy with unproblematic cases of causation is irrelevant because we have no issue with many properties being causally relevant to a single event occurring. This contrasts with our requirements for unique and determinate reference. I conclude that unmediated causal theories of reference do not as of yet provide a satisfactory answer to the CDQ.

CHAPTER 4

This chapter overviews certain naturalistic theories of mental content, namely those that started as causal theories, gradually grew into causal-informational theories, and finally took the form of teleosemantics, which is now the predominant philosophical theory of mental content. Much of the development of these theories happened in reaction to issues with indeterminacy and misrepresentation. The chapter concludes that even if teleosemantics could resolve its issues with indeterminacy, it would have a deeper issue: an inability to properly account for rational thought and behaviour.

Section 4.1

This section briefly introduces causal theories of mental content and their connection to the project of naturalisation. I overview Fred Dretske's causal-informational account of mental content and its struggles with accommodating the possibility of misrepresentation.

Section 4.2

This section begins with an explanation of how and why causal theories had to introduce the notion of function to deal with the problem of misrepresentation. Given causal theorists' commitment to naturalisation, biological functions are an obvious candidate for this role. The result of developing the notion of biological function into a fully fledged theory of mental content is teleosemantics. Two kinds of theories are described: producer (input-oriented) and consumer (output-oriented). Both struggle with content indeterminacy.

Section 4.3

This section serves to uncover the methodological assumptions that lead to a larger problem for teleosemantics, which will be the focus of Section 4.4. I isolate these methodological assumptions through a back-and-forth between Angela Mendelovici (2013, 2016) and Marc Artiga (2013), who supposedly disagree over whether theories of content should allow for reliable misrepresentation. The real disagreement is related to the way we access mental contents and, ultimately, what the object of inquiry of a theory of content should be.

Section 4.4

This section connects the conversation between Mendelovici and Artiga to similar discussions between externalists and internalists about mental content. The main tension in these discussions arises due to the incompatibility of two plausible principles: the idea that we have "armchair" access to mental contents on the one hand, and the idea that there are cases in which external factors determine mental contents on the other. I adapt anti-externalist considerations by Boghossian (1994, 1997) to argue that teleosemantics struggles to describe rational thought and behaviour as such. This means that the notion of content provided by teleosemantics differs both from our first-person judgements and from our common sense third-person judgements about mental content. It becomes suspect that the object of inquiry of teleosemantics is mental content and not some other phenomenon. The conclusion is that teleosemantics, along with previous iterations of causal theories of mental content, is unable to adequately answer the CDQ.

CHAPTER 5

The final chapter of this thesis considers theories that advance the idea that phenomenal properties give rise to representational properties of mental states. These theories have recently risen in prominence, mainly due to the perceived failure of alternative theories of content. I will argue that they cannot meet the sceptical challenge, and that they are unable to account for identity between contents.

Section 5.1

This section introduces phenomenal intentionality theories (PITs). PITs state that at least some mental states have their representational contents in virtue of their phenomenal properties. I overview some of the most common arguments in support of PITs. These include arguments based on the first-person intuitive plausibility of the dependence of content on phenomenology and the fact that PITs can presumably avoid the indeterminacy issues that plague alternative theories of content.

Section 5.2

This section examines how PITs fare when confronted with the sceptical paradox. I first consider Kripke's own brief comments on theories that base meaning on "an irreducible experience, with its own special quale, known directly to each of us by introspection" (1982/1995, p. 41). I then expand these considerations to acknowledge theories that take phenomenality to inherently present subjects with accuracy conditions. There seem to be two ways of conceiving of phenomenal properties. The first takes phenomenal properties to be the part of our experience responsible for *what it seems like the world is* while undergoing the experience; I call these *impure* phenomenal properties. The second takes phenomenal properties to be the "what it's like" aspects

of undergoing an experience; I call these *pure* phenomenal properties. I conclude that a theory of content based on impure phenomenal properties would violate the non-circularity requirement, while pure phenomenal properties are unsuited to be the basis of representational content.

Section 5.3

In this section I argue that PITs cannot account for identity between contents. This is because, as I demonstrate, even relatively conservative theories of phenomenal intentionality lead to a contradiction when conjoined with the assumption that two mental states can have the same representational content. The argument proceeds from a colour-continuum case in which a subject reports that the contents of their experiences of adjacent shades are the same. One can easily infer a contradiction from these reports. While other theories of content may challenge the truth of the subject's reports, PITs do not have this option due to their reliance on introspection. I end this section by arguing that PITs have no way of avoiding this methodological reliance on introspection without abandoning the basic thrust of the phenomenal intentionality project. This leads me to conclude that PITs are not a promising avenue for answering the CDQ.

1. KRIPKENSTEIN'S PARADOX

INTRODUCTION

Kripke's *Wittgenstein on Rules and Private Language*[10] (1982) deals with the following question: In virtue of what does someone mean something by an expression?[11] Contrary to what one might assume given the prominence of Wittgenstein's name in the title of this book, Kripke is not overly concerned with being faithful to the arguments presented in *Philosophical Investigations* (Wittgenstein, 1953/2009). Kripke introduces his own argumentation by saying it should be taken as "Kripkenstein's" argument – "Wittgenstein's argument as it struck Kripke" (1982, p. 5). This does not only mean that the argument is not necessarily found in Wittgenstein, but that it does not reflect Kripke's own philosophical views either ("With few exceptions, I am *not* trying to present views of my own", 1982, p. 5). Kripke is simply interested in presenting an unusual argument that came to him while reading Wittgenstein's work.

The conclusion of Kripkenstein's argument is surprising and even paradoxical: There are no facts in virtue of which one can be said to mean anything. The argument is paradoxical because, if it is successful, it should not be possible for *any* arguments to be given – proposing an argument

10 From this point on I will use the abbreviation "WRPL" for "Wittgenstein on Rules and Private Language."

11 Notice that this question is a more specific formulation of the CDQ.

without words or sentences having meaning is simply impossible. So, the argument is self-refuting, and its conclusion obviously suspect.

Whatever the issues with Kripkenstein's paradox are, much insight can be gained from the arguments that lead to the "incredible conclusion" that there is no such thing as someone determinately meaning something. This chapter will be dedicated to working out those insights in a manner that allows for their application in the rest of this thesis.

The chapter is structured as follows. Section 1.1 contains a straightforward reconstruction of Kripke's sceptical paradox. Section 1.2 overviews solutions to the paradox considered by Kripke – all of them discarded for various reasons. Section 1.3 outlines the lessons to be taken from the argument and clarifies what its targets are. The basic lesson to be learned is that an adequate theory of content must respect three requirements: the extensions it confers to representations must be appropriate; it must explain the normative force of representations; and it must do all this in a non-circular manner. The section ends with a few notes about the shift from scepticism about meaningful occurrences to scepticism about meaning *simpliciter*.

1.1 THE SCEPTICAL CHALLENGE

Kripke's sceptical argument takes the form of an unusual scenario. The scenario is the following: Kripke, like most people today, is confident in his ability to add. He also has no doubt about what he means by "plus": he means the addition function. However, there are additions he has never performed – for example, there is surely a number such that it is larger than any number he has ever added. Imagine for ease of exposition that 57 is the largest number Kripke has ever added, and that consequently 68 plus 57 is one of the sums he has never performed before. Faced

with this new problem, Kripke answers that 68 plus 57 equals 125. Along with being sure of the mathematical correctness of his solution, Kripke also believes himself to be *correct in his usage of the word*, in the sense that he intends, as he always did before, to use "plus" to denote the addition function, and that that function yields 125 when applied to 68 and 57 (the answer is correct "in the metalinguistic sense" – Kripke, 1982, p. 8).

A strange sceptic approaches Kripke and challenges his answer: The answer Kripke should have given is that 68 plus 57 equals 5! He insists that Kripke has always used the word "plus" to mean not addition, but *quaddition*, a strange function that works exactly like addition up until the addends are smaller than or equal to 57; otherwise, the output is 5.

Quaddition (*): $x * y = x + y$ if $x, y \leq 57$;

$x * y = 5$ otherwise.

What the sceptic is saying seems immediately absurd. Kripke meant to *add,* not *quadd,* both in this instance and in all past instances; what the sceptic is saying is false. It seems clear to him that whenever he performs a new addition, he does not take an "unjustified leap in the dark" (Kripke, 1982/1995, p. 10). But because of how this example was crafted, there are no past instances of Kripke adding that he can cite to defend his claim that he was adding and not quadding. His past behaviour is perfectly consistent with him having quadded this whole time (and also with him meaning to perform infinitely many other deviant functions similar to the quaddition function).

The sceptic's strange claim and the fact that we have no immediate reply to falsify it are troubling. As the problem obviously generalises, the sceptic can cast doubt over whether anyone means anything by any expression – "it seems the entire idea of meaning vanishes into thin air" (Kripke, 1982/1995, p. 22). But in virtue of what is the sceptic's claim that Kripke was quadding, not adding, false? In other words, what makes it the case that Kripke uses the word "plus" to mean addition and not something else?

1.2 DISCARDED "STRAIGHT" ANSWERS TO THE SCEPTIC

The rest of chapter 2 of Kripke's book is devoted to searching for and discarding various answers as to which facts could falsify the sceptic's claim. Unfortunately, none of the "straight" answers to the sceptic considered by Kripke – that is, none of the theories that aim to preserve the standard notion of meaning[12] – resist under pressure.

The first instinct one has is to try to *define* adding, or explain the procedure "behind" adding, in a way that excludes the possibility of quadding (Kripke, 1982, pp. 15–17). So, for example, someone might say: "I am adding and not quadding because, when I add two numbers, I envision two heaps of pebbles representing those numbers. To get the sum of the numbers represented by these pebbles, I simply put them together and count; the resulting number is the sum." The sceptic could then say that that's exactly what one should do if they are quadding; take two heaps of pebbles and *quount* them, where *quounting* works just like counting, except when the number of what is being counted exceeds 57, in which case the result of our counting should always be 5. This kind of trick can be repurposed for any other interpretation, definition, or procedure we might use to explicate what we mean by "plus." The sceptic can always propose a new deviant meaning for the words used in the interpretation – a deviant use we do not seem to be able to rule out.

Another proposal that may immediately come to mind is that somehow a speaker's *intention to add* is what determines that they are, in fact, adding and not quadding (Kripke, 1982/1995, p.

12 What "standard notion of meaning" means here is similar to what Hattiangadi defines as "semantic realism" (2007, pp. 12–13). Semantic realism amounts to the idea that representational content is to be understood in terms of correctness conditions and truth conditions; and that ascriptions of specific mental states or meanings are factual, that is, true or false. Hattiangadi individuates semantic realism as the target of Kripkenstein's attack.

21). But the sceptic can immediately ask what makes it the case that the speaker *intends* to add and not quadd. Our explanation of what makes it the case that one's intentions have the content that they do does not seem to fare any better than our explanation of what makes it the case that one means a certain thing by "plus." The problem is evocative of some considerations given by Wittgenstein in *Philosophical Investigations*:

> This was our paradox: no course of action could be determined by a rule, because every course of action can be brought into accord with the rule. The answer was: if every course of action can be brought into accord with the rule, then it can also be brought into conflict with it. And so there would be neither accord nor conflict here. […] For what we thereby show is that there is a way of grasping a rule which is not an interpretation, but which, from case to case of application, is exhibited in what we call "following the rule" and "going against it." (1953, p. 87, § 201)

The relevant portion of this quote concerns the idea that there must be a way of "grasping a rule" *that is not itself an interpretation*. In Kripke's book, the grasping of a rule is taken to be analogous to our meaning something. In other words, there must be a way of meaning something that is not an interpretation, as we'd otherwise slip into a vicious regress.

We can generalise this point. Attempts at determining what Kripke means by "plus" will fail whenever they include anything that is itself vulnerable to an additional attack by the sceptic. As has been shown, this includes definitions, descriptions of procedures, and intentions. What all of these have in common is that they are content-laden. It seems clear, then, that attempting to determine what one means by relying on other contentful states is misguided and will inevitably lead to either vicious circles or vicious regresses.

An interesting consequence of the fact that we cannot rely on other content-laden states to explain away Kripkenstein's paradox is that the sceptic does not only target attributions of meaning to *speakers*, but attributions of contentful mental states as well. Kripke puts it in terms that were more appropriate to the time the book was written – he specifies that he assumes no "behaviourist" limitations as to what kind of facts can be cited to answer the sceptic (Kripke,

1982/1995, p. 14). In other words, the evidence for why he means the addition function by "plus" can also come from introspection, or, as Kripke says,

> Whatever 'looking into my mind' may be, the sceptic asserts that even if God were to do it, he still could not determine that I meant addition by 'plus.' (Kripke, 1982/1995, p. 14)

Paul Boghossian's review of the problem posed by Kripke and of some of the reactions to it includes a somewhat detailed explanation of why mental content is also targeted by the sceptical argument (Boghossian, 1989, pp. 509–514). The general lesson is the following: All bearers of intentional content are victims of Kripkenstein's monster.

One last point that needs to be made explicit here is that the sceptical challenge is not an epistemological challenge, but a metaphysical one (Kripke, 1982/1995, p. 21). The question is not about how we *know* that we mean addition by "plus," but rather *what makes it the case* that we mean addition by "plus."

Kripke considers several specific theories of meaning and their ability to rise to the sceptic's challenge. I will not go into detail about why Kripke believes none of the theories he considers are up to the task, as that would take up too much space. I will only briefly describe some selected critiques he puts forward. This will hopefully clarify what a theory of meaning should be able to do to avoid Kripkenstein's paradox.

Dispositional theories propose that what a speaker means by a word is determined by facts about their dispositions. Kripke dismisses dispositional theories on roughly two grounds: First, dispositions do not seem to be the kind of thing that can tell us how one ought to use a word. The way one is *disposed* to act does not tell us how one *ought to* act. This is an issue for Kripke because, as he set it up, the sceptic asks why Kripke does not make an "unjustified leap in the dark" when he says that 68 plus 57 is 125. In other words, meaning addition by "plus" somehow *justifies* our answer, and the facts that determine Kripke meaning addition by "plus" should be able to account

for that normative force. But facts about dispositions are not normative; so dispositionalism will not do.

A related issue stems from the notion of error, or more precisely, from the idea that if something *represents*, it should be possible for it to *misrepresent* too. A theory of meaning should be able to account for incorrect uses of language. Dispositionalism will naturally struggle with this requirement: If one's meaning "plus" is fixed by their disposition for using the word "plus," then the meaning of "plus" will include all their (very human) dispositions to make mistakes when tired, to forget to carry, to refuse to add when the numbers are too large and so on. The difficulty here is that if someone's actual dispositions to add fix the meaning of "plus," "plus" does not denote the addition function, and the dispositionalist has no way of categorising mistakes *as* mistakes; the speaker was using "plus" in accordance with their dispositions, which fix the meaning of "plus." There is no obvious way of distinguishing between the uses of "plus" that are correct from those that are incorrect, no way of distinguishing the "good" meaning-constituting dispositions from the "bad" non-meaning-constituting dispositions which we'd prefer to exclude from the extension of "plus." Assuming that there is a way to distinguish between the meaning-constituting and the non-meaning-constituting facts would amount to begging the question.

Of course, the version of dispositionalism we have criticised until now is very crude and it can be enhanced with idealisations and ceteris paribus clauses. For example, we could say that I mean addition by "plus" in virtue of my dispositions to add on the condition that I am well-rested, have enough time to devote myself to adding, and so on. However, even very detailed and intricate refinements of the dispositional theory fall into one of two traps: they either beg the question by

assuming what the correct extension of a term is, which usually renders the theory unable to account for misrepresentation, or they do not yield the appropriate extension of the term.[13]

Kripke also briefly considers the possibility that what one means by a sign could be determined by an introspectively accessible irreducible experience "with its own special *quale*" (1982/1995, p. 41). He states that even if we granted that meaning something by "plus" had such a special quale, which is doubtful, that would not resolve the sceptical challenge. The big issue is similar to the one faced by dispositional theories: there is no way for this type of answer to account for the normative aspect of meaning. In other words, there is nothing about having a certain *quale* associated with a meaning that can determine that I *ought* to use a sign in a certain way. Facts about qualia are not normative; and we are looking for facts that can explain why one is guided in their application of a term. There seems to be no sense in which a unique *quale* can tell anyone they ought to answer "125" when asked what 68 + 57 is. Kripke is especially quick in his dismissal of *qualia*-based theories of meaning because they had few supporters at the time. Times have changed, and a new kind of theory of content has emerged: Theories of phenomenal intentionality argue precisely that the phenomenal feel of experience determines representational properties. I will devote more attention to this kind of answer to the problem of content determination in Chapter 5, which is dedicated to the rise of phenomenal theories of intentionality.

13 Kripke makes this exact argument in (1982, pp. 27–32), and more obliquely in (1982, pp. 31–35). For a newer analysis of subsequent dispositionalist accounts and their failures to respond to the sceptical challenge, see (Hattiangadi, 2007, pp. 105–120)

1.3 TAKING STOCK: KRIPKE'S ARGUMENT, ITS INTERPRETATIONS, AND ITS CONSEQUENCES

The argumentation presented by Kripke is dense and intricate, so I will try to summarise it here. The assumptions that lead up to the sceptic's challenge are the following: As we usually understand it, a person meaning something by a word seems to grasp something analogous to a rule by which she then applies that word in future cases. She is guided by the rule in her application of the word in potentially infinitely many new cases, and whatever she grasps also determines whether the applications of the word are correct and how the word should be used. When I have grasped the meaning of the word "horse," I have (it seems) grasped a rule which allows me to apply it correctly whenever I want to refer to horses (though, of course, I may still make a mistake). However, the sceptic argues, there are no facts that can account for anything resembling the "instructions" we grasp when we mean something. In other words, there seems to be no good answer as to which facts make it the case that I mean horses by "horse." Kripke considers several theories that individuate different meaning-determining facts. All these theories fail for one or more of the following reasons:

- The theory proposes facts that underdetermine facts about meaning; they are consistent both with one's meaning addition and quaddition (and infinitely many other deviant functions) by "plus"
- The theory proposes facts that cannot account for the "instructive," "guiding" force that is characteristic of meaning, i.e., they don't reveal why one *should* answer "125" when asked what 68 + 57 is.
- The theory proposes facts that concern other content-laden states, that is, they are not a "way of grasping a rule which is not an interpretation."

We have now clarified the basic challenge theories face in order to keep the sceptic at bay: finding the facts in virtue of which speakers mean something by a sign without falling into regress. These facts, the argument goes, should meet three requirements. The first requirement is that meaning-determining facts should yield the appropriate *extensions* of expressions meant by speakers; so, these facts should determine that what Kripke means by "plus" is the addition function and not some deviant, quus-like function(s). The second requirement is that meaning-determining facts should account for the normative force of meaning, that is, the facts in virtue of which Kripke means the addition function by "plus" should make it so that Kripke *ought to* use the word "plus" in a certain way. Kripke assumes that whichever facts determine the extensions of our terms also have to be the ones providing the justification for why I ought to use those terms a certain way; he believes that these two requirements are related (1982, p. 11). The third requirement is that the facts in virtue of which one means something by a sign should not involve other content-laden phenomena, such as thoughts, intentions, or descriptions. We must avoid falling into the trap of relying on a "rule for interpreting a rule" (Kripke, 1982/1995, p. 17). The third requirement should remind us that the argument concerns contentful mental states as well – in other words, though the debate has usually focused on the meaning of utterances, the issue is with representational content at large.

We can now more precisely formulate the three requirements an adequate theory of representational content should meet, according to Kripkenstein:

- Extensionality: An adequate theory should ensure that representations have the appropriate extensions.
- Normativity: An adequate theory should ensure that representations are appropriately normative.

- Non-circularity: An adequate theory should ensure that representations (ultimately) do not have the content they do in virtue of other contentful states.

The fact that the sceptical paradox poses these three requirements is relatively uncontroversial. However, readers of WRPL diverge on many of the details required to flesh out these three requirements; for example, there are large differences in how the normativity requirement has been understood. Some of these interpretative differences will become relevant later in this work – Chapter 2 is in large part dedicated to working out the details and the impact of the normativity requirement.

In the interest of precision, one note on a common interpretation of Kripke's argument should be made here. Many philosophers have made an implicit move: shifting the argument's target from attributions of meaning to speakers (Kripke means the addition function by "plus") to meaning *tout court* ("plus" means the addition function). In other words, many have taken the paradox to have implications not only for the factuality of speakers *grasping* token meanings and concepts – e.g. it does not only impact speaker meanings – but for meaning itself. The move is characterised by a common shift from "nobody means anything by any expression" to "there are no meanings," which is prima facie a more radical conclusion. For example, McDowell says that the sceptic's opponent "has renounced the right to attribute meaning to expressions at all" and that the sceptical solution presumes that "we must reform our intuitive conception of meaning, replacing the notion of truth-conditions with some notion like that of justification conditions" (McDowell, 2002, pp. 50–51). McGinn takes the sceptic as "demanding an answer to the question 'what does meaning/reference consist in?", or, in other terms, "the skeptic is in effect asking for a non-semantic analysis of what it is for a word to refer to a particular object" (McGinn, 2002, p. 81). Boghossian puts the argument in the following terms: "Having a meaning is essentially a matter of possessing a correctness condition. And the skeptical challenge is to explain how

anything could possess that" (2002, p. 149). More recently, Anandi Hattiangadi individuates the target of the sceptic's argument as "semantic realism" (2007, pp. 12–13). Semantic realism amounts to the idea that representational content is to be understood in terms of correctness conditions and truth conditions; and that ascriptions of specific mental states or meanings are factual, that is, true or false. In other words, she takes Kripkenstein's argument to question the standard notion of meaning and representation, and not *just* the factuality of ascriptions of meanings to speakers.

Regardless of the ubiquity of the assumption that the sceptic targets not only ascriptions of meanings to speakers but meaning *tout court*, the argumentative shift from one to the other is not obvious in *WRPL*. Daniel Whiting is seemingly aware of the distinction, but apparently believes that the sceptical argument targets both ascriptions of speaker meaning and standing, "linguistic" meaning:

> According to Kripke's Wittgenstein's skeptic, there is simply "no such thing" as meaning. […] Kripke sometimes talks of what an expression means – that is, of linguistic meaning – but more often he talks of what a speaker means by an expression – that is, of speaker meaning. I focus here on linguistic meaning. (Whiting 2024, p. 124 and footnote 2)

Crispin Wright is one of the rare philosophers to address the distinction explicitly. He does so by noticing that the sceptical argument demonstrates that nobody means anything by an expression, which is different than demonstrating that no "impersonal" meanings exist. However, he does not provide justification for inferring one from the other beyond the following:

> "Remember that the argument will have involved an extensive idealization of your knowledge of your previous behavior and mental history: you will have been granted perfect recall of all such facts. […] Hence, in the presence of the idealization, there can be no facts about what anybody means by any expression. And it is impossible to see how, consistently with that admission, there might yet be facts about what expressions, as it were impersonally, mean." (Wright, 1984, p. 763)

Wright's argument can be reconstructed roughly as follows:

(I). There are no facts about what anybody means by any expression. (This is the conclusion of the sceptical argument.)

(II). If there are no facts about what anybody means by any expression, then there are no facts about meanings.

So,

(III). There are no facts about meanings.

(II) is an undefended assumption which is implicit in the reactions of most of the philosophers who reacted to *WRPL*. Why has so much of the literature assumed, implicitly or explicitly, the passage from (I) to (III)?

To understand why the non-factuality of ascriptions of meaning to speakers affects meaning *tout court*, we must turn to some methodological considerations about the philosophy of meaning. Meanings *tout court*, whatever our theory of such meanings might be, are postulated as having a certain theoretical role. That theoretical role of meaning is that it should be part of what explains the meaningfulness of signs broadly understood, e.g., sentences, words, but also linguistic utterances, written text, singular thoughts, and so on.[14] While this may seem almost vacuous, it does put some constraints on our theorising. The basic assumption is that *there are* meaningful things whose meaningfulness requires explanation; otherwise, it is unclear how or why inquiry into meaning would even begin. The existence of these meaningful occurrences is the primary reason it is explanatorily useful to postulate the existence of meaning. If there were no meaningful occurrences, it is unclear what the explanatory role of meaning would be. What evidence would prompt us to believe in such strange things?

14 There may be controversy over what kinds of things count as signs, that is, as meaningful instances. However, that is not a question to be settled here, as its answer does not affect the broader point made in the paragraph.

This methodological point should not be taken to imply that meaningful tokens are somehow metaphysically more fundamental than meaning; that is an independent matter, to be settled on independent grounds. It should also not be taken to mean that the aim of a theory of meaning is *exclusively* the explanation of such meaningful tokens. It is unquestionable that semantics has evolved far beyond the scope of explaining the meaningfulness of natural languages. The real impact of the methodological considerations made above is that the existence of at least some actual meaningful occurrences is the starting point of inquiry for a philosophy of meaning, and that absent those occurrences, it becomes unclear what role the notion of meaning is supposed to play. If there are no meaningful utterances or thoughts, meaning's explanatory raison d'être vanishes.

It is now clear why the conclusion of Kripkenstein's sceptical paradox – that nobody ever means anything by a sign – brought many to conclude that there are no meanings *tout court*. If there are no occurrences of meaningful signs, it seems as if there is no reason to postulate standing meanings. However, given that the Kripkensteinian paradox only directly affects the meaningfulness of single expressions, and since I have no need to take a position about the existence of standing meanings to gauge whether the CDQ has been answered, the present work will focus on meaningful occurrences. In other words, when testing the adequacy of a theory of content, the standard applied here will be the following: Is the theory able to explain in virtue of what a single meaningful occurrence has its content? When it comes to linguistic meaning, the question will take the form of wondering what makes single utterances (or similar meaningful linguistic expressions) have the content that they do. When we turn to mental "meaning" in Chapters 4 and 5, the question will take the form of wondering what makes single mental representations have the content that they do.

CONCLUSION

To counter the meaning scepticism advanced by Kripkenstein, the theories considered in this thesis will need to explain in virtue of what signs have meanings. In other words, to successfully counter Kripkenstein, a theory needs to outline the metaphysical conditions under which a representation has a content.

As has been explained in this chapter, the meaningful "signs" in question may be either mental or linguistic; as long as something may be said to have a determinate content, Kripkenstein will have been "defeated." Chapters 3, 4, and 5 are dedicated to exploring whether a selection of recent theories within the philosophy of language and mind have managed to avoid the kind of vulnerabilities highlighted by Kripke's sceptic.

When considering contemporary theories of representational content in the rest of this thesis, I will often come back to ideas presented in this chapter, and in particular to the "three requirements:" extensionality, normativity, and non-circularity. This framework will help determine whether new approaches have managed to evade the sceptic's argument; and, if they fail, it will illuminate the reasons why. I will advance the thesis that all attempts at answering the CDQ considered in this thesis fail, in many cases in ways that can be directly traced back to failures in meeting one of the three requirements.

Before directly engaging with contemporary theories of content, it is important to clarify what the normativity requirement posed by the sceptical paradox entails. The next chapter serves to advance the understanding of this controversial portion of Kripkenstein's argument.

2. THE NORMATIVITY OF MEANING

INTRODUCTION

The argument put forward by Kripke in *Wittgenstein on Rules and Private Language* has been interpreted in many different ways since its publication. The central argument in WRPL*,* also known as Kripkenstein's paradox, is usually taken to show that nothing determines whether speakers mean anything by a sign. It is much less clear *how* the argument reaches its conclusion. It can be safely said that Kripke highlights two issues that can be separated: the difficulty that extant theories have in specifying, without falling into vicious circularity, which facts determine the extension of terms; and the difficulty that extant theories have in explaining the normative aspect of meaning in a non-mysterious way. This chapter focuses on the latter difficulty, which has been referred to as "the normativity constraint" (Wright, 2002; Glüer et al., 2024), "the normativity argument" (Zalabardo, 1997), and "the argument from normativity" (Guardo, 2009).

The normativity requirement as it originally appears in *WRPL* can be reconstructed as follows. Kripke notices that we have a set of intuitions about meaning that need to be accounted for in our response to the sceptic's challenge.[15] He says that "…we think of ourselves as *guided* in our application" of a rule when it comes to new instances (Kripke, 1982/1995, p. 17); he feels as if meaning is something that "*instructs* [him] what [he] ought to do in all future cases" (p. 22); and

15 For a reconstruction of the sceptical challenge and of subsequent interpretations in the literature refer to Chapter 1.

that the sceptic puts pressure on the notion that our usage of language is *justified* and *not arbitrary* (p. 23). Meaning-determining facts, then, should account for these intuitions; they should determine what I *should do* given that I mean something by a sign (p. 24).

Kripke goes on to argue that extant theories of meaning struggle to account for the "guiding" force of meaning, or the way we feel as if we "should" use words in a certain way. The basic intuition Kripke relies on is vividly described on p. 22 of WRPL:

> Even now as I write, I feel confident that there is something in my mind - the meaning I attach to the 'plus' sign - that instructs me what I ought to do in all future cases. I do not predict what I will do […] but instruct myself what I ought to do to conform to the meaning.

The most precise illustration of how Kripke believes this normative feature of meaning leads to paradox can be seen in his subsequent critique of dispositionalism: he says that even if facts about speakers' dispositions could fix the extensions of terms properly, they are nevertheless the wrong kind of facts because they are *descriptive* and not *normative* (1982/1995, p. 37). Facts about dispositions can clarify how speakers *will* or *do* behave, but not how they *should* behave; because of this, they cannot be an adequate determiner of meaning.

The normativity requirement is taken to be a crucial aspect of Kripkenstein's argument and has been discussed at length by Paul Boghossian (1989), Hattiangadi (2007), Allan Gibbard (2012), Hannah Ginsborg (2022), and many others. Though there is agreement about the importance of this part of the argument, the details regarding how exactly a theory of content should meet the normativity requirement differ considerably.

The aim of this chapter is to argue that meaning is not genuinely normative, and that because of this, the intuitions we have about being guided in our use of language should not be explained via standard truth-conditional meaning, but through the existence of linguistic conventions that regulate felicitous language use.

The chapter is structured as follows. Section 2.1 presents two ways the normativity requirement has been interpreted, which correspond to two ways the Kripkensteinian argument can be interpreted. Roughly, the first interpretation assumes meaning to be genuinely, categorically[16] normative in the way moral norms are thought to be, while the other takes meaning to be "weakly" normative, in the sense that it only presents us with semantic correctness conditions. Whether one or the other interpretation should be adopted depends on whether meaning is in fact genuinely normative or not. However, standard truth-conditional meaning is not genuinely normative, since we do not have a categorical obligation to say the truth. Section 2.2 explores whether there could be other sources of categorical normativity in our linguistic practices; in particular, I consider the notion of linguistic correctness. In Section 2.3, I defend the idea that linguistic correctness is constitutive of conventional meaning, which means that if one wants to speak, they should do so in a linguistically correct manner. However, this is not enough to conclude that the norms presented by linguistic correctness are categorical. In Section 2.4, I investigate whether we can obtain categorical norms from constitutive norms by comparing the discussion of meaning norms to that of moral norms. There have been attempts to derive categorical moral norms from what is constitutive of being an agent; however, such attempts have been convincingly thwarted by the arguments in David Enoch's "Agency, Shmagency" article (2006). I conclude that there are no categorical norms related to language, meaning that it is most charitable to interpret Kripke's argument as *not* assuming that there are genuine meaning norms of any kind. However, the intuitions about feeling guided and justified in our language use still need to be explained; I argue that linguistic correctness should be used for this task.

[16] A definition of categorical norms will be provided in Section 2.1.

2.1 TWO INTERPRETATIONS OF THE NORMATIVITY REQUIREMENT

The goal of this section is to provide a more precise understanding of what constraints the normativity requirement poses. There are different ways of interpreting what the requirement is, which leads to different constraints. I will consider two main ways of understanding the normativity requirement here: one which takes it as expressing that for expressions to have meanings, they must have correctness conditions; and one which takes it as expressing that meaning is categorically normative, that is, that it involves norms that are inherently motivating.

A popular proposal is to interpret the "meaning is normative" slogan as "meaningful expressions have *correctness conditions*." In other words, meaning is normative in the sense that there is a correct (and incorrect) way of using language. The job of responding to the sceptic, then, amounts to showing in virtue of what do meaningful expressions have correctness conditions. The idea that the existence of correctness conditions implies that meaning is normative is now commonly known as "the simple argument" for the normativity of meaning (Glüer et al., 2024). The argument is first found in Boghossian (1989); Whiting (2016) also relies on this argument. Proponents of the simple argument argue that a straight answer to the sceptical challenge should consist in explaining how anything could have correctness conditions, where correctness conditions are usually understood in terms of truthful use (Boghossian, 1989, p. 513). For example, I *should* say that 68 + 57 equals 125 because it is correct to do so, in the sense that it is *true* that 68 + 57 equals 125. I will refer to correctness conditions understood in terms of truthful use as *semantic correctness conditions* in this work. Proponents of the simple argument believe that meaning is normative precisely because there are semantic correctness conditions. This notion of normativity

is easily applicable to mental content, too; our beliefs, thoughts, and even perceptions have correctness (or accuracy) conditions in virtue of being true or false.

This interpretation of the normativity of meaning flattens the distinction between the extensionality and normativity requirement within Kripkenstein's argument: providing an answer to the former automatically provides an answer to the latter.[17] If the normativity of meaning can be understood as the mere existence of semantic correctness conditions, explaining how the extensions of terms are determined immediately explains how we *should* use them: We should apply terms precisely to their extensions. So, for example, if we determine what makes it the case that the extension of "horse" corresponds to horses, we automatically determine how one ought to use "horse" – they ought to use it to talk about horses. The notion that there is a correct way to use words implies that there is an *incorrect* way to use them: if I point at a piece of cheese and say "that's a horse," I am speaking incorrectly. As Boghossian puts it, "the notion of the extension of an expression just is the notion of what it is correct to apply the expression to"[18] (1989, p. 530).

This simplifies the work of Kripkenstein's opponent considerably. If the simple argument is the right way to interpret the normativity requirement, providing an appropriate example of how semantic correctness conditions (or extensions) are fixed would solve the sceptical paradox. By this interpretation, Kripkenstein's argument is not particularly strong; all it does is demonstrate that the particular theories he has listed and critiqued are not up to the task, that is, they have not

---

[17] The extensionality requirement in Kripkenstein's argument represents the challenge in answering what fixes the extensions of our terms. For a more in-depth explanation, refer back to Chapter 1.

[18] The terms "apply" and "application" are usually used, and will be used in this work, as a placeholder for instances of the relation that (presumably) obtains between representations and the world. Hattiangadi clarifies this point: "The expression 'applies correctly' is a placeholder for the various semantic relations an expression can have to the world: it stands for either 'x refers to a', 'x denotes a', or 'x is true of a'"(2007, p. 52).

successfully individuated which facts determine the extensions of the terms used by speakers. This is quite different from a strong *a priori* argument that shows that no *possible* theory could be up to the task.

A second, stronger interpretation of the normativity requirement is proposed by Hattiangadi (2007). This fortified interpretation leads to Kripkenstein's argument gaining *a priori* status; it becomes an argument that rules out the possibility that *any* theory may individuate the facts that determine what a speaker means by a sign. Hattiangadi proposes that a way of interpreting Kripkenstein's paradox is to see it as analogous to certain arguments against reductionism about moral facts (2007, pp. 43–45), and in particular with a revised version of G. E. Moore's Open Question Argument (Moore, 1903/1922, pp. 10–21).

The argument against the reducibility of moral facts to natural facts can be summarised as follows:

(I).Moral facts are normative (prescriptive or inherently motivating). (Assumption)

(II).Natural facts are not normative (prescriptive or inherently motivating). (Assumption)

(III).Normative facts cannot be reduced to non-normative facts. (Assumption)

(IV).Moral facts cannot be reduced to natural facts.

The idea that the normative cannot be reduced to the non-normative follows a tradition leading back to Hume, who expresses this idea in terms of the impossibility to derive or deduce statements containing an "ought" from statements containing an "is",[19] an idea that is to this day called

[19] The relevant text can be found in the *Treatise on Human Nature*, book III, part 1, section 1, where Hume argues for the independence of morality from reason. The very last paragraph of the section contains a

"Hume's law."

After this first step, following a line of argument defended by J. L. Mackie (1977/1990, pp. 37–49), one can move from the conclusion that moral facts are irreducible to natural facts to the conclusion that there are no moral facts at all:

(V). If moral facts cannot be reduced to natural facts, then they are non-natural.

(VI). Moral facts are non-natural.

(VII). There are no non-natural facts.

(VIII). There are no moral facts.

If moral facts cannot be reduced to natural facts, they are a sui generis class of unnatural facts that has some peculiar and unexplained property, namely inherent action-guidance (or something similar). These facts are "queer," create epistemic problems, render some causal relations mysterious, and lack appropriate philosophical explanation. The conclusion is that we should forgo moral facts altogether and accept that all our usual moral judgements are untrue;[20] this is known as moral error theory within ethics.

---

famous passage that is usually cited in defence of Hume's law (italics mine): "In every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, *is*, and *is not*, I meet with no proposition that is not connected with an *ought*, or an *ought not*. This change is imperceptible; but is, however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given, for what seems altogether inconceivable, *how this new relation can be a deduction from others, which are entirely different from it.*"

[20] Moral error theory usually takes judgments about moral properties, such as "stealing is wrong," to be false. However, depending on the position one takes regarding talk about non-existing entities, these kinds of judgments may simply be meaningless. For our purposes, we can simply commit to the idea that the argument demonstrates, at least, that it is never true that semantic properties are instantiated. I am grateful to Andrea Bianchi for making me notice this issue.

We can see how this line of argument can analogously be applied to semantic facts if we assume that they are normative:

(I).Semantic facts[21] are normative (prescriptive or inherently motivating). (Assumption)

(II).Natural facts are not normative (prescriptive or inherently motivating). (Assumption)

(III).Normative facts cannot be reduced to non-normative facts. (Assumption)

(IV).Semantic facts cannot be reduced to natural facts.

(V).If semantic facts cannot be reduced to natural facts, then they are non-natural.

(VI).Semantic facts are non-natural.

(VII).There are no non-natural facts.

(VIII).There are no semantic facts.

If semantic facts are normative, that means that they are prescriptive or inherently motivating, and so they cannot be reduced to natural facts that lack these properties. But the existence of sui generis, "queer" semantic facts is unacceptable, as it creates epistemic and causal problems. If semantic facts are neither reducible to natural facts nor can we accept their existence as *sui generis* facts, we are left with one option: to negate their existence. As Hattiangadi puts it, "Kripke could argue that semantic facts cannot be non-natural because they would be unknowable, without the assumption of a peculiar epistemology, and because they would be metaphysically queer; quite unlike anything else in the universe" (2007, p. 50). Analogously to what Mackie proposes about moral judgements, then, we should accept that semantic judgements such as "Kripke means the addition function by "plus"" are untrue. We can call this stance "semantic error theory" in analogy with moral error theory.

[21] By "semantic fact" I mean a fact of the matter about an expression's or mental state's meaning.

In this form, the sceptical argument is a much stronger *a priori* argument against semantic realism: It leads to the conclusion that there simply are no semantic facts, meaning that no true theory of such facts can be formulated. For this strengthened sceptical argument to go through, however, there needs to be plausible justification of the assumption that semantic facts are normative in a similar way in which moral facts are normative. As we have seen, the simple argument seemingly provides support for this assumption: Semantic facts are normative because they are inseparable from semantic correctness conditions, which account for the "guiding force" Kripke was looking for. The fact that it is correct to apply "horse" to horses means I *ought to* apply "horse" to horses – therein lies the normativity of semantic facts.

There have been critics of the simple argument for the normativity of meaning. Specifically, it has been pointed out that semantic correctness conditions do not necessarily have any genuine normative force (Glüer & Wikforss, 2015; Hattiangadi, 2007, 2009), i.e., they do not determine what speakers *ought to do* in a robust sense. For example, Hattiangadi (2007, pp. 181–183) argues that semantic correctness as it is understood within the simple argument creates no *categorical* obligations; it can only create obligations in connection with external desires or motivations, that is, it can only create hypothetical obligations. To better understand this criticism of the simple argument, we should first define categorical and hypothetical norms. We may use Copp and Morton's definition of these two types of obligations:

> **Categorical obligations:** Obligations that are in force independently of a person's contingent desires or attitudes.
>
> **Hypothetical obligations:** Obligations that are in force in virtue of a person's contingent desires or attitudes. (Copp & Morton, 2022)

Hypothetical obligations can be described in a characteristic "if [you have desire/attitude X] then you ought to do Y" form, and importantly, they are reducible to non-normative facts. For example,

even though it is true that I ought to make dinner *if* I am hungry, I have no categorical obligation to make dinner; the "ought" here is completely dependent on the fact that I am hungry. This hypothetical obligation is the result of two (apparently) non-normative facts: the fact that I am hungry and the fact that hunger can be satisfied by making dinner.

Because hypothetical obligations are easily created and just as easily vanish, as they vary with personal attitudes, they are not examples of "robust normativity." There is nothing robustly normative about my obligation to make dinner; the obligation disappears as soon as my desire is satisfied or changes. In contrast, robust normativity figures in "ought" statements that are categorical – they are in force regardless of what persons feel, desire, or are interested in. Unlike hypothetical obligations, categorical obligations are thought to be *irreducible,* following Moore's and Mackie's insights (Copp & Morton, 2022). We can now see that semantic correctness conditions do not create categorical obligations, but only hypothetical ones. The obligation to apply "horse" to horses is dependent on speakers' desires to speak, and to speak truthfully; it is merely a hypothetical norm (Hattiangadi, 2007, pp. 189–190). Absent a desire to speak, we have no reason to apply any term correctly. Semantic correctness conditions don't present us with inherently motivating norms.

This brings us to a crucial point: For the stronger version of Kripkenstein's argument to succeed, meaning must exhibit genuine categorical normativity.[22] If semantic correctness conditions do not create categorical obligations, that means that proponents of the simple argument got something wrong; meaning is *not* normative in virtue of the existence of semantic correctness conditions, because semantic correctness conditions do not provide categorical

[22] Recently, this point has been contested by Hattiangadi (2024b, 2024a), who argues that weak normativity is sufficient to conclude that intentionality is in principle irreducible to the physical.

obligations. In other words, the simple argument provides no reason to believe that semantic facts are robustly normative; and if semantic facts are not robustly normative, that is, if they don't create categorical obligations for speakers, then there is no reason to believe that they are irreducible to natural facts. The *a priori* version of Kripkenstein's argument cannot go through.

To summarise the points made until here, Kripke's argument can be interpreted as an argument against the existence of semantic facts due to their normative character, analogously to how some philosophers argue that the inherently normative character of moral facts should lead us to reject their existence. The simple argument – that is, the argument that the existence of semantic correctness conditions is a sign of the normativity of content – gives support to the idea that semantic facts are normative. Critics of the simple argument, however, argue that the existence of semantic correctness conditions is not a sufficient reason for concluding that meaning is robustly normative. This criticism highlights a significant disanalogy between moral facts, which are supposed to be categorically normative, and semantic facts, which are apparently only hypothetically normative. If this criticism is warranted, semantic correctness conditions do not provide us with the kind of categorical obligations needed for the strengthened sceptical argument to go through.

Interpreting the normativity of meaning in terms of semantic correctness conditions leads to a significant weakening of Kripke's argument. In a sense, the simple argument can be reinterpreted as an oblique rebuttal of the idea that meaning is strongly normative: If the normativity of meaning simply is nothing more than the existence of semantic correctness conditions, and semantic correctness conditions do not provide us with robust normativity, our intuitions regarding the existence of semantic oughts – such as Kripke's feeling that he is "guided" or "instructed" in his use of language – rest on shaky grounds.

2.2 ALTERNATIVE SOURCES OF NORMATIVITY: A NEW TYPE OF CORRECTNESS[23]

Now that it has been established that semantic correctness conditions do not present us with categorical semantic norms, the pressing question at this point is whether we have any other reason to believe that meaning is genuinely normative. Is there a way to honour Kripke's original intuitions about being "guided" and "instructed" when speaking? This section will investigate whether there is a viable alternative way of justifying the notion that meaning is normative. An alternative source of meaning normativity could not only honour the intuition that meaning "guides" and "instructs," but would also vindicate the modified Open Question argument presented previously. In other words, if we could find an independent reason to believe meaning is genuinely normative, that would strengthen the argument against semantic realism and Kripkenstein's sceptical position.

One possibility is to consider whether the normativity in question might stem from a different kind of correctness, that is, a correctness that is not understood in terms of truthful use. Some philosophers have suggested that there are two different types of meaning-related correctness: correctness in virtue of truthful use and correctness in virtue of use in accordance with meaning[24] (e.g. Millar, 2002; Buleandra, 2008; Reiland, 2023). This is interesting because the primary focus in the literature about meaning normativity has been on what these authors call

---

[23] This section, along with Sections 2.3 and 2.4, borrows from a paper I published in *Phenomenology and Mind* (Papic, 2023).

24 The quoted philosophers have not used the exact terminology I am using in this work but have made the same distinction. For example, Millar just distinguishes true application and application in accordance with the meaning of an expression, which is precisely how linguistic correctness is defined here.

semantic correctness as a possible source of normativity. In other words, the alternative we are looking for might lie in the idea that the source of genuine meaning normativity is this other kind of correctness: linguistic correctness.

As was mentioned previously, correctness conditions are typically defined in terms of truthful use: One speaks correctly if they speak truthfully. Defenders of this distinction between semantic and linguistic correctness point out that there are intuitions supporting the idea that we can use language correctly even if we (intentionally or not) stray from true application. For example, whenever I lie there is a sense in which I am using language correctly, even though I am saying something false, because I am speaking in accordance with the words' established meanings.[25] I can also use an expression incorrectly even if I am saying something true. Typical examples include speakers who misspeak, e.g. someone who uses "arcane" while believing it is synonymous with "ancient." Substituting the two can contingently generate true statements which are nevertheless examples of incorrect use, like "this book contains arcane stories;" the book may very well contain both ancient and arcane stories, but the use of the sentence is incorrect in a sense that is not directly related to the truthfulness of the statement. Another sense in which one's use of language can be correct without being true is related to the fact that we can distinguish between appropriate and inappropriate ways of asking questions, giving orders, greeting someone, and so on, which are not truth-bearing portions of language.

25 This distinction has been made at least as early as in Anselm of Canterbury's De Veritate: "A statement then is right and true either because it is correctly formed or because it fulfils its function of signifying correctly. The former belongs immutably to it, the latter is mutable. The former it always has, the latter not always. The former it naturally has, the latter accidentally and according to use." (1998, p. 155). The first type of correctness is said to belong even to false statements, as long as they signify something. I owe this observation to Paolo di Lucia, who kindly directed me towards his 2011 paper which contains an insightful analysis and categorisation of the notions of correctness present in Anselm's work.

These examples indicate that there are two different senses in which we might speak (in)correctly: one that is directly linked to true and false use of language and one that is not. To clarify this distinction, whether it turns out to be substantial or not, we can preliminarily define the two types of correctness as follows:

SC: A statement S is **semantically correct** if S is true.

LC: A statement S is **linguistically correct** if S is used in accordance with its meaning.

"In accordance with its meaning" is somewhat imprecise, but it captures the basic idea behind linguistic correctness. A more precise proposal is to identify linguistic correctness with appropriate use-conditions, that is, with the conditions of felicitous usage of an expression. For example, an utterance of "bon appetit" is linguistically correct if it happens at the beginning of a meal; but the utterance has no semantic correctness conditions because it cannot be true or false.

There are both supporters and deniers of the idea that there might, in fact, be two different types of correctness that are relevant to meaning. Typically, those who disagree that the notion of correctness is ambiguous insist that it is impossible to distinguish linguistic from semantic correctness (Reiland, 2023). Indrek Reiland argues that the resistance to the idea that linguistic correctness is separate from semantic correctness comes from the implicit assumption that people can privately imbue words with meaning through their intentions (2023, pp. 2201–2202). The reasoning of deniers of the distinction can be summed up as follows: There is no such thing as linguistic incorrectness because utterances always mean whatever the speakers intend them to mean, and if they stray from publicly established norms for the usage of an expression this should always be interpreted as a type of linguistic innovation. In other words, "misuses" do not exist: If I say "good morning" to my husband right before we start eating dinner, what I am doing is not using the words "good morning" incorrectly but trying to introduce a new word from my personal

idiolect into English. The meaning of words in this idiolect is determined by my individual intentions.

The idea that meaning is imbued into words and sentences via speakers' intentions is certainly not a necessary background assumption for the discussion at hand. Consequently, there is no reason to assume that a distinction between linguistic and semantic correctness cannot be made. I will assume the distinction is intelligible and will spend the next section investigating whether linguistic correctness may exhibit the genuine normativity that would vindicate the normativity requirement in Kripkenstein's argument.

2.3 LINGUISTIC CORRECTNESS, USE-CONDITIONS, AND CONSTITUTIVE RULES

This section will examine linguistic correctness in more detail. I will argue that linguistic correctness understood in terms of use-conditions is constitutive of linguistic practice. Later, I will also argue that the fact that linguistic correctness is constitutive of linguistic practices can be used to account for Kripke's intuitions that we feel "guided" and "instructed" in our use of language. However, as we will see, this will not be enough to lend credence to the idea that meaning is normative in any genuine sense.

The idea I am defending here is the following: The use-conditions for a linguistic expression make the meaningful utterance of that expression possible, in the sense that I cannot even participate in the practice of language if I don't speak in accordance with the expression's use-conditions. The standard definition of constitutive rules characterises them as making new

types of behaviours possible (Searle, 1969/2011, p. 35). So, the use-conditions of an expression are constitutive of the conventional meaning of that linguistic expression.

Before continuing, several clarifications should be made. First, by "conventional meaning" I mean the standard, stable meaning of expressions, separate from contextual aspects or the individual intentions of speakers in a communicative situation. This is important because we want to be careful not to veer too far into the realm of pragmatics and away from meaning; we are investigating the normativity of *meaning*. While pragmatics may describe linguistic behaviour that is subject to norms, those are not the norms we are looking for. The conventional features of linguistic expressions are usually taken to be within the realm of semantics (see, e.g., Gutzmann, 2015, pp. 1–3; Kaplan, 1999, pp. 3–4; Korta & Perry, 2024; S. Kripke, 1977, p. 263). However, defining semantics as the study of conventional features of language is controversial, as is typical of distinctions drawn between semantics and pragmatics. Daniel Gutzmann notes, "*truth-conditionality* as a criterion for semantics does not characterise the same aspects of meaning as the criterion of *conventionality*" (2015, p. 3), and "the choice of whether we want to base semantics on conventions or on truth conditions leads to different sets of phenomena respectively falling within the scope of semantics" (2015, p. 5). To avoid going into too much detail about the semantics/pragmatics distinction, I will take it that it is safe to assume conventional aspects of language are part of semantics proper, and so have to do with (literal) meaning.

A second clarification concerns how linguistic correctness should be defined and understood. Reiland proposes a generic definition of linguistic correctness which may be adapted to different theories: using an expression in accordance with its meaning is using it while being in its *"use-conditions"* (2023, p. 2193). Use-conditions could take on the following preliminary form: "Saying *S* is linguistically correct when certain conditions are satisfied." In simple terms, this means that there will be occasions in which it is linguistically appropriate to use an expression, and

occasions in which it is not linguistically appropriate to use an expression. Reiland leaves use-conditions to be further defined.

One way use-conditions could be fleshed-out is through reliance on the new tools provided by use-conditional semantics. Semantics has sometimes been understood to exclusively study meaning as provided by a truth-conditional analysis. Any other meaningfulness found in language was posited as belonging to the domain of pragmatics. This basic criterion for distinguishing semantics from pragmatics has sometimes been represented as "pragmatics = meaning – truth-conditions" (Gazdar, 1979, p. 2). As has been mentioned, though, it has since been recognised that there are conventional aspects of meaning that go beyond truth-conditions. For example, "goodbye" is an expression that has a well-established conventional meaning, but whatever is expressed by "goodbye" is neither true nor false. It seems intuitive, then, that conventionally established meaning encompasses something more than purely truth-conditionally understood meaning.

The appearance of David Kaplan's unpublished paper titled "On the Meaning of 'Ouch' and 'Oops,'"[26] (1999) inspired several philosophers to work on use-conditional semantics, that is, on a semantics that can accommodate features of language that are determined by conventions of use and not purely by truth conditions (Gutzmann, 2015; Potts, 2004; Predelli, 2013). Kaplan's framework is designed to deal with expletives, indexicals, and other components of language which are unsuited to a truth-conditional analysis, but which have "a stable, conventional meaning, or perhaps better, a stable, conventional use; we say hello when we meet, goodbye when we part" (1999, p. 4). His proposal is that it makes sense to construct tools for a formal semantics that

[26] The transcript of Kaplan's lecture can be found at the following link as of 15/08/2025: https://eecoppock.info/PragmaticsSoSe2012/kaplan.pdf

includes some conventional aspects of meaning which are not analysable in truth-conditional terms:

> I now believe that by attending to rules of use – the right sort of rules of use – we can extend our formal semantics, and thus even our logic, to systematically account for the ignored semantic phenomena, and with surprising and, I hope, illuminating results. […] So, here's my method: I don't ask what the expression means, for example, I don't ask, "What does goodbye mean?" Instead I ask, what are the conditions under which the expression would be correctly or accurately used? (Kaplan, 1999, p. 3)

Kaplan, then, wanted to extend semantics to include considerations about rules of use. The use-conditions for some words, such as "goodbye" or "oops," intuitively provide us with information about the correct and incorrect ways of using them – and clearly, this is not *semantic* correctness, as there are no true or false utterances of "goodbye." It should be underlined, then, that use-conditions are a good candidate for what determines linguistic correctness.

Truth-conditions and use-conditions can coexist. Following Kaplan's basic idea, Gutzmann (2015) develops a "hybrid semantics" that includes both truth-conditions and use-conditions. The goal of Gutzmann's project is to build a framework that would enable us to apply the familiar tools of formal semantics even to non-descriptive, but still conventionally determined, features of language. In his framework, truth-conditions of propositions are based on sets of possible worlds in which a proposition is true. On the other hand, use-conditions are provided by the sets of contexts in which an expression is "felicitously" used (Gutzmann, 2015, p. 18). It is safe to say that felicitous usage can model what we have, up until now, referred to as linguistic correctness.

As was mentioned previously, the obligation to speak in a semantically correct way depends on contingent desires, such as the desire to be honest, for example. The notion of speaking in accordance with a term's use-conditions seems to be more significant, as it is inseparable from conventional meaning. It is difficult to conceive of a language without use-conditions: given a set

of use-conditions in my language, I ought to speak in accordance with them, if I want to speak at all. This suggests that the "oughts" derived from linguistic correctness are inherent to language in a way that the "oughts" derived from semantic correctness are not.

It might be helpful to rely on an example in order to clarify what this type of obligation may consist in and why it is different from an obligation to speak truthfully (that is, in a semantically correct way). If we take a non-referring expression such as "goodbye," it is clear that it has no truth-conditions. However, it is also clear that there are definite use-conditions which regulate its linguistically correct use: It is felicitous to say "goodbye" to people with whom we are parting, it is infelicitous to say "goodbye" when we're sitting down to eat. The use-conditions for "goodbye" provide us with something that is intimately tied with the conventional meaning of the word and with being able to use it in the English language. If I do not adhere to use-conditions, it can be doubtful that I am speaking at all and not merely making noises.[27] This lands credence to the idea that use-conditions, and in turn linguistic correctness, provide constitutive rules for the meaningful use of language. If violating the norms of linguistic correctness means that what I am doing does not count as speaking, this suggests that the norms of linguistic correctness are constitutive of the practice of making meaningful utterances (or, simply said, they are constitutive of speaking)[28].

---

[27] In comments to this text, Anandi Hattiangadi has challenged this by asking the following question: why would this be the case? Can't I speak infelicitously, yet still meaningfully? By saying "goodbye" when I've just arrived at a party, haven't I said something meaningful? My view is that in this case, nothing conventionally meaningful has been said. Use conditions delineate the border between appropriate and inappropriate uses of an expression within a linguistic community, and I believe that this fixes conventional meaning. When I say "goodbye" while arriving at a party, I might be in the process of making some sort of joke, which would be meaningful; I believe this muddles our intuitions. It's easier to see why infelicitous usage is meaningless in more extreme cases, such as someone saying "Gute Nacht" out of nowhere while having brunch with friends. If the use of an expression strays too far from what is considered felicitous usage, it can't count as being meaningful, similarly to how throwing a pawn across the room doesn't count as playing chess.

[28] I am using the word "speaking" for ease of exposition, but the same reasoning can be applied to all forms of linguistic tokens, including e.g. handwritten text, instances of sign language, code, and so on.

Obligations, even ones stemming from constitutive rules, can be violated. Violating our obligation to use language according to its use-conditions has some interesting parallels to violating the constitutive rules of games. While straying from use-conditions of a language may be a sign that we're not *speaking* it anymore, straying from central constitutive rules of a game is equally a sign that we're not *playing* it anymore; however, in both cases violations may be used to innovate or constitute a new practice. This suggests that the possibility of linguistic innovation does not interfere with the idea that use-conditions are constitutive of meaning, just like the possibility of innovation within the rules of a game does not interfere with the idea that games are constituted by their rules.

One might object to the idea that linguistic correctness is constitutive of meaning by arguing that the obligation to speak in a linguistically correct way is dependent on a desire to communicate or being understood. If one does not care to communicate or be understood, they have no obligation to speak in a linguistically correct manner; but it could be argued that they may still *speak*. And if one can speak without adhering to the norms of linguistic correctness, then linguistic correctness is not constitutive of meaning. In response to this objection, one could say that communication is widely accepted as being (at least) one of the primary functions of language. Because of this, it is difficult to conceive of participating in language in any way that precludes the desire to communicate something. When our aims are not to communicate or be understood, it is possible to question whether we are speaking at all. In other words, one can respond to the objection that without conforming to the norms of linguistic correctness, one does not speak: Any attempt to speak without adhering to these norms is just making language-like sounds. Desires related to such an integral function of language are not external to the practice. If you want your actions to count as playing chess, you should move the bishops diagonally across the board. If you want your actions to count as speaking, you should speak in a linguistically correct manner.

2.4 IS LINGUISTIC CORRECTNESS GENUINELY NORMATIVE?

As has been shown in the previous section, the idea that linguistic correctness is constitutive of speaking (that is, of making meaningful utterances) can be plausibly defended. However, this is not sufficient for our purposes. The question we set out to explore is the following: Are there any *categorical* meaning-norms? And could linguistic correctness be a source of these categorical meaning norms? Constitutive rules do not automatically give rise to categorical norms; the rules of chess are constitutive of chess, but I am not in any way categorically obligated to follow them. Just because I need to move bishops diagonally *in order to play chess* does not mean I categorically ought to move bishops diagonally, full stop. I may not want to play chess at all. In other words, the obligation to move bishops diagonally depends on the desire to play chess, which makes it a hypothetical norm, even though it is constitutive of chess. In a similar manner, one could say that the obligation to speak in a linguistically correct way depends on the desire to speak at all. If this is the case, the norms provided by linguistic correctness are hypothetical, even though they are constitutive of speaking.

Returning to the parallel between moral and semantic norms, some philosophers have suggested that in the case of categorical moral norms, they can be derived from facts about what is constitutive of being an agent. Given the analogies between semantic and moral norms that we have relied on until now, it may be illuminating to see whether there have been any successful attempts in the reduction of categorical *moral* norms to constitutive rules. David Enoch (2006) presents a survey of these attempts. He individuates several goals of the theories he covers, one of which is responding to Moore's Open Question Argument and reinstating the possibility of a naturalistic explanation of moral norms.

For our purposes, we may lay out the structure of the arguments deriving categorical oughts from constitutive rules surveyed by Enoch as follows: There are some essential normative features of agency,[29] in the loose sense that without these features, agency would not be possible. Since agency is constituted by these features, one cannot be an agent and avoid the normative force posed by them. So, the constitutive norms of agency are genuine norms (and not contingently dependent on our desires and interests). We have apparently derived genuine norms from constitutive rules.

"The Problem" with these kinds of argument, as Enoch calls it, is the following: Even if we individuated what is constitutive of agency, the normative force posed by those constitutive norms depends on whether a person is interested in being an agent (2006, pp. 177–180). If a person is uninterested in being an agent or acting, it seems as if the "oughts" derived from the constitutive rules of agency do not apply to them. In other words, the norms we derived from constitutive rules of agency are not categorical: they depend on a person's desire, and specifically, on their desire to be an agent.

Enoch considers several possible answers to "The Problem" that plagues attempts at deriving categorical norms from what is constitutive of agency, the most promising of which relies on the unavoidability of agency (2006, pp. 187–190). The idea is the following: What if it makes no sense to question someone's interest in being an agent because having interests is impossible if one is not an agent? There is something seemingly self-defeating about the idea that one could *choose* not to act. Enoch proposes a plausible counterargument to this idea. Even if we are somehow

29 The account of which exact features are the agency-constituting ones will vary – this is simply a placeholder – but the ones Enoch takes from existing literature as examples are the desire for self-knowledge, good self-constitution, and "motives and capacities constitutive of agency." Note the normative nature of each of these proposals.

forced into a practice, e.g., we are unable to avoid being an agent, this does not translate into having a categorical reason to follow the constitutive norms of that practice. There is still a need for an independent reason to act. He imagines a scenario in which we'd be forced to play chess – would we have a categorical obligation to follow the rules of chess in that case? Unless we have an independent reason to want to play chess, we have no obligation to play a certain way, even if playing is unavoidable[30] (Enoch, 2006, pp. 185–186). So, even if a practice is unavoidable, that does not mean we have categorical reasons to follow its constitutive rules.

We can now finally apply this reasoning to the constitutive rules of meaning. If I want to speak, I should follow the norms of linguistic correctness; but this normativity is hypothetical, that is, dependent on my desire to speak. If I am uninterested in being a speaker, there is no sense in which I ought to speak in a linguistically correct manner. I have no categorical obligation to speak in any kind of way. The constitutive norms of meaning are not categorical.

When applied to the norms of meaning, the "unavoidability" line of argument runs into the same difficulty outlined by Enoch. While the idea of agency being unavoidable has some merit, not being a speaker is more easily conceivable. While opting out of language altogether is rare, it does not seem to be impossible, and so we could not argue along the unavoidability line even if it were promising. However, as we have seen, even if a practice is unavoidable, that does not make its constitutive rules categorically normative. So, there is no way to derive categorical meaning

[30] One might object that if one is forced to play chess, that means that they are by definition following the constitutive rules of chess; breaking the rules of chess would automatically mean that the player wasn't playing and so would contradict the starting assumption of unavoidability. However, as Enoch notes when considering this issue, the fact that we have to follow the rules of chess does not mean we *ought to* or have reason to do so (2006, pp. 189–190).

norms from the constitutive rules of speaking, and more generally, we have little reason to believe that meaning is genuinely normative.

I would like to end this section by returning to the reasons why Kripke posed the normativity requirement in the first place. The original intuitions he leaned on were that there is a sense in which I *ought to* speak a certain way, and that I feel *guided* or *instructed* in my application of words I competently use. These intuitions have some force, and I believe they should be explained, even if we've concluded that meaning is not categorically normative. Fortunately, linguistic correctness presents a promising way of explaining these intuitions. When the sceptic asks why it feels as if I ought to use expressions a certain way, one can account for this intuition by explaining that linguistic correctness is constitutive of meaning broadly understood, and that *if* one wishes to speak, then they *ought* to speak in a linguistically correct manner. There is no categorical obligation to speak, but whenever we do wish to speak, we feel the force of the constitutive norms provided by linguistic correctness. We can return to the analogy with games: If one wishes to play chess, they ought to adhere to its rules. The norm here is hypothetical, but it is unavoidable if one wants to engage in the practice.

Some may find this way of explaining the "normativity of meaning" unsatisfactory because it places the source of our obligations not within standard truth-conditional meaning but within a broader, conventional meaning. In a sense, this objection is warranted: The basis for Kripkenstein's argument is that whatever facts determine the extensions of our terms should *also* account for the normative character of meaning, and that no theory is able to provide this kind of explanation. The same facts were supposed to account for both the normativity requirement and the extensionality requirement (see Section 1.2 for more detail; also, Kripke 1982, p. 11). I am proposing a strategy that separates these two requirements: The facts that account for the "normative character of meaning" may very well be different from the facts that determine the

extensions of our terms. However, it is important to note that the "normativity" derived from linguistic correctness is not genuine, categorical normativity, and that I am in effect arguing that Kripke's intuitions about feeling "guided" and "instructed" by *meaning* should not be taken at face value. What I propose, contra Kripke, is that we feel guided by certain norms of use that constitute our linguistic practices.

CONCLUSION

This chapter examined the scope and influence of the normativity requirement posed by the sceptic in Kripke's *WRPL*. The goal was to clarify 1) how the normativity requirement should be understood, 2) the way different interpretations of the normativity requirement affect the sceptical argument, and 3) whether there are any genuine meaning norms. Much of the chapter has been dedicated to the third point, and in particular to the exploration of linguistic correctness as a possible avenue for understanding the normativity of meaning. I have argued that even though it is plausible that linguistic correctness is constitutive of our linguistic practices, there is no way to derive categorical meaning norms from constitutive rules. The conclusion is that there is no good reason to believe that there are categorical, non-hypothetical meaning norms. Meaning is not (genuinely) normative.

How does this affect the sceptical argument? As was described at the beginning of this chapter, the normativity requirement of the sceptical argument can be interpreted in two ways: as assuming that meaning is categorically normative, or as assuming that meaning is normative in the more superfluous sense provided by the existence of semantic correctness conditions. The sceptical argument is certainly stronger if one can show that meaning is genuinely normative. To

recall, if meaning is categorically normative, Kripkenstein's argument can be interpreted as an analogue to arguments in support of moral antirealism. If meaning is not categorically normative, as was concluded in the section immediately preceding this one, the sceptical argument is weaker: it merely demonstrates that no extant theory of content is successful in individuating which facts determine the correctness conditions of a representation, that is, which facts determine its extension. This means that there is significant hope for the sceptic's opponent, as this weaker argument gives us no reason to believe that new theories must encounter the same difficulties as the ones considered in *WRPL*. The sceptical argument is not *a priori*, and it can be defeated by explaining how representations have correctness conditions (in a non-circular way).

The key takeaways of this chapter are threefold: first, meaning is not genuinely normative, in the sense that there are no categorical norms to be found either in our linguistic practices or stemming from abstract, truth-conditional meaning. Consequently, it is charitable to avoid interpreting Kripkenstein's argument and its normativity requirement as assuming that meaning is genuinely normative. Second, given the fact that meaning is not genuinely normative, Kripkenstein's argument is best understood in the way Boghossian originally did: The sceptical challenge boils down to explaining how anything could possess (semantic) correctness conditions (1989, p. 515). This means that the sceptical challenge should be interpreted as a difficulty in finding the facts in virtue of which something has semantic correctness conditions. Lastly, the intuitions we (and Kripke) have about being "guided" and "justified" in our use of language can be explained by linguistic correctness better than by semantic correctness. The rest of this thesis is dedicated to considering whether some new theories have, in fact, successfully resolved Kripkenstein's paradox as it has been interpreted here.

3. CAUSAL THEORIES OF REFERENCE

INTRODUCTION

One of the most prominent debates in the philosophy of language in the 20th century has concerned the workings of reference. Reference is a relation that is usually said to obtain between words and things; it is what explains how words can represent something. Words refer to their referents. For example, when I utter "Saul Kripke wrote *Naming and Necessity,*" the referent of the name "Saul Kripke" is the man called Saul Kripke. Having recognised that words *do* refer, the question becomes: How do words get their reference? In virtue of which facts does the reference relation between words and referents obtain?

The workings of reference are relevant to the central goal of this thesis, which is to examine how representational contents could be determined without falling victim to the Kripkensteinian paradox. The standard semantic realist picture takes (roughly) this shape: Sentences are meaningful in virtue of their truth-conditions, and they have truth-conditions in virtue of the referents of the words they contain (and their syntactic properties). In Michael Devitt's words, "philosophers often are interested in reference because they take it to be the core of meaning" (1998). If a theory could successfully explain the workings of reference, that would provide a partial answer to the CDQ,[31] because we'd be in the position to determine what makes it the case that words refer to – and so *represent* – things.

---

31 CDQ stands for "Content Determination Question," as was established in the introduction.

A central point of contention in the debate about reference is whether it is always mediated by descriptive content or whether it can also be direct, that is, unmediated by any descriptive content. If reference is mediated by descriptive content, that means that we always refer by associating certain descriptions with referring terms; the associated description is what determines reference. I will not engage with theories that posit that reference is always mediated by descriptive content here.

Theories of unmediated reference – *direct* theories of reference – might prove resistant to the sceptical challenge. The basic intuition supporting the possibility of direct reference is that we can successfully use proper names and natural kind terms to refer even if we have very limited, or even false, beliefs about the objects or kinds in question. For example, I may mistakenly believe that Albert Einstein invented the nuclear bomb; and I may even have no other beliefs associated with the name Albert Einstein (this example is taken from Kripke, 1980/2001, p. 85). My ignorance about Einstein does not seem to prevent me from referring to him – I am still talking *about* Einstein. In other words, even when the appropriate descriptive content is missing, words can still refer.

This chapter will be dedicated to examining whether direct causal theories of reference can meet the sceptical challenge. To reiterate, the challenge was to individuate the facts in virtue of which speakers mean something by a sign, and to provide an answer that respects the three requirements: extensionality, normativity (understood as the existence of correctness conditions), and non-circularity.

There are several reasons one should be interested in causal theories, and in particular direct causal theories of reference, as possible candidates for this task. First, their reductive approach to semantic phenomena means that they aim to explain meaning in terms of physical facts, which would make them compliant with the non-circularity requirement. If unmediated

reference were possible, we'd have a straight and non-circular answer for the sceptic: Speakers mean something by a sign (they *refer to something*) in virtue of certain causal relations. Secondly, Kripke himself – who likely thought the sceptical paradox had some merit – was a defender and pioneer of the causal-historical theory of reference. Curiously, he never explicitly addressed how the sceptical paradox interplays with his own theory of reference. Other philosophers have made that connection later (e.g. Maddy, 1984; R. B. Miller, 1992).

This chapter will be structured as follows. Section 3.1 briefly overviews the basic structure and goals of causal theories of reference. In Section 3.2, I will address a serious obstacle theories of unmediated reference face: the qua problem. The qua problem highlights that purely causal theories struggle with reference indeterminacy. I overview and critique Penelope Maddy's attempt at resolving Kripkenstein's paradox by using causal theories of reference. I conclude that Maddy's work suffers from issues that are characteristic of causal theories of *mental* content, which will be considered in detail in Chapter 4. Section 3.3 contains an overview of Max Deutsch's (2023) recent promising attempt at resolving the qua problem; he proposes that causal relevance is the key to how reference gets fixed during baptisms. In Section 3.4, I argue that Deutsch's solution to the qua problem ultimately fails because causal mechanisms alone are simply insufficient for the purpose of singling out determinate reference. I conclude that while purely causal theories of reference have no fundamental issues that prevent them from resolving the sceptical challenge, they have not yet resolved it.

3.1 CAUSAL THEORIES OF REFERENCE

In *Naming and Necessity,* among other things, Kripke argues for a causal-historical theory of reference, wherein speakers can refer to objects or kinds regardless of the descriptions they associate with such objects or kinds. The story goes as follows: Reference is initially fixed by an act of dubbing in the presence of an object or kind. Afterwards, it is passed on via a causal-historical chain that obtains between speakers. To put it somewhat reductively, there is a history of causally linked events that connect the act of naming an infant "Saul" and my utterance of that name 80-odd years later, and this is what allows me to refer to the philosopher who wrote *Naming and Necessity* (Kripke, 2001, pp. 91–92). What I believe about Kripke plays no role at all in my successful referring to the man.

Two mechanisms, then, are central to the workings of reference for the causal-historical theory: "baptisms", i.e., the initial fixing of reference via a dubbing act, and "reference borrowing", that is, the passing on of reference between speakers through communication. More will be said about the details of these processes later in this chapter. At this introductory point, it is only important to note that neither the introduction nor the borrowing of reference require that the referent *satisfy* any description present in speakers' minds. Kripke believes that reference does not (usually) obtain in virtue of things going on in speakers' minds, but in virtue of the existence of a certain kind of chain of communication (2001, p. 93). However, Kripke does say that the borrower needs to *intend to* refer to the same thing as the speaker who is lending the reference to successfully refer, meaning that reference borrowing cannot happen without a certain type of intention being present (2001, p. 96).

Regardless of what Kripke believed when delineating his own version of the theory, the causal-historical theory of reference appeared in a philosophical context that was strongly

physicalist, and that context influenced its subsequent development. Given the rise and success of the natural sciences in the 20th century, this attitude is unsurprising. Much of analytic philosophy endorsed the notion that, fundamentally, there is nothing "over and above" the physical world examined by natural sciences. Content, whether of mental states or of linguistic expressions, was (and still is) one of the phenomena that are most difficult to subsume under the physicalist approach. Because of this, the appearance of the causal-historical theory of reference provided a promising strategy for those who wanted to naturalise semantic phenomena.

The goal of naturalising content is related to the resolution of the sceptical challenge. To reiterate, the sceptical challenge consists in providing a non-circular explanation of how representational content is determined. In Chapter 1, we put this goal in terms of individuating the facts in virtue of which speakers mean something by a sign; in other words, individuating meaning-determining facts. A theory that succeeds in naturalising content will, by definition, explain how representational content is determined non-circularly, because it will individuate *natural* meaning-determining facts. So, a successful naturalisation of content could resolve the sceptical paradox.

The basic idea behind how causal-historical theories of reference can provide a physicalist picture of meaning is the following. Speakers can "baptise", i.e., fix the reference of a word, without associating any descriptive mental content with the object or kind being baptised. What is required for a speaker to refer is the obtaining of a certain type of causal contact between a speaker and an object or a kind (or a sample of a kind, to be precise). Causal relations are often assumed to be natural and explainable in purely physical terms. Given all of this, we have sketched a naturalistic story for the way in which at least some of our words come to have their reference.

The rest of this chapter is dedicated to considering whether this sketch can be completed and transformed into a theory of representational content that is paradox-proof. I will argue that

all attempts at providing a direct causal theory of reference are either circular – they sneak semantic elements in at some point – or they do not yield determinate reference.

3.2 THE QUA PROBLEM FOR CAUSAL-HISTORICAL THEORIES OF REFERENCE[32]

Causal theories of reference became a popular alternative to descriptivist theories of reference in the second half of the 20th century for a variety of reasons. However, whether it is actually possible for speakers to successfully refer without *any* descriptive content in mind is not clear. Kripke seems to imply that no such descriptive content is necessary – after all, we can be radically wrong about the object or kind we're talking about. For example, in a distant future we may find out that cats are not animals, or even material beings; we still successfully refer to them *now*, even though we think they are animals. During a baptism, it is the causal contact between speaker and cat-sample, along with the speaker saying "I baptise *this "*cat"" (or something similar), that enables us to denote cats with "cat." Our beliefs about cats do not seem to play a role in fixing reference.

If we maintain that reference can be fixed purely by causal contact with an object at baptism, the question that immediately arises is: Causal contact *with what?* As it turns out, ostensive acts such as the ones used during baptisms – "I baptise *this* "cat"" – do not automatically single out a referent, since there are usually many, many things in front of a speaker during any act of

[32] This section, along with Sections 3.3 and 3.4, borrows from a paper published in *Erkenntnis* (Papic, 2024).

baptism. This fact poses a significant obstacle to the possibility of purely causally determined reference. This weak spot in the causal-historical theory of reference is known as the *qua problem.* Briefly, the qua problem for causal theories of reference is this: What makes it the case that a baptism fixes reference in virtue of causal contact with one object/property and not any of the other objects/properties present during the act? Why should the reference of "cat" be grounded in the object in front of me *qua* cat and not, for example, *qua* mammal?

Resolving the qua problem has proved to be anything but trivial. As was implied previously, the issue stems from the proposed mechanism for reference fixing. The first mention of the qua problem by name can be found in Kim Sterelny's article "Natural Kind Terms" (1983); however, Kripke attributes (and dismisses) similar concerns to Peter Geach (1954).[33] The qua problem affects reference-fixing for both proper names and natural kind terms, which represent the paradigmatic cases in which reference is said to be fixed causally.[34] I will now outline how the qua problem arises for both kinds of cases.

The qua problem for proper names arises in the following way. When a speaker attempts to introduce a new proper name, it is unclear how the mere utterance of a name during an act of baptism fixes the reference of that name to an object *as a whole* and not, for example, in its momentary time-slice, or in the object along with a bubble of air surrounding it, or in an

33 "According to Geach, since any act of pointing is ambiguous, someone who baptizes an object by pointing to it must apply a sortal property to disambiguate his reference and to ensure correct criteria of identity over time-for example, someone who assigns a reference to 'Nixon' by pointing to him must say, 'I use "Nixon" as a name of that man', thus removing his hearer's temptations to take him to be pointing to a nose or a time-slice." (1980, p. 115, footnote 58. Italics mine.)

34 The qua problem also afflicts the fixing of reference of non-natural kind terms, if one believes that direct reference fixing is possible in those cases. However, since this possibility is controversial and so does not represent causal theories at their strongest, I have chosen to focus on natural kind terms only. Regardless, the qua problem arises in a completely analogous manner for non-natural kind terms as well.

undetached part of the object, or in one of the kinds the object is a sample of, or in anything else the baptiser might have had direct causal contact with at the moment of naming. Simply put, regardless of the details of our metaphysics of choice, speakers are in causal contact with many things during any single baptism, and it is unclear which one of these things fixes the reference of a newly introduced term; reference is left indeterminate. So, if direct (purely causal) reference is possible for proper names, the following must obtain:

(I). The act of baptism must identify the referent not *qua* kind but *qua* object; and

(II). The act of baptism must identify the referent as a single, appropriate object and not as anything else.

The temptation one has is to say that the baptiser needs to have some sort of description in mind to uniquely fix reference; for example, they would have to say, "I dub this "Saul"" while *having an intention to name the person in front of them* to successfully name the baby "Saul." But this would negate the effort to demonstrate that there are cases of direct reference, since we're (once again) supposing reference must be mediated by a description; making this compromise soils the supposedly purely causal mechanism of reference fixing.

The qua problem for natural kind terms presents as follows. In the case of natural kind terms, purely causal theories propose that reference can be fixed through causal contact between a baptising speaker and a sample of a natural kind. For example, the reference of the kind "tiger" is fixed via contact with an individual tiger (Devitt & Sterelny, 1987/1999, p. 88). The qua problem presents itself in two ways: First, it is unclear why causal contact with an individual fixes the reference of the newly introduced term in a *kind* and not in the individual itself. So, when the word "tiger" is introduced via baptism – "I will name *this* "tiger"" – it is unclear what makes it the case that the speaker has named a kind and not provided a proper name for an individual animal (an inverse and analogous problem can be applied to proper names as well – the name "Saul" used in

an earlier example might as well have been a newly introduced kind term). Second, most things are instances of many natural kinds at once. For example, the individual tiger in front of the speaker is a tiger, but is also a feline, a mammal, a vertebrate, and an animal (among many other things). Given any single thing's membership in many natural kinds, it remains unclear *which of many kinds* is named via the baptism; reference is left indeterminate. So, if direct reference is possible for kind terms, the following must obtain:

(I). The act of baptism must identify the referent not as an object but as a kind; and

(II). The act of baptism must identify the referent as a single, appropriate kind and not as anything else.

In their discussion of the *qua* problem, Michael Devitt and Kim Sterelny ultimately conclude that reference cannot be completely direct and that a "hybrid theory," something in-between a descriptive and a purely causal theory, is needed to resolve it (Devitt and Sterelny 1999 p. 91). In other words, they believe that reference must be mediated by descriptions present in the speaker's mind: This introduces intentional content as the determiner of reference and precludes any chance we had of resolving the sceptical paradox, since it means that the theory is not able to meet the non-circularity requirement.

Devitt and Sterelny are far from convinced that introducing descriptive content in speakers' minds to fix the qua problem resolves all issues, particularly when it comes to natural kind terms. For example, there are historical cases in which there have been significant shifts in the descriptions associated with kind terms due to scientific discoveries – for example, the word "light." The first person to introduce the term "light" likely had something completely different in mind than what we know about the properties of light today; it is unclear whether the descriptions they had in mind when introducing the term were satisfied by anything in front of them. Yet, we seem to think that people today and people in the 17th century were talking about the same thing

when using the word "light." This intuition gives support to the idea that even the simplest descriptors present in baptisers' minds may turn out to be irrelevant for the fixing of reference.

More importantly, Devitt and Sterelny notice that reference can't *ultimately* be grounded by descriptions because the content of descriptions requires further grounding (1999, pp. 60, 93). Descriptive theories "pass the referential buck:" they explain reference in terms of other referring entities. This leads them to assume that there must be some basic, fundamental expressions that do not require descriptions for their reference to be fixed. However, they do not specify how these fundamental expressions themselves can obtain their reference while avoiding the qua problem.

Richard B. Miller poses the qua problem in a slightly different way (1992). Miller says that the issue arises because during the act of baptising, one can only have causal contact with an individual thing; and that is a problem because kind terms always refer to more than one is ever in causal contact with (1992, p. 426). In essence, a singular causal contact between one speaker and one object strains to explain the generality of kind terms. It is important to notice that this is not how Sterelny and Devitt described the issue; first of all, as Miller has defined it, the problem cannot be applied to proper names, which Sterelny and Devitt do.[35] However, it could be argued that the original grouping of cases involving proper names and cases involving kind terms is misleading, as they present significant differences, which could in turn justify distinguishing between the two types of qua problem: one for proper names and one for kind terms.

---

35 When he first defines the qua problem, Miller says "referring expressions almost always refer to more than what is actually present to any speaker"(1992, p. 426). This could be stretched to apply to proper names as well – at the baptismal act of a person, we want to refer to the individual as a whole and not only to the time slice that we're in perceptual contact with. However, this description of the qua problem does not capture the spirit of what Devitt and Sterelny were saying. At baptisms, we are also in the presence of undetached parts of a person, and of a myriad of other mereologically gerrymandered objects which are intuitively not designated by the proper name. This has nothing to do with the fact that the name needs to refer to something that is not present to us at that moment. We are, in fact, in actual and present contact with way too many things.

But even if we focus only on the qua problem as it relates to kind terms, Miller seems to include assumptions that are missing in earlier interpretations of the problem. Devitt and Sterelny only assume that there is an issue in grounding (or fixing) the reference in one kind rather than in another whenever an object is in fact the member of multiple kinds at the same time (which is nearly always the case). They do not go on to say that this happens *because* of the generality of those terms, or because we're not in contact with everything those terms should apply to. This is an additional assumption that is not necessary for the qua problem to be formulated.

The qua problem bears some similarities to Kripke's sceptical paradox. The shared issue is the following: There is a difficulty in univocally determining what speakers mean by a sign. In the case of the rule-following paradox, all possible theories of representational content are targeted; the qua problem, on the other hand, afflicts causal theories of reference. In *WRPL*, it is quickly uncovered that any attempts at determining what a speaker means by a sign via additional intentional content – be it descriptions the speaker associates with the sign, beliefs, or past intentions to use the sign a certain way – are bound to fail, as they simply shift the sceptic's target without providing a definitive solution. This explains why Devitt and Sterelny's hybrid solution to the qua problem, which relies on descriptions present in baptisers' minds at the time of reference fixing, cannot avoid the issue. For the qua problem to be definitively resolved, there must be a convincing proposal delineating how reference can (at least sometimes) be fixed without invoking additional intentional content.

Kripke never considers a causal solution to the sceptical paradox. Direct causal theories should have the immediate advantage of not requiring intentional states for the determination of reference. The reason for Kripke's omission could be that he felt it was evident that a causal theory would immediately run into the issues highlighted by the qua problem (even if the naming of the problem succeeded the 1982 publication of *Wittgenstein on Rules and Private Language*); or he could

simply have believed that it is impossible to exclude intentions' role in the fixing of reference. The fact that he requires an intention *to* baptise and an intention *to* use a word in a certain way when borrowing occurs points towards the latter.[36]

An early instance of proposing causal theories of reference as a solution to Kripkenstein's paradox can be found in Penelope Maddy's paper, "How the Causal Theorist Follows a Rule" (1984). Maddy explicitly investigates whether causal theories of reference can resolve the sceptical paradox presented in *WRPL*. After considering the problems posed by the sceptical argument, she argues that the resulting indeterminacy presents itself in all theories that posit the referent of an expression as satisfying a requirement, usually specified by a sense or a description of some sort (Maddy, 1984, p. 464). She notices that the causal theory of reference does not assume that senses or descriptions are needed for reference, and because of this, she wants to assess whether it can resolve the sceptical paradox.

Maddy uses the terminology of "baptisms" and "reference borrowing." She acknowledges that baptisms as they have been described until now seem to be insufficient to pick out a unique referent (Maddy, 1984, p. 464). An act of baptism is ostensive, in the sense that the imagined act involves a sort of "pointing" and the idea that the newly given name applies to everything that is "like this." How could the new referent be determined without relying on descriptions or intentional states? Maddy separates the issue into two stages: the determination of what "this" in

---

[36] Andrea Bianchi suggests that Kripke does not consider a causal solution because it seems hopelessly implausible for "plus," which is the example with which he introduces the sceptical paradox. However, while Kripke acknowledges that the paradox is most easily described using the word "plus," he insists that the problems posed by the sceptic "apply throughout language and are not confined to mathematical examples" (1982/1995, p. 19); he mentions that the paradox can be constructed for the words "count" (p. 16), "table" (p. 19), "red" and "color" (p. 20), "cube" p. (42). If Kripke did believe the causal solution could work for some sections of language, e.g. kind terms and/or proper names, it would have been an odd choice not to mention it. Regardless, it is odd that he does not mention causal theories in *WRPL*.

"like this" is; and the determination of what "like" in "like this" is (1984, p. 464). The first issue concerns the individuation of the sample *as an individual sample of x*, while the second issue concerns what determines the connection between samples of x that make them *like* each other. This is the point at which Maddy engages with the issues posed by the qua problem, though she does not name it as such.

For the first stage of her answer, Maddy suggests that neurological states of the baptiser could be the thing that determines what they are in causal contact with. Neurological states are not intentional and would, as such, be immune to the sceptic's attacks. We could speculate that people's ability to perceive develops over time through the development of neural structures called "cell assemblies." Maddy imagines this process is gradual, with people first developing cell assemblies for simpler features of objects and then integrating those features into a larger and more complex cell assembly which allows them to perceive, e.g., triangles *as* triangles.[37] The development of these neural structures is what presumably allows us to see things as units and not as chaotic sensory information. It also allows us to perceive similarities between objects of perception, to remember them, and so on (Maddy, 1984, p. 465).

Given this initial sketch about the inner workings of perceptual mechanisms, Maddy believes we can give an answer to what individuates the sample we are pointing towards during a baptismal act. When the baptiser says "gold" in the presence of a sample of gold, we can differentiate this act of reference fixing from another that picks out a different sample because contact with gold stimulates a different neural cell assembly than contact with yellow things or

37 The information available to Maddy in 1984 is likely outdated and should be compared to newer neurophysiological findings related to perception, but the basic assumption – that our ability to perceive is based in physical substrates that develop through stimulation and experience – can be preliminarily granted and adapted as needed.

metal things. If contact with gold consistently stimulates a corresponding neural cell assembly (and not others), we can use this fact to claim that the sample picked out by the baptismal act is determinate (Maddy, 1984, p. 465). In other words, Maddy believes that the existence of a certain pattern in the neural state of a speaker is good evidence that the baptiser denotes "gold" and not "metal" or "yellow" during the act of baptising (which would hopefully be associated with different neural patterns). We do not need to be aware of such mechanisms for them to exist and, ultimately, for them to determine which sample is picked out during a baptism. To sum up, Maddy's idea is that we can determine what the target of a baptism is by determining what the baptiser perceives; and what they perceive is a matter of neurological fact. So, we can individuate the target of a baptism through facts about the baptiser's neurological state.

Maddy then goes on to analyse the "like" portion of the issue: What does it mean when the baptiser picks out a sample and fixes the meaning of a new term to whatever is "like" it? A natural answer would be to say that every sample that stimulates the same cell assemblies as the ones stimulated by the initial contact at baptism is what determines which things are *like* the initial sample, and, ultimately, what the extension of the referring term is (Maddy, 1984, p. 469). However, this leads to several counterintuitive consequences. First, there will be cases in which samples of a kind do not stimulate the correlated cell assemblies, e.g., in cases where it is difficult to perceive the sample at all – a horse in the pitch-black darkness will not be perceived as a horse; but we still want it to be in the extension of the word "horse." Similarly, people often misperceive. Far away cows may look like horses, which only means that the object in question – a cow – stimulated the neural patterns usually associated with horses; but we do not want any cows in the extension of the word "horse." In other words, we know that mistakes in perception are possible; but if it is my perception that determines the extension of terms, there is no way to explain the possibility of mistakes in sorting (Maddy, 1984, pp. 469–470).

The second objection is related to the first and, ultimately, to the possibility of misrepresentation – a speaker's neural state cannot tell us how a word *should* be used, just how it *will* be used. This echoes the issues related to the normativity requirement, which were tackled in Chapter 2; to reiterate, Kripke believes that a theory must not predict how a speaker *will* use a term, but how they *should* use a term (1982/1995, p. 37). In Chapter 2, the problem was framed as a difficulty in demonstrating how anything could have semantic correctness conditions, that is, conditions in which a representation is semantically correct (true) and conditions in which a representation is semantically incorrect (false). While the focus of Chapter 2 was on linguistic expressions, the same idea can be easily applied to mental content: Just like utterances, mental representations such as perception and belief have correctness conditions, and those correctness conditions should be understood in terms of truth. Because of this, we can say that perceptions should have semantic correctness conditions, that is, a theory should be able to account for the fact that one can correctly or incorrectly perceive something. Since Maddy relies on perception to ground reference, and her understanding of perception does not provide ways of accounting for misperception – conditions under which a representation is semantically incorrect – there is no space for correctness conditions in her theory. And, as Boghossian succinctly said, "having a meaning is essentially a matter of possessing a correctness condition. And the sceptical challenge is to explain how anything could possess *that*" (1989, p. 515).

Because of this issue with explaining misrepresentation, Maddy is pushed to introduce an external factor that contributes to how reference is fixed. "The [cell] assembly is a fallible detector of phenomenal similarity, and the phenomenal similarity is a fallible detector of membership in the kind"; the fallibility here is meant to leave room for misrepresentation (1984, p. 471). But *which* kind does this neurological substrate detect? To explain away the undesirable, deviant kinds, she commits to a "ready-made world", i.e., to the existence of objects and kinds that are natural, that figure in ultimate scientific explanations, and are objective. Samples of gold, then, are *like* each

other in virtue of having certain objective natural properties. Maddy uses the notion of naturalness in this sense to explain what is perceived, and, ultimately, what fixes reference (Maddy, 1984, p. 471). It will be a cat and not an undetached part of the cat that figures in the best scientific explanation of the perceptual experience of the baptist, because the property of being a cat is more natural than the property of being an undetached cat part. Similarly, the property of being a cat figures in the best scientific explanation of my perceptual capacity to recognise cats (whether I am aware of this or not). This is enough, Maddy states, to fix the reference of "cat" in cats.

We can sum up Maddy's story of how reference is fixed as follows:[38] During a baptism, the baptiser perceives something. Our best scientific theories say something about the neural state the baptiser is in when they perceive something. This neural state is caused by a determinate thing, which is a member of a natural kind, and *that* natural kind is what determines the reference of the newly introduced term. In other words, it is certain facts about the baptiser's neural states, coupled with certain facts about the object causing those states baptiser, that initially fix the reference of a term.

If we grant that some properties and individuals are metaphysically more natural than others, does Maddy's story work? There are two dimensions to consider here: whether her story works as an answer to the qua problem, and whether her story works as an answer to Kripkenstein's paradox. As an answer to the qua problem, Maddy's proposal does work to some degree, given we grant two big assumptions are true: the claim that some properties and individuals are more natural than others; and that there are neural mechanisms that work as detectors of those natural properties and individuals. The first is a theoretical assumption, while the latter is an (at

[38] I am here referring to Maddy's proposal as a "story" or "picture" because she is explicit about it being a sketch of a more detailed solution to be developed in the future.

least partly) empirical one. Given the truth of these two assumptions, Maddy's theoretical structure yields determinate reference. However, this answer is too restrictive: If the most natural property detected by the speaker's neural structure fixes a new term's reference, it is going to be difficult to explain how anything other than the most natural (detectable) property or individual present at baptism is ever named. So, while this setup yields determinate reference, it is unclear that it delivers the appropriate extensions for our terms.

More importantly, there are several vulnerable points in Maddy's story when it is taken as an answer to Kripkenstein's paradox. The most pressing issue Maddy's theory has is that her account seemingly "passes the referential buck" – it relies on the content of perceptual states to determine reference. In other words, it does not meet the non-circularity requirement: the theory grounds content in more content. Maddy is careful not to talk about perception directly and opts for explaining the content of perception via a presumed underlying neural substrate that "identifies" what is perceived; she hopes in the reduction of the perceptual to the neural. However, she uses covertly intentional language when talking about neural substrates: "A cell assembly, understood as a *detecting device*, can be seen as *measuring* several equally natural kinds" (Maddy, 1984, note 24, italics mine). If the neural substrate is already taken as a "detecting device" that "measures," we are assuming it can sort instances of what it is measuring; "detect" is synonymous with "discern." And like all detecting devices, it represents something: the thing it detects; and it can make mistakes in its detection: it can misrepresent. What this means is that in explaining the representational capacities of referring terms, Maddy has shifted the burden to the representational capacities of the neurological substrates that underlie perception, which themselves remain mysterious. In virtue of what does a neural substrate detect (represent) gold and not something that looks exactly like gold? And in virtue of what does a neural substrate detect (represent) at all? These are questions for theories of mental content, which will be examined in some detail in

Chapter 4. The thing to note here is that grounding referential content in mental content does not resolve the challenge posed by Kripkenstein by itself: it just shifts the explanatory burden.

Another problem for Maddy's theory, which it shares with many other causation-centric theories of mental content, is the distality problem, which puts pressure on uniquely identifying the (one) desired cause of a neural state from the long chain of causes that leads to that neural state occurring (Schulte, 2023, pp. 22–23). Maddy defined the target of a baptism as the thing that causes a certain neural response. But any neural response is going to have many causes beyond the sample in front of the speaker, both more distal and more proximal; for example, the neural response is caused by the sample itself, but also by the light waves reflected by the sample and traveling through space and towards the speaker's eyes, and by the retinal image created by those light waves, and by signals traveling from the sensory apparatus towards the central nervous system and so on. Which one of these causes is the target of a baptism? The theory outlined by Maddy does not have the tools to answer this question. We will return to the problem of distality in more detail in Chapter 4, where it will be discussed in the context of more recent theories of mental content which engaged with it directly.

The assumption that our capacity for perception is what grounds reference is in tension with some externalist intuitions. Our cell assemblies (or any other supposed neural mechanism that underlies perceptual recognition) are surely not sensitive to all natural kinds. We do not need to be *aware* of our capacity to recognise or differentiate kinds, but even so, there is no justification for presupposing that our neural structures reflect the natural world in any significant way. The reference of our kind-terms seems to be determined by the actual thing we run into, and not by our perceptual machinery. While Maddy uses naturalness to account for the role of the world in determining reference, her setup still initially relies on the idea that there are neural substrates that underlie a process of recognition; and it is this process that singles out what a speaker is baptising.

This is an issue because we might want reference to be fixed in things we are not able to reliably recognise (or which our neural substrates are not able to reliably detect). These are, again, issues and considerations that would be more relevant within a discussion of causation-centric theories of *mental* content. For the purposes of this chapter, we can conclude that Maddy's theory has some difficulties resolving the Kripkensteinian paradox due to the issues listed above.

3.3 MAX DEUTSCH'S SOLUTION TO THE QUA PROBLEM

If one does not wish to ground purely causally determined reference in perceptual capacities or their neural correlates, there are alternatives. Such an alternative is developed by Max Deutsch, who argues that there is no qua problem for purely causal theories of reference in the first place (2023). This section outlines the way in which Deutsch proposes that reference is fixed purely causally. His argument is the following: We usually agree that things can cause effects *qua* one of their properties and not *qua* others. For example, a kettle boils water due to its heat and not due to its colour. Given this assumption, we can simply say that an individual object caused a dubbing act *qua* the relevant property (heat) and not *qua* the irrelevant properties (colour etc.), thus fixing the reference of the term. The qua problem has apparently vanished. Notice that there is no mention of mental content in this story: only causal relations are invoked.

Deutsch explicitly takes inspiration from Miller's paper on the qua problem (1992). Miller argues that causal contact with certain kinds can have an unmediated effect on our capacities to recognise members *of that kind,* and that we gain this ability *due to that kind* and not due to other kinds with which we may have been in contact with at the same time. The development of these recognitional capacities – whether we are aware of them or not – can ground the reference of our

kind terms. It is not difficult to see that Miller's proposal will have issues similar to the ones that emerged from the discussion of Maddy's paper. However, Deutsch does not follow Miller in basing his argument on recognitional capacities, or on anything else about the speaker's psychological or neurological state. Deutsch limits himself to borrowing the insight that we have no issue admitting that some things can have an effect in virtue of some of their properties and not in virtue of other properties. Deutsch argues that this characteristic of causal relations, called causal relevance, is enough to see that the qua problem is not a problem at all: it doesn't even emerge in the first place.

I will argue that, unfortunately, the qua problem persists. It can neither be dissolved, as Deutsch proposes, nor solved, as Miller hopes. Causal mechanisms, and in particular the facts about causal relevance appealed to by Deutsch, are insufficient to dissolve the qua problem. This is because it is not generally the case that a *unique* property is causally relevant in purely physical cases of causation; but this is precisely what is required to avoid the referential indeterminacy highlighted by the qua problem.

To reiterate, the qua problem arises in connection to the proposed mechanism for causal reference grounding, namely the "baptism" or "dubbing" event at which a new object or property is named for the first time. If reference is grounded by causal contact with something, there is a difficulty in specifying *in virtue of what* a new term is grounded in a specific property or object rather than in another, since many properties and objects are always present at baptism. Deutsch engages with the qua problem as it concerns kind terms. He argues that because causes bring about their effects in virtue of some properties and not others, reference can be grounded in a specific property in virtue of the dubbing act being caused by that very property; and a dubbing act is caused by a specific property because that's just how causation works. There is no mystery here,

just like there is no mystery in the case where water boils due to the heat of the kettle and not due to its colour.

If the qua problem can truly be dissolved, causal theories come much closer to being a viable response to the sceptical challenge. If reference is grounded purely in causal mechanisms in *some* cases, we will have an answer to the CDQ in those cases: The content of referential expressions is determined by a certain kind of causal contact at baptism. This kind of answer to the CDQ is not circular, as it does not rely on facts about additional contentful states in its answer to how extensions are determined. It also does not seem to preclude the possibility of misrepresentation. Once the reference of, e.g., "gold" is fixed via causal contact with a gold-sample, speakers can misapply the word "gold" by saying of things that are not gold that they are gold. The theory has no difficulty accommodating this fact.

How does Deutsch purport to demonstrate that causal mechanisms are sufficient to determinately ground reference? Deutsch relies on some of the general characteristics of causal relations and proposes that these general characteristics are sufficient to explain the qua problem away. The example he uses is that of hot acid being poured in a glass beaker, causing the beaker to break. In this case, it seems unproblematic that the liquid broke the glass *qua* hot thing and not *qua* acidic thing (2023, p. 1811).[39] Now imagine a dubbing event of some sort: Adam is walking through a forest and sees a strange large creature. He says, "I dub this an 'elephant.'" In virtue of what is the newly introduced term's reference grounded in the object *qua* elephant and not, for example, *qua* mammal, or *qua* loud thing, or *qua* grey thing? Deutsch says that the answer is simple: The dubbing event was caused by the sample *qua* elephant and not by the sample *qua* mammal.

39 This example is borrowed from Miller (1992).

No further story is needed, as in the beaker case. The dubbing event can be seen as analogous to any other physical event that involves causation.

In short, Deutsch argues that "elephant" refers to elephants because it is the fact that the thing in front of Adam is an elephant that *causes* him to say "I dub this an 'elephant'" (Deutsch, 2023, p. 1811). The thing *qua* elephant is what causes him to speak and behave as he does. This analysis of the dubbing event, he argues, rests upon an uncontroversial fact about causation: that causes bring about their effects in virtue of some properties and not others.

The key to Deutsch's dissolution of the qua problem is the idea that "some features of a sample (its elephantness, for example) can be causally relevant with respect to an act of reference grounding even when its other features (its loudness, e.g.) are not (Deutsch, 2023, p. 1813). The notion of causal relevance is what supports the idea that the liquid's heat is causally relevant to the glass beaker breaking, while the liquid's acidity is causally irrelevant to the glass beaker breaking. Deutsch believes that if we can rely on causal relevance to individuate *the* cause of an event, then we should be able to rely on causal relevance to individuate *the* cause of a dubbing event.

3.4 WHY DEUTSCH'S DISSOLUTION OF THE QUA PROBLEM DOES NOT WORK

Deutsch's dissolution of the qua problem does not work due to causal relevance being insufficient for the individuation of a *unique* cause of a dubbing act. I will show that many properties are always causally relevant to a dubbing act. Since reference is supposedly fixed by the property that is causally relevant to the dubbing act, and since there are many such properties, Deutsch's theory does not yield determinate reference, and the qua problem persists.

For precision's sake, some clarifications about Deutsch's position need to be made first. While Deutsch states that the assumptions he makes about causation are uncontroversial, this is not entirely accurate. The assumption that "causation is *qua* certain properties and not *qua* others" (2023, p. 1812) commits Deutsch to a theory of causation that allows for a certain level of fine-grainedness of the causal relata. Donald Davidson's (1967) view, for example, according to which the causal relata are events understood as individuals existing at a particular time, would not be suitable for this; Davidsonian events do not seem to single out anything that could plausibly be *the* property *causing* the dubbing event in any single grounding of reference. Another view, the one that takes causal relata to be *facts*, also seems to be incompatible with properties being causes; facts are not usually thought of as composed of properties (they might be composed of *concepts,* or they might *exemplify* properties, but they do not involve properties directly). Jaegwon Kim's (1976) theory of events, which defines them as being instantiations of properties by objects at times, is better suited to Deutsch's approach. To ensure some level of precision and clarity in this discussion, we can minimally assume that the causal relata are events, and that these events are instantiations of properties which may or may not be causally relevant. However, my argument does not depend on the truth of this specific view on the metaphysics of causation. The only requirement I need is the same one Deutsch needs: that properties may be understood, within the theory of causation of our choice, as causally relevant or irrelevant to an effect.

Deutsch's assumption that some properties are causally relevant while others are not needs to be paired with an account of what differentiates the causally relevant from the causally irrelevant properties. Though Deutsch does not provide an account of causal relevance, he cites David Braun (1995, p. 450), Miller (1992, p. 429), and Maddy (1984, p. 465), who all discuss (in slightly different ways) counterfactual accounts of the distinction between causally relevant and causally irrelevant

properties.[40] Deutsch is also explicitly inspired (2023, pp. 1808, 1811, 1814) by Miller's (1992) argument, which is given in terms of a counterfactual analysis of causal relevance. Although Deutsch discusses the differences between his and Miller's argument (2023, pp. 1814–1816), he does not seem to believe that they differ in the treatment of causal relevance. Given all this, it can be reasonably assumed that he would accept an analysis of causal relevance in counterfactual terms.

A simplified way of characterising the distinction between causal relevance and irrelevance in counterfactual terms can be summed up as follows:

> ***Causal relevance of a property F:*** *c* which is F causes *e* which is G; and had *c* not been F, *e* would not have been G;
>
> ***Causal irrelevance of a property F:*** *c* which is F causes *e* which is G; and had *c* not been F, *e* would still have been G.

where *c* and *e* are events, and F and G are properties instantiated by those events.[41]

In the case of the glass beaker, the causal irrelevance of the acidity and the causal relevance of the heat is supported by the presumed truth of the following two counterfactuals:

> ***Relevance of heat:*** Had the liquid in the beaker been cold, the glass would not have broken.

---

40 It should be noted that while Braun describes counterfactual approaches as the "natural" way of understanding causal relevance (1995, p. 449), he later goes on to criticise them and propose his own, essentialist analysis of causal relevance. His novel proposal is that a property is causally relevant to an effect iff it is an essential property of the cause (p. 453). An essentialist analysis of causal relevance cannot work for Deutsch's dissolution of the qua problem – there are too many essential properties of any given cause, and if the causally relevant properties are the essential ones, that does not restrict the number of referential candidates to a sufficient degree. Because of this, I believe it is charitable to assume that Deutsch did not have an essentialist analysis of causal relevance in mind when citing Braun.

41 These counterfactuals are not supported in cases of overdetermination, such as those in which two causes acting simultaneously would have independently caused one and the same effect (think of two snipers aiming and shooting at a single target). Since there is no reason to think that dubbing acts are generally overdetermined, and since overdetermination is a rare occurrence, it can be set aside for now.

Irrelevance of acidity: Had the liquid in the beaker been non-acidic, the glass would still have broken.

In the case of the glass beaker, this counterfactual account of causal relevance delivers the right result, i.e. that it was the heat and not the acidity that "made a difference" in causing the effect of the glass breaking. According to Deutsch, the truth of these counterfactuals is supported by additional physical facts about the nature of the substances involved, e.g. the glass's breaking point with respect to a certain temperature and other such physical properties (2023, p. 1813). While the above counterfactuals support the relevance of one property and the irrelevance of another property, they do not tell us that the heat is *the unique cause* of the breakage; there are additional causally relevant properties here – properties without which the effect would not have occurred.

In the case of reference grounding, however, a newly introduced term must refer to a *unique* property in order for reference to be determinate. Deutsch suggests that referential determinacy can be explained by causal relevance. If Deutsch were to use the account of causal relevance we have assumed in the dubbing example, he might argue that the referential determinacy of the introduced terms is grounded in causal facts that support the following counterfactuals:

Relevance of elephant [**elephant**]: Had the thing not been an elephant, Adam would not have uttered "I will dub this an 'elephant.'"

Irrelevance of loudness [**loudness**]: Had the thing not been loud, Adam would still have uttered "I will dub this an 'elephant.'"

If these two counterfactuals are true, they support the conclusion that it was the thing's property of being an elephant, and not loudness, that caused the utterance "I will dub this an 'elephant.'" It should be noted here that, just as in the case of purely physical causation, these counterfactuals do not support the conclusion that *only* the property of being an elephant caused the dubbing utterance; and that is precisely what we need to avoid the referential indeterminacy presented by

the qua problem. As will be shown soon, there is always a multiplicity of causally relevant properties in dubbing cases, as in any other case of causation. Causal relevance does not provide us with the uniqueness required to solve (or dissolve) the qua problem.

The issue with purely causal reference grounding was the indeterminacy that resulted from multiple properties being equally good candidates for reference. The difficulty was singling out a unique property that would cause a dubbing event, thereby grounding reference in that property and not in others. What is needed for determinate reference in this framework, then, is the causal irrelevance of *all but one* property.

In cases of purely physical causation such as the beaker example, it is implausible that a unique property is causally relevant. While the heat of the liquid is causally relevant to the breaking of the beaker, so too is the beaker's shape, the thickness of the glass, the pressure in the room, and many other conditions *sine qua* the beaker would not have broken. Deutsch is right that nothing like the qua problem occurs in typical cases of causation, but this is simply a consequence of the fact that we do not require a unique causally relevant property to have brought about an effect. Saying that a basketball bounced both due to its shape and due to its density does not lead to any kind of indeterminacy. In contrast, if one wants to defend the possibility of purely causal reference grounding, there needs to be some causal mechanism that supports the singling out of a unique property that will ground reference. Otherwise, reference remains indeterminate. Deutsch's analogy with the beaker suggests that he believes that causal relevance will do the trick; however, causal relevance does not seem to deliver the uniqueness that is needed for referential determinacy.

Deutsch mentions that there might be issues with singling out *the* cause in cases of causal overdetermination – i.e., in cases where we have multiple distinct and individually sufficient causes of one and the same effect. He also states that these are issues that concern the metaphysics of causation more generally, and which have little to do with the specifics of reference. However, the

cases we have been considering here are not cases of overdetermination; the basketball's shape *alone* is not sufficient to cause the ball's bouncing, nor is its de

nsity. Having multiple causally relevant properties is not the same as having multiple distinctly sufficient causes.

One might respond that there is something akin to a unique causally relevant property in the cases of straightforward causation: it is what we would call "the cause" as opposed to "the background conditions." For example, most people would say that in the case of a house fire, the cause was some triggering event, such as a candle tipping over, rather than some background condition, such as the presence of oxygen in the room, even though there wouldn't have been a fire in the absence of either. However, there are two reasons why this cannot help Deutsch's case. First, there are many cases of causation in which background conditions are contrasted with *the causes* and not *the cause* of an event, and as such, do not supply the uniqueness that is necessary in the case of reference. Second, it is generally accepted that the distinction between the cause(s) and the background conditions is dependent on context, and thus not a matter of objective fact. In other words, what is the cause and what are the background conditions shifts based on the interest of the observer or, more generally, pragmatic reasons (Frisch, 2023; Lewis, 1973; Schaffer, 2005). We have no way of isolating a single cause in a way that supplies an objective causal ground for determinate reference.

Having established that we have no reason to believe events have a single cause as determined by what is causally relevant to them, we can show why that's also the case for dubbing events specifically. In other words, it can be shown that for any single dubbing event, there will be many properties that are causally relevant to its occurrence. A first and obvious example is that all intermediate links in the causal chain that connects the elephant encounter and the utterance are

causally relevant to the dubbing event. For the sake of argument, let's assume that [**elephant**] is true, i.e., that the property of being an elephant is causally relevant to the occurrence of the dubbing event in question. There are a number of intermediate causal steps between the encounter with the elephant and the dubber's utterance. For instance, the elephant must be *perceived* by the dubber, requiring the appropriate functioning of intricate perceptual and cognitive mechanisms, along with the necessary sound-producing machinery that enables speech and thus the dubbing utterance, etc.[42] What this means, in short, is that there are many links in this causal chain without which the dubbing act would not have occurred. Each of these intermediate causes is causally relevant:

> ***Relevance of perception***: Had Adam not perceived the thing in front of him, he would not have uttered "I dub this an 'elephant.'"

Many properties other than being an elephant are causally relevant to this dubbing event. There seems to be no way of excluding these intermediate steps from what is causally relevant to the dubbing act. There is no single cause of the dubbing act, and so no way of securing determinate reference through purely causal reference grounding.

One possible objection to this argument goes as follows. While there are many intermediate causes between Adam's encounter with the elephant and his utterance, these have little to do with the relevant sample in front of Adam. In other words, should we not limit ourselves to considering properties of the relevant sample only? Unfortunately, this kind of restriction does not help: there are properties *of the sample* in question that are causally relevant to the dubbing event, but which should not be part of what grounds the reference of the newly introduced term.

---

42 In this particular example we are assuming that the dubbing act occurred via an utterance. Of course, speech is not generally necessary to dub; a person may dub in written form, for example. However, these contingencies do not affect the overall argument that leads to this problem; there will always be intermediate steps between the cause and the effect in the cases under consideration.

For example, spatiotemporal properties of the elephant, such as the fact that it is located in front of Adam and not on the other side of the globe, are causally relevant to the dubbing act; had the elephant not been within a perceivable distance from Adam, the dubbing act would not have occurred. The size of the sample is also causally relevant; had the elephant been microscopic, the dubbing act would not have occurred. However, it does not seem plausible that that individual elephant's size or location within time and space should be part of what determines the reference of "elephant." However we go about restricting the pool of causally relevant properties, there are simply too many of them.

In addition to the fact that causal relevance does not deliver the uniqueness we need in the dubbing case, it is questionable whether [**elephant**] and [**loudness**] are true. Is it true that, had the thing in front of Adam not been an elephant, Adam would not have uttered "I will dub this an 'elephant'"? Maybe, maybe not; this clearly depends on facts about Adam's intentions. For example, we can easily imagine that Adam previously intended to name the next kind of animal he saw "elephant," regardless of what it was. If this were the case, the counterfactual would be false; had the object Adam encountered not been an elephant (but, for example, a mouse), Adam would still have dubbed it "elephant" due to his intention to name *a kind of animal* "elephant." Whether or not [**elephant**] and [**loudness**] are the case varies with the intentional states of the dubber.

Additionally, any explanation of why [**elephant**] and [**loudness**] are true requires us to invoke intentional states, unless we wish to say that causal relations are mysteriously brute. As Deutsch himself notes, it is not enough to simply state that there is a cause of the dubbing act; there needs to be a fuller story to account for *why* and *how* one specific property turns out to cause the dubbing act (2023, p. 1813). Deutsch's proposed solution for "filling in" the causal story in the case of dubbing invokes some intentional states, including wanting and trying to dub, for example (2023, pp. 1813–1814). However, Deutsch insists that the *content* of these intentions is completely

irrelevant to the grounding of the reference of the newly introduced term: "No particular intentional state or descriptive conception is *required* for securing determinate reference" in a metaphysical sense (2023, p. 1814). In other words, Deutsch argues that the relationship between a dubbing act and a property that caused it is not brute because we can explain it via facts about the dubber's intentions, even though the contents of these intentions have nothing to do with the actual reference of the new term. However, it is implausible that the content of the dubber's intention is irrelevant to the grounding of reference. If Adam never intended *to dub* elephants, but simply intended *to hunt* them, no dubbing act would have occurred. The content of these intentions, then, matters not only in an explanatory sense, but in a metaphysical sense: There would not have been a dubbing act without there being an intention *to dub*. The intention to dub is causally relevant to the dubbing event occurring.

Even if [**elephant**] and [**loudness**] were true, it can easily be shown that the content of the dubber's intentions is also causally relevant. For the sake of argument, assume that [**elephant**] is true. In this case, it seems as if the content of Adam's intention is also causally relevant:

> ***Relevance of intention****:* Had Adam not intended to dub something, then Adam would not have uttered "I will dub this an 'elephant.'"

It is very plausible that a dubbing act would not have occurred had the dubber lacked the intention to dub a thing. If we only have causal relevance to go by, there seems to be no way of excluding intentional states and their descriptive content from the set of causally relevant properties. Consequently, we have no way of excluding them from what grounds reference within Deutsch's framework.

There are several distinct issues for Deutsch's proposed dissolution of the qua problem here. First, the causal relevance of intentional states further severs the analogy between dubbing cases and purely physical cases of causation, since no intentional states need be causally relevant

in purely physical cases. Second, the relevance of the dubber's intentional state demonstrates that there is *at least* one other property in addition to the property of being an elephant that is causally relevant to the dubbing event. This means that, once again, there is no unique causally relevant property that would allow for the grounding of determinate reference in purely causal terms. The last difficulty presented by the relevance of intentions is that, once we've seen that causal relevance does not deliver a unique property that would ensure determinate reference, intentional states start looking like the best way out of this problem. If we already grant that intentions are causally relevant to the grounding of reference, why not grant that it is the content of these intentions that metaphysically ensures determinate reference? In other words, why not amend the framework and adopt a hybrid causal theory at this point?

Deutsch argues that reference is grounded in *the* property that is causally relevant to the dubbing act. However, as has been shown here, multiple properties are causally relevant to bringing about a dubbing act. So, causal relevance does not allow for individuation of a unique property that a newly introduced term refers to. At best, it allows us to individuate a large collection of properties that are causally relevant to the dubbing act; but this cannot help us secure determinate reference.

Nothing like the qua problem arises in purely physical cases of causation because, unlike in reference-grounding scenarios, there is no general need to ensure that a unique cause caused an effect. It is precisely the requirement of uniqueness, which is characteristic of attempts at securing determinate reference, that generates the qua problem for causal theories. Contrary to Deutsch's hopes, causal relevance is insufficient to secure such uniqueness; there is simply too much that is plausibly causally relevant to any given dubbing act. A particular difficulty is presented by the fact that it seems impossible to exclude the content of speakers' intentions from what is causally relevant to dubbing acts. This leaves Deutsch and other defenders of purely causal theories of

reference in a position where they need to find additional causal mechanisms that would, on one hand, exclude intentions and their contents from what grounds reference, and on the other, be sufficient to ensure determinate reference. It is difficult to say what these mechanisms might be. The qua problem for purely causal reference grounding persists.

CONCLUSION

The aim of this chapter was to consider whether purely causal theories of reference could face the sceptical challenge posed by Kripke in *WRPL,* which was to determine in virtue of what speakers ever mean anything by a sign. Purely causal theories of reference state that reference can in some cases be fixed *just* through causal contact between a speaker and a thing during a baptismal act; and reference can be borrowed through causal contact between the baptist and other speakers. If this is the case, speakers mean something by a sign in virtue of facts about certain causal relationships obtaining.

As has been shown, even though the basic ideas behind purely causal reference fixing are promising, these theories have not yet resolved the sceptical challenge. Purely causal accounts of reference run into several issues when they refine which kind of causal relationship needs to obtain between speakers and objects during baptisms. There is a temptation to say that an object's causal effect on speakers' perceptual capacities or their neural correlates determines reference; this is the basic idea behind Maddy's and Miller's proposals (1984 and 1992, respectively). Basing reference fixing on perceptual capacities, however, "passes the referential buck;" it explains referential content in terms of mental content, which goes against the non-circularity requirement of the sceptical challenge. Basing reference fixing on the neural correlates of perceptual capacities can go

wrong in two ways. If one implicitly interprets neural correlates as having properties which are in essence intentional, that is, as having built-in representational capacities, we are once again passing the referential buck and providing a circular explanation. If one does not interpret neural correlates as already intentional, it becomes difficult to see how such states can explain the possibility of misrepresentation, meaning that there are no conditions under which a state counts as incorrect; this goes against the normativity requirement. In any case, causal theories of reference that isolate certain mental states or their neural correlates as what fixes reference run into issues that are characteristic of naturalistic theories of *mental* content. These issues will be discussed in their proper context in Chapter 4.

It is possible to resist the temptation of invoking mental content or brain states in explaining what causally grounds reference. Deutsch (2023) develops such a proposal, which ultimately fails to yield determinate reference. His idea is that the referent of a new term is fixed in virtue of the properties that are causally relevant to the act of baptism. This proposal fails because relying exclusively on causal mechanisms, and in particular on the notion of causal relevance, is insufficient for individuating a single referent: There is always a multiplicity of causally relevant properties during any single baptism. Referential indeterminacy, in addition to being undesirable in and of itself, goes against the extensional requirement of the sceptical challenge.

There is nothing about the basic structure of causal theories of reference that prevents them from providing an answer to the sceptical challenge in the future. What remains to be seen is whether theories of causation provide additional tools that could be used to construct a theory of reference that yields appropriate results. Such a theory would explain reference in a way that allows for both determinate extensions and the possibility of misrepresentation, all without passing the referential buck. At this time, I do not see a way in which the mechanisms of causation could yield this kind of result.

4. CAUSAL AND TELEOLOGICAL THEORIES OF MENTAL CONTENT

INTRODUCTION

The central question of this thesis – the CDQ[43] – pressed us to say what makes it the case that representations have the content that they do. The question can be answered "language-first" or "mind-first"; until now, this work has engaged with the first kind of approach. As has been noted in the previous chapters (and also by Kripke in 1982, pp. 22, 42, and throughout *WRPL*), there is a tendency to want to explain the representational properties of language via the representational properties of mental states. So, for example, it may be tempting to say that the word "horse" means (represents) horses if its usage is properly linked to a speaker's capacity for recognising horses, as Miller proposed (1992). Of course, the act or state of recognising a horse is already representational and – importantly – includes the possibility of *misrepresentation.* This path leads us to sharpening the focus of the CDQ from representation *tout court* to mental representation. It might be that determining what makes it the case that mental representations have the content that they do is the key to resolving the sceptical challenge and to answering the CDQ. Once mental representation is sorted, one can use the determinate contents of mental states

[43] To reiterate, "CDQ" stands for the *content determination question*: What makes representations, linguistic or mental, have the content that they do?

to explain the determinate contents of linguistic expressions. The main task is to conjure up a plausible explanation of how *any* representations come to have the contents that they do.

To repeat Boghossian's interpretation of the sceptical challenge, "having a meaning is essentially a matter of possessing a correctness condition. And the skeptical challenge is to explain how anything could possess that" (2002, p. 149). Many kinds of mental states seem to have correctness conditions: Singular perceptual states can be accurate or inaccurate, while beliefs and thoughts can be true or false.[44] This makes such mental states assessable in terms of correctness. So, a false belief is incorrect, while an accurate perception is correct. This is just an oblique way of describing the fact that mental states can represent and misrepresent the world. But in virtue of what do mental states have the representational contents that they do?

The question of what determines mental content has been prominent in analytic philosophy for nearly a century, and the literature on this topic is immense. The present work seeks to investigate whether contemporary theories of mental content can eschew the problems raised by Kripke's sceptical challenge and give a satisfying answer to the CDQ. This chapter considers causal-informational and teleological theories, whose development has been historically intertwined. Some historical and background considerations must be laid out first, since much of the theoretical machinery of newer theories has been constructed to avoid objections raised against older theories. These background considerations will not be in any way comprehensive and only serve to provide context for properly understanding contemporary approaches. As will be shown in Section 4.2, the problems of theories of mental content follow a familiar pattern. Theories struggle either with issues of content indeterminacy – the theory implies that there are too many

[44] Whether other kinds of mental states such as emotions and desires can be evaluated, and what such evaluations would look like, is more controversial. We can ignore such controversies here and focus on the kinds of mental states which are usually taken to be semantically evaluable and, as such, representational.

inappropriate candidates for what the content of a mental state is, or it simply yields disjunctive, deviant contents – or they cannot account for the possibility of misrepresentation. These issues mirror the struggle to meet two of the requirements posed by the sceptical challenge: the extensionality and the normativity requirement, respectively. As will become apparent, the theories considered here have not produced a satisfactory solution to the sceptical challenge.

After having shown that the contemporary versions of these naturalising theories are not up to the task of resolving Kripke's sceptical paradox, I will argue that teleosemantics – arguably the most popular theory of mental content to date – fails to properly account for representational content's role in explaining rational thought and behaviour. The programme of naturalising mental content, it seems, faces great difficulties on many fronts.

Some notes about the title and contents of this chapter are in order. The title of this chapter is not "theories of mental content" but "causal and teleological theories of mental content." This framing delimits the discussion in some important ways.[45] First, there is a kind of theory of mental content that is omitted here: theories that explain the contents of mental states in terms of phenomenal properties. These theories, which are usually called phenomenal intentionality theories, are excluded from this chapter because they will be covered in detail in Chapter 5. The choice to dedicate phenomenal intentionality theories a separate chapter is partly motivated by the fact that they do not aim to naturalise content, which is one of the main goals of the theories covered in this chapter. The other reason is that this chapter serves as a very long precursor to the one dedicated to phenomenal intentionality theories. The failures of causal and teleological theories of mental content are the main reason for the recent rise of the phenomenal intentionality

[45] To be entirely precise, this thesis excludes much more than is explicitly noted in this paragraph. In particular, this work does not engage with either primitivist or eliminativist theories of mental content.

programme, and the particular manner in which they fail has informed and shaped these new competitor theories.

A second note concerns the fact that this chapter does not go into the very influential view according to which the mind should be understood as a computational system. Computational theories of the mind are connected with functionalist and causal approaches to the mind, but their focus is more on *what counts as a mental state or mind* and not on the CDQ; Gualtiero Piccinini (2004) provides a helpful overview of the rise of computational theories, their goals, and the reasons they have ignored or struggled to account for mental content in a non-circular manner. To put it very briefly, Piccinini convincingly argues that "whether the mind is computational and whether the mind has content (and how it manages to have content) are different problems that need to be solved independently of each other" (2004, pp. 403–404). Since the focus of this thesis is representational content, the discussions directly related to computational theories of mind have been excluded from this chapter.

This chapter is divided into four sections. Section 4.1 briefly presents the goals and history of causal theories. Most of the section is devoted to the information-based causal theory developed by Fred Dretske in *Knowledge and the Flow of Information* (1981/1982). Two crucial problems faced by Dretske's theory will recur periodically for its descendants: the struggle to explain the possibility of misrepresentation and the indeterminacy of content. Section 4.2 explains how and why causal theories grew to incorporate the notion of function, eventually leading to the creation of teleosemantics. Biological functions are a promising theoretical tool that allows theories of content to explain how misrepresentation is possible. Indeterminacy, as will be shown, is a much harder problem to get rid of. Section 4.3 considers some methodological assumptions held by teleosemanticists through the lens of a conversation between Angela Mendelovici and Marc Artiga. It will become apparent that different methodological approaches may lead to the isolation of

completely different phenomena we may want to call "mental content." In particular, the object of inquiry of teleosemantics does not correspond to what we have access to introspectively and ordinarily identify as mental content. In Section 4.4, I argue that these methodological assumptions lead to an unacceptable consequence: Teleosemantics does not have the tools to adequately describe rational thought and behaviour. While it may be acceptable that teleosemantics does not support or match what we believe about mental contents based on introspective access, its inability to explain behaviour is more alarming. What teleosemantics individuates as mental content cannot play the normal explanatory role we usually associate with mental content, that is, it cannot figure in common-sense psychology.

4.1 CAUSAL THEORIES OF MENTAL CONTENT: THEIR HISTORY AND THEIR PROBLEMS

Causal theories of mental content are theories that explain how beliefs, thoughts, and other intentional states come to have the content they do via some sort of causal connection between the mental state and the object that is being represented by the state (Schulte, 2023, pp. 21–22). The basic idea is that things sometimes cause their representations – the presence of an orange on the table causes me to believe that there is an orange on the table – and that this kind of occurrence is at the basis of representation. This section is dedicated to overviewing some notable causal theories of mental content and the problems they ran into.

Traditional defenders of causal theories of mental content are nearly always explicit about their intent to *naturalise* content, unlike defenders of causal theories of reference. Dennis Stampe calls his work an "attempt to naturalize contents" (1977, p. 51). Fred Dretske playfully describes

his project in the preface to *Knowledge and the Flow of Information* as attempting to "bake a mental cake using only physical yeast and flour" (1981/1982, p. xi). Jerry Fodor states that the goal for a "serious" theory of representation is to specify naturalistic conditions under which something represents something (1984, p. 232). Given these naturalistic motivations, seeking to explain representation through causation is an obvious choice, since causation is widely held to be natural.[46] Another motivating factor, specifically in Fodor's case, is that representations seem to have a causal role when behaviour is concerned: Beliefs and desires *cause* humans (and some animals) to act in a certain way (1987/1993, pp. x–xii). So, representational mental states are seemingly part of the natural, causal chain of events, and our theories should reflect this fact.

Stampe is often cited as one of the initial popularisers of causal theories of content, and one of the central figures who laid the foundations for subsequent causal theories of mental content specifically (Adams & Aizawa, 2021). In his 1977 paper "Toward a Causal Theory of Linguistic Representation" he begins by stating his conviction that causal relations can explain representational content, whether mental or linguistic. As has been mentioned previously, the representational or intentional has the peculiar characteristic of being about an object; Stampe's idea is that this aboutness can be analysed through causal relations between the representation and the thing it represents.

Stampe outlines some more reasons causal theories are attractive in addition to their naturalness. He argues that the success we have in navigating the world would be a mystery if our representations of it did not involve the world's objects in some tangible, real way; and causal relations seem to be *real* in a way that would satisfy most scientifically-minded naturalists (1977, p.

---

[46] I have found no evidence to the contrary. Some are causal eliminativists, but I have not found an account that defends the idea that causation exists *and* is unnatural.

43). Another reason to pursue a causal explanation of representation is that there are clear-cut cases in which it seems to be the best explanation of representational properties. Consider a photograph of a person who has an identical twin. The photograph is a representation of *that person*, and not of their twin, even though the twin has (let's assume) identical properties to those of their sibling. A plausible explanation of this fact is that there was an appropriate kind of causal connection between the photographed twin and the camera used to produce the photograph, and that no such causal connection obtained with the other twin (1977, p. 43).

Stampe also immediately sets out the problems causal theories must tackle: selecting the *right* causal relations, i.e. those that determine content vs. those that do not; accounting for misrepresentation; and explaining how non-referring terms work. He notices that in order for causal theories to account for misrepresentation, they might need to invoke teleology, e.g., the notions of normalcy conditions or functions (1977, pp. 44, 51). Whether the notion of normalcy conditions or of function can, in turn, be cashed out naturalistically will become the object of contention in the decades that followed.

The biggest struggle for the causal theorist will be to delineate which causes should be taken as the ones responsible for the contents of single representational states, and which ones are irrelevant to that purpose. Similarly to the *qua problem* for causal theories of reference, this is in part due to the fact that one is in causal contact with many things at any given time, and many of these things seem to be involved in causing any single representation. Causation is common and representation is not; there is much more causing than representing happening in the world. As will become apparent soon, "the recurring problem is that the attempts to separate content-determining from non-content-determining causes threaten to smuggle in semantic elements" (Adams & Aizawa, 2021). In other words, causal theorists must specify one or more principles

that privilege some causes of a representation over others, and must do so without relying on "semantic elements" – without violating the non-circularity requirement.

To make the issue of separating content-determining from non-content-determining causes more vivid we can use the example of simple visual perceptual representation of a ripe strawberry. There is a myriad of causes of a perception that go from distal to more and more proximal – until we reach the most proximal cause of the representation, which is likely going to be some neural signal. Why should we consider the strawberry "out there" as the content-determining cause of the perception? How can the intuitively irrelevant sections of the causal chain, such as the neural signal or the perceiver's parents meeting in the 1970s, be dismissed as part of what the representation is about? This is traditionally known as the *distality problem* and mirrors the extensional requirement of the sceptical challenge.

Another notorious issue for causal theories of mental content – often called the *disjunction problem* – similarly highlights how difficult it is to select for content-determinant causes. There are many things which might cause a representation of X other than X itself – images of X, being bonked in the head, being asked about X, the administration of psychoactive substances. The issue is the following: Why is the object of my representation X and not X *or* a hallucination of X? The disjunction problem asks us what makes it the case that we represent X and not some of the many deviant, disjunctive properties that seem to be equally eligible. More importantly, the disjunction problem is linked to the issue of misrepresentation. If misrepresentation is possible, I will sometimes have a mental state representing an X while I am in fact in causal contact with a Y.[47] In other words, I can *mistake* a Y for an X. But why not interpret my mental state as *correctly*

[47] In this example it is assumed that I have been in contact with Xs in the past.

representing Xs *or* Ys? According to the scenario, I am in causal contact with a Y, after all. Which principle can grant us that the Y in question is not involved in determining the content of my representation? If no such principle can be found, there is no way of characterising anything as misrepresentation. This issue and others like it mirror the normativity requirement of the sceptical challenge.

The task of causal theories of mental content, then, is to find the right content-determining causes. The desiderata for a successful theory are, by now, familiar: We want mental representations to represent the right things, and we want an account that explains how correctness conditions arise – without which we lose the possibility of misrepresentation.

One of the early attempts at selecting the relevant content-determining causes of representation was made by Dretske (1981/1982). Dretske's main project was to provide an account of knowledge based on it being belief caused by information. Because of this, his account and others similar to it have sometimes been named causal-informational theories of mental content. Information is "out there" and is assumed to precede observers, even if it cannot be defined without reference to an observer (or "receiver," if we borrow terminology from communication theory, as Dretske does) (1982, pp. 44–45). Information, within Dretske's system, is supposed to stand in for a sort of *natural meaning,* a naturally occurring source of what will (in a subsequent specification) constitute intentional content. In other words, information is supposed to bridge the gap between belief and the world, along with providing a basis for truth; it is what allows us to learn about the world (1982, p. 44). "Roughly speaking, information is that commodity capable of yielding knowledge, and what information a signal carries is what we can learn from it."

The notion that interests us here is the one of *informational content,* which is, according to Dretske, the specific information about the world carried by "signals" (1982, p. 65). The notion

informational content is supposed to capture is that of "what-it-is-we-can-learn from that signal or message" (1982, p. 47). Dretske defines informational content as follows:

> **Def:** A signal *r* carries the information that source *s* is F iff the conditional probability of *s*'s being F, given *r* (and the knowledge of the receiver, *k*) is 1 (but, given *k* alone, is less than 1). (Dretske, 1982, p. 65).

What this means, roughly, is that a signal carries information about something having a specific property exactly when it is (probabilistically) certain that something has that property *given* the occurrence of the signal, relative to the knowledge of the receiver of the signal. To use a simple example, imagine that you see a single set of fresh footprints in the snow. The footprints signal that someone has recently passed there (assuming you have background knowledge about how footprints in the snow work). Had the footprints not been there, the probability of someone recently having passed there (relative to your background knowledge) would have been lower. The footprints, then, carry information – the information that someone has passed there. They carry it *for* a specific observer, but that does not mean that the information itself is observer-dependent.

The mention of the knowledge of the receiver in the definition of informational content might seem alarming at first, as Dretske aims to give a naturalistic explanation of content (and knowledge!) via that of information. Invoking knowledge in the definition of informational content, then, could lead to circularity. He clarifies that the definition is recursive and that one eventually arrives at a point where no prior knowledge is necessary for a signal to carry specific information. (Dretske, 1982, pp. 86–87)

Dretske must refine this definition to rule out unwanted consequences about what information a signal may carry. As it stands, the definition is too permissive. While Dretske allows signals to carry quite a lot of information – if a signal carries information that F, it also carries everything that can be logically derived from that – he wants to defend the idea that signals *do not*

carry information about contingent correlations, even if they happen with perfect regularity (1982, pp. 71–75). We want the position of mercury in a thermometer to carry information about the temperature of the room it is in, but not about a distant room that (by pure chance) always has the same temperature. So, correlation between A and B – even if it is perfectly regular – is not enough for A to carry information about B. What this means is that Dretske must go beyond the extensional aspect of informational content. In his own words, "our definition of information distinguishes between extensionally equivalent pieces of information. In this respect, and to this degree, statements describing the information carried by a signal are *intensional* descriptions" (Dretske, 1982, p. 76, italics mine).

The way Dretske restricts what information signals carry is by using counterfactuals of this sort: Had the temperature of my room been different, the thermometer would have given a different reading; but the thermometer would be unaffected by changes in temperature in a distant, unrelated room. This is true in virtue of certain laws of nature, that is, in virtue of certain *nomic regularities* (1982, pp. 76–77). This kind of counterfactual is supposed to exclude cases of arbitrary, random, non-nomic coextensionality; just because there is some random regularity out there in the world between property F and property G does not mean that Fs carry information about Gs. In other words, Dretske relies on a distinction between arbitrary and non-arbitrary regularity in the world for his theory to deliver the desired results, i.e. for things in the world to carry the information we deem to be intuitively relevant. To complete this picture, he pairs it with his own account of the laws of nature that, contra some of his contemporaries, views them as inherently different than mere true generalisations (Dretske, 1977).

How does information and informational content relate to representational content? Dretske recognises that information as he characterises it cannot *misrepresent*, which is one of the features of representation we want to preserve; there is no such thing as false information or mis-

information (1982, p. 45). This is because, as was mentioned earlier, information is a kind of "natural meaning" in the sense proposed by Grice – smoke *naturally means* fire, but only if there actually is a fire; there is no way for smoke to misrepresent that there is a fire (1957). Dretske is very careful to separate information and informational content from what he calls "meaning" and "semantics," and states that "the task is to describe the way structures having a semantic content can be developed out of information-bearing structures" (1982, p. 175). Even greatly restricting how much information a signal carries, which Dretske does in an effort to individuate the propositional content supposedly carried by belief (1982, p. 176), does not yet allow for misrepresentation (1982, pp. 190–191). To introduce the possibility of error and finally obtain semantic content from informational content, Dretske distinguishes between a period of concept-formation – during which a subject learns to track a certain type of information in their environment – and a subsequent period in which an acquired concept is deployed in specific instances (1982, pp. 193–197). It is only in this second phase that misrepresentation becomes possible.

This two-step concept-formation mechanism described by Dretske has been convincingly criticised by Fodor (1990, p. 41). First, Fodor points out that there can be no principled distinction between the pre- and post-learning phases, which is an issue because the distinction between truth and falsity hinges on it. Second, this framework does not leave room for potentially innate, non-learned representation to occur, since no learning period would exist in such cases. Most damningly, Fodor points out that it seems unclear that the learning period solves the issue of misrepresentation at all. Imagine that a subject encounters many cows during the learning period, and that only cows cause (i.e., lawfully correlate with) certain responses that allow the subject to acquire the concept "cow" at the end of the learning period. Now imagine that after the end of the learning period, the subject encounters a horse in the dark but believes they encountered a cow. Fodor asks: What would have happened had this encounter happened during the learning

period? It seems like we have no principled reason to believe that it would not have elicited a "cow" response; so, we have no principled reason to exclude it from what, in fact, forms the concept "cow." The supposed error is subsumed into the correct application of a disjunctive "horse or cow" concept. The possibility of misrepresentation has once again disappeared. This kind of issue – the fact that "mistakes" tend to be subsumed into correct applications of a different, deviant concept – repeats itself in different iterations of causal theories. Causal-informational theorists, then, had to introduce new notions to allow for misrepresentation, the most prominent one being that of function.

Before moving onto the role functions have developed in causal theories of mental content, an important critique by Robert Cummins should be mentioned. Cummins puts forward an argument that affects some causal theories of mental content, including the sketch of Dretske's theory presented here, as long as we accept some assumptions about human psychology. He defines his target as those theories proposing that "the contents of the semantic primitives in the human scheme of mental representation are determined by their role in detection," where detection is the proper instantiation of a representation after causal contact with a property (1997, p. 535). He then argues that any theory of this form is incompatible with the fact that the detection of distal properties is mediated by "theory," that is, by knowledge about which (more) proximal properties reliably indicate that the distal property is present. To reliably detect cats we need a "theory of catness": we need to know what cats look like (Cummins, 1997, p. 536). To know what cats look like – to know that cats are generally furry, quadrupeds, smaller than humans, whiskered – we must have acquired the concept "cat" already (along with many other concepts). But if one has to have the concept "cat" to reliably detect cats, they must already have a mental state with the appropriate content in place before detection becomes possible; so reliable detection cannot be what ultimately explains how the content of representations of cats is fixed. In other words, causal theories of mental content cannot provide a non-circular account of representation.

Cummins' argument can be countered by arguing that not all reliable detection must be mediated by theory; this is the objection Robert D. Rupert raises (2001). Rupert argues that it is possible that detection happens via implicit theory. For example, detection could happen via innate concepts, or other non-learned forms of representation (2001, pp. 503–505). While this may well be the case, it does not help us in our task of finding what ultimately fixes the content of representations: We have "passed the referential buck" (or, to be precise, the representational buck) once again, this time shifting the burden onto the existence of innate representational states. This is not an adequate answer to the CDQ, since it relies on primitives that account for content; it does not respect the non-circularity requirement. What we need is an account of representation that tells us how the content of single representational states is determined without relying on other contentful states.

4.2 FUNCTIONS, MISREPRESENTATION, INDETERMINACY: THE INCEPTION OF TELEOSEMANTICS

Simple causal-informational theories struggle to account for the possibility of misrepresentation. Many have thought that the key to resolving this problem lies in the notion of *function,* including Stampe, who kicked off the programme of explaining intentionality via causal relations (Stampe, 1977; but also Dretske, 1986). Functions lend themselves to being talked about in terms of normal and abnormal functioning or malfunctioning; therein lies the hope that one can work this into an appropriate notion of misrepresentation. Functions may also restrict the scope of causes to be considered relevant to the determination of content. Stampe optimistically writes:

> A representation will ordinarily stand in the specific causal relationship to – that is, will be a representation of – any number of things back through a causal chain. Certain considerations about

> the mode of representation (considerations of a functional nature) will select some one of these links as the one that is the thing (if any) seen, or referred to, or depicted, or whatever the appropriate modal may be. (Stampe, 1977, p. 54)

This section will quickly overview some notable examples of how functions have been worked into causal theories of mental content. A particularly influential approach uses certain *biological* functions to explain mental content.

Dretske eventually included functions into his theory (1986). After having grappled with criticisms of his 1981 proposal, he opens this article by centring the question of misrepresentation: "How is it possible for physical systems to misrepresent their surroundings?" (1986, p. 17). The physical systems Dretske is most interested in are humans, whom he believes to be the source of underived representational capacity. In other words, he continued seeking a reductive answer to the fundamental question of what makes it the case that something represents something else. As Dretske knew in 1981, natural "meaning" – which he equates with informational content – simply will not do, since there is no such thing as misinformation. So, what is "nature's way of making a mistake" (Dretske, 1986, p. 18)?

The solution Dretske proposes is that some signals have the function of indicating a certain condition through their being a certain way. This allows for misrepresentation to be defined as a signal's failure to carry information in a way that conforms with its function to carry that information. Of course, Dretske recognises that the success of this characterisation of meaning will heavily depend on whether the notion of function can be naturalised. Many functions seem to depend on people's intentions and beliefs. For example, the function of a thermometer is to show the temperature of what surrounds it because its creator intended for the thermometer to do so. A naturalistically described configuration of glass, mercury and so on does not have a function outside what has been imparted on it by a human agent; and what has been imparted on it is further dependent on the desires, beliefs, and interest of the humans creating and using it. To avoid a

circular explanation of representation, then, functions must be analysable in a way that does not "smuggle in" any semantic elements. In other words, the functions we are looking for should not exist in virtue of human purposes and desires.

An obvious place to look for a truly natural function is in the biological workings of organisms. Dretske focuses on biological systems he describes as having an "information-gathering role" that developed through the organism's adaptation to the environment (1986, p. 25). Intuitive candidates for this type of system are perceptual mechanisms, which can be described as having developed to have the function of gathering information. To illustrate this, Dretske describes the (now notorious) example of magnetotactic bacteria, which have receptors that make them sensitive to the Earth's magnetic field (1986, pp. 26–27). The receptors are advantageous because they help the bacteria travel towards the Earth's poles and away from oxygen-rich environments which are toxic to them. Bacteria in the northern hemisphere are prone to travel north; bacteria in the southern hemisphere are prone to travel south. This polarity is important to the example. A southern magnetotactic bacterium transported to the northern hemisphere will continue travelling towards the south, leading it away from a favourable environment and towards its death. Dretske argues that this is a rudimentary case of misrepresentation. The bacteria's receptors have evolved *with the function to indicate* the location of oxygen-free water; this is the reason the receptors have been selected for. A southern bacterium in the northern hemisphere *misrepresents* the south as being an oxygen-free location.

Dretske anticipates an objection that will plague many subsequent functional analyses of mental content: the fact that natural functions are not as determinate as we might want the corresponding intentional content to be (1986, p. 28). For example, it is unclear whether the magnetotactic bacteria's receptors have developed the function to indicate oxygen-free locations and not the function to indicate the prevailing magnetic pole or, even more directly, the presence

of a certain kind of magnetic field (or a deviant, disjunctive property encompassing all of the above). In addition to this, the function of a system that has evolved to indicate can be analysed in a way that reduces it back to having mere *natural* meaning: If the receptors are seen as having the function to indicate the presence of a magnetic field, they will never misrepresent. The receptors simply naturally indicate – in the sense that they carry information about – the orientation of a magnetic field in the immediate environment, just like smoke indicates the presence of fire. This kind of natural meaning does not allow for misrepresentation, as has been established in the previous. The possibility of error has once again slipped through our fingers. The upshot is that the introduction of functions, even natural ones, does not immediately resolve the issues of misrepresentation and indeterminacy that plagued causal theories of mental content.

Theories about mental content that involve biological functions are usually called biosemantic or teleosemantic theories. Many of these theories involve a causal component, whether explicitly or implicitly. The moniker "teleosemantics" groups together disparate theories which differ in many respects. These theories have two basic things in common: the goal of naturalising content and the idea that content is somehow metaphysically dependent on naturally occurring functions. Teleosemantics is currently the most popular account of mental content, though that might be because it is often perceived as the "only game in town" by those who seek a naturalistic answer to the CDQ (Neander, 2017, p. 89). However, as has been foreshadowed in the previous paragraph, merely tacking functions onto a causal theory à la Dretske does not yield the desired results. I have used the example of Dretske's work, which had developed from a simpler causal theory to one that involves functions, to illustrate the history of the issues faced by naturalistic theories of mental and explain what motivated their transformation. This history could have been told in much more detail, which has been omitted for the sake of brevity. The crucial point is that the problem of misrepresentation, and the closely connected problem of indeterminacy, are at the forefront of what prevents causal theories of mental content from

succeeding in their goals. As will be shown shortly, this fact shaped the evolution of teleosemantic theories as well.

I have shown that the introduction of functions into causal theories is necessitated by the theoretical need to account for misrepresentation. This move corresponds, in the language used in this work, to an attempt to resolve the normative challenge posed by the sceptical paradox. The extensional challenge remains: We want theories to attribute *the right contents* to representational states. Teleosemantic theories famously struggle with this part and in particular with indeterminacy – Karen Neander lists six (!) determinacy challenges for teleosemantics (2017, p. 150). Marc Artiga writes about the teleosemantic project, "the significance of the indeterminacy problem is hard to exaggerate, since it has played a leading role in establishing the idea that this naturalistic project might be doomed to failure" (2021, p. 471). This section will overview some of the teleosemanticists' attempts to rise to the challenge.

Most teleosemantic accounts state that the relevant functions should be understood in terms of what has happened in evolutionary history. In other words, functions emerge from a historical process that happens under selective pressure. For example, extant perceptual mechanisms have been selected for due to their properties contributing to the survival and reproduction of the organisms that had them. The function of perceptual mechanisms, then, is to do whatever contributed to the survival and reproduction of the organisms that had them. This view is often called the *etiological* theory of biological function (Schulte & Neander, 2022). Teleosemanticists believe that there are biological systems whose function is to produce mental representations, and that the way these systems have been selected for throughout evolutionary history is part of what determines the content of single representations. To give an actual example of this kind of position, we can cite Ruth G. Millikan's summary of her theory: "Cognitive systems are designed by evolution to make abstract pictures of the organism's environment and to be

guided by these pictures in the production of appropriate actions" (1993/1995, p. 11). This is the gist of etiological teleosemantic theories.

Teleosemantic theories are usually divided into two large camps which can be roughly described as input-oriented and output-oriented teleosemantics (Schulte, 2023, p. 35). The two camps differ in terms of which biological mechanism is more directly responsible for determining the content of a representation: the one that creates (produces) the representation or the one that utilises (consumes) the representation. Input-oriented teleosemantics, also commonly called "producer teleosemantics," argues that the content of a representation is determined by the function of the representation-producing mechanism. The idea, roughly, is that a representation has content X because there is a mechanism with the function to create that representation in response to the presence of Xs. So, a magnetotactic bacterium's magnetosomes (the producers of representations in this case) produce representations with the content "oxygen-free location" because they have evolved to have the function to indicate oxygen-free locations. Dretske's amended theory (1986) is of the producer type, along with the more recent proposal put forward by Neander (2017) and J. R. G. Williams (2020).[48]

Output-oriented teleosemantics, also known as consumer teleosemantics, centres the role of the *use* an organism makes of representations to determine their content. Consumers are, by definition, the systems an organism uses in response to the representations a producer produces. Output-oriented teleosemantics frames representations as signals sent from a representation-producing mechanism to a representation-consuming mechanism. For example, a frog's motor system (the consumer of a representation in this case) will activate tongue-snapping behaviour as

[48] Williams confines this kind of teleosemantic analysis to lower-level representational states (e.g. perception) and argues belief and other higher-level representations should be treated differently.

a response to a fly-representation produced by the visual system. This is just a very technical way of saying that frogs flick their tongues when they see a fly. Representational content is determined by the proper function of the consumer-producer system. Specifically, a representation's content is determined by the conditions that contributed to the selection of the consumer-producer system; Millikan calls these adaptive circumstances "normal conditions." Normal conditions are the conditions under which a system successfully performs its proper function (Millikan, 1993/1995, pp. 28–29, 57).[49] So, the frog represents something like "there is food here" because what was selected for was tongue-snapping behaviour in response to a representation of food, whenever food was actually there (the food being actually there was the normal condition in which this system evolved). Representation-producers have an important role in output-oriented theories, but ultimately, the function of mechanisms which *use, respond to,* or *exploit* representations is primary. Consumer-oriented semantics is defended by Millikan (originally in 1984) and David Papineau (originally in 1984).

The main difference between input- and output-oriented teleosemantic theories, then, is in their understanding of what the "representational system" that is selected for is. For input-oriented teleosemantics, the representational system is whatever produces representations; for output-oriented teleosemantics, the representational system *includes* the consumers of what the representation-producing system produces. Both kinds of theories basically state that the content of representations is determined by whatever caused the representational system (whether it includes consumers or not) to be selected for.

[49] Normal conditions need not be *statistically* normal, i.e., they aren't necessarily common. Millikan uses the example of sperm to illustrate this: though their proper function is to fertilise an ovum, it is statistically extremely unlikely that any single sperm actually achieves this goal (1984, p. 29).

While teleosemantic theories may have a somewhat plausible route to dealing with misrepresentation, they struggle with indeterminacy. Artiga differentiates between the horizontal and the vertical problem for teleosemantics, which are two problems that correspond to specifications of the distality and the disjunction problem (2021). The horizontal problem puts pressure on teleosemantics to individuate which link in a causal path causes a mental state (is it a fly? The light bouncing off the fly? The retinal image? And so on). Predictably, this input-side problem is more urgent for input-focused, producer theories. The vertical problem puts pressure on teleosemantics to individuate what a system was selected for: has the frog's tongue-snapping response been selected for because it was advantageous for the frog to ingest the fly? Or an insect? Or a small, nutritious thing? And so on. This is an output-side problem and is more relevant to output-focused, consumer theories. There have been attempts at dealing with this indeterminacy – such as Artiga, 2021; Martínez, 2013 – but none have been accepted by the wider philosophical community as successful solutions. The struggle with indeterminacy has become so ingrained that it led to the development of a position arguing that teleosemantics should embrace the indeterminacy of contents (Bergman, 2023, 2025). Karl Bergman's suggestion is that while teleosemantics yields indeterminate contents, the indeterminacy in question is "well-behaved" and relatively contained. In any case, teleosemantic theories do not yet yield determinate mental contents, whether one believes this to be an acceptable result or not.

4.3 RELIABLE MISREPRESENTATION AND PRE-THEORETICAL ACCESS TO MENTAL CONTENT

Even if one were to successfully grapple with issues of indeterminacy, teleosemantics has a deep philosophical difficulty which has not been sufficiently appreciated: It cannot adequately describe intuitively rational behaviour and thought processes as such. The reasons why this problem arises for teleosemantics are methodological and can be connected to long-standing debates between externalists and internalists about content. In this section, I will use a back-and-forth dispute between Mendelovici and Artiga, who argue about the importance of accommodating reliable misrepresentation, to put into focus the methodological assumptions that are ultimately responsible for this deep problem for teleosemantics. After these methodological assumptions have been crystallised, I will turn to their consequences in Section 4.4.

Mendelovici proposes an argument against teleosemantics and all other "tracking theories," that is, theories that understand representation in terms of an organism's capacity for tracking features of the environment (2013, 2016).[50] In her two papers, the second of which is a reply to Artiga's reply to her first paper on the topic (2013), Mendelovici argues that – intuitively – it should be possible for a theory of representation to account for *reliable misrepresentation*. She then argues that teleosemantic theories fail to do so.[51] She concludes by highlighting some negative

[50] Mendelovici's argument is explicitly against "tracking theories," which she defines as theories that take mental representation to be "a relation of causation or correlation holding between mental representations and things in the world in content-endowing circumstances, e.g. circumstances in which the tokening of a representation is useful, adaptive, or involves a sufficiently strong causal connection" (2013, p. 422).

[51] Even though her argument targets "tracking theories," I will continue describing Mendelovici's argument as if it targets teleosemantics specifically both because teleosemantics is the focus of this section and because Artiga's reply defends teleosemantics from her argument.

consequences these theories face because they are unable to accommodate reliable misrepresentation. Mendelovici starts by defining reliable misrepresentation as follows:

1. Tokens of R represent some objects O as having property P (representation)
2. Most of these objects O do not have property P (non-veridicality)
3. Tokens of R do (or would) non-veridically represent objects O as having property P in the same circumstances on separate occasions (reliability). (2013, p. 423)

These are the conditions under which a (type of)[52] representation R counts as reliably misrepresenting the world. It does seem at least intuitively conceivable that such reliable misrepresentations are possible. Mendelovici specifies that cases of reliable misrepresentation are cases in which the representation usually "tracks" some *other* existing property – just not the property it represents (Mendelovici, 2013, p. 424).

To illustrate her point, Mendelovici uses the example of standardised aptitude tests (SATs), which are tests administered to evaluate and measure scholastic aptitude. She says that while SATs *are supposed to* detect (represent) scholastic aptitude, it can be argued that they reliably detect something else, such as hours of preparation for the test and parents' income. The difference between a measurement's accuracy and its precision clarifies the distinction between what the SATs *should* ideally measure and what they *do* (reliably) measure. A measurement is precise if the device used to measure gives the same results every time; accuracy, on the other hand, is about how close the measurement is to an idealised theoretical value. In other words, there is a difference between the reliability of a measurement and whether it represents what it is measuring accurately.

[52] The notion of reliable misrepresentation necessitates talk of *types* of representations because it is impossible to define "reliability" without being able to group together some collection of token representations; it is the collection that can be said to reliably represent or misrepresent.

The SATs are reliable – they give reproducible, repeatable results every time they are administered – but they are inaccurate as indicators of scholastic aptitude. This sheds light on why Mendelovici believes that reliable misrepresentations usually track something else: It is difficult to obtain reliably repeatable results without the existence of *something* being tracked. Importantly, since reliable misrepresentation usually tracks something, their existence can easily be beneficial to the organism; it can be used for the purposes of survival.

Why can't teleosemantics account for reliable misrepresentation? Artiga sums up Mendelovici's argument:

> Since teleosemantics claims that the representational content of R is determined by the state that existed in the past and had a causal influence in the selection of the representational system, it seems to be committed to the usual existence of the represented state. Only something that existed could have had this causal influence. But if the represented state has usually existed in the past, then the representation has been true most of the time. So there seems to be some tension between the teleosemantic view and the existence of systematic misrepresentation. (Artiga, 2013, p. 268)

This is a somewhat simplified version of Mendelovici's argument, which is intricate and difficult to parse at times. However, the basic tension is clear: Teleosemantics says that what determines representational content is some property whose detection was beneficial to an ancestral organism. Within teleosemantics, misrepresentation is defined as a failure of the evolutionary function of representational systems. This setup leaves no room for the development of reliable misrepresentation, which while inaccurate, can be beneficial to the organism. If a property is represented in a stable manner and this is useful to the organism, leading to the representation being selected for, that representation cannot be non-veridical within teleosemantics.

Teleosemanticists must interpret these kinds of situations as instances of veridical representation.[53]

Mendelovici uses the example of colour representation to better defend her case and to stress some unpleasant consequences a theory faces when it is unable to accommodate reliable misrepresentation. There is long-standing philosophical disagreement about whether colour representations are real or illusory.[54] It is also uncontroversial that we reliably represent the world as having colour properties – for example, we reliably represent ripe strawberries as being red. As we've seen, a consequence of teleosemantics is that we can infer that a property is real from the fact that it is reliably represented. Given the uncontroversial fact about colour being reliably represented, if we assume teleosemantics, we can infer colour realism; and this seems unacceptable (Mendelovici, 2013, pp. 437–440). To clarify, this is the structure of the reductio ad absurdum argument presented by Mendelovici:[55]

1. If property P is reliably represented, then P was instantiated at some point in the past. (consequence of teleosemantics)

---

[53] I am excluding some complex counterexamples teleosemanticists might propose because they are irrelevant to the general issue highlighted by Mendelovici. I am only going to include one somewhat obvious counterexample here to placate possible objections to the initial judgement about teleosemantics' incompatibility with reliable misrepresentation. One could argue that teleosemantics can accommodate reliable misrepresentation in cases in which a representational system was selected for one purpose, but the environment has significantly changed and it now serves a different purpose. For example, it is possible that an organism's representational system has evolved to detect and react to a predator's call. Imagine that the ancestral organism's environment contained not only predators but also prey that evolved to mimic the predator's call as a defence mechanism. It is possible that the predators have, in the meantime, become extinct, and that the descendant organisms now reliably misrepresent the calls made by prey as "There is a predator close by," even though a predator is never nearby. However, as Mendelovici points out, these cases are "unstable" and teleosemantics must accept that the representation will eventually – under evolutionary pressure – come to have the content "There is prey close by." In other words, these cases of reliable misrepresentation are temporary, unstable, and they do not seem to provide a plausible story for all possible cases of reliable misrepresentation.

[54] For an overview of the debate see Maund (2024).

[55] Mendelovici presents this argument in a different form, but I believe this presentation is faithful to the original version.

2. Colour properties are reliably represented. (from introspection and empirical investigation into the reliability of the representation)
3. Colour properties were instantiated at some point in the past. (from 1, 2)

Premise 1 is equivalent to stating that reliable misrepresentation is impossible within teleosemantics, and Mendelovici's earlier argumentation shows that this is a consequence of teleosemantics. Premise 2 is seemingly uncontroversial; we represent colours, and we do so similarly in similar conditions. She takes the derivability of conclusion 3 – which is basically realism about colour properties – to mean that 1 should be abandoned, and in turn, that teleosemantics should be abandoned. It should not be so "easy" to establish realism about a property.

Artiga argues that contrary to what Mendelovici is saying, the move from a property being reliably represented to it being instantiated is not unacceptable. He motivates his objection as follows: Within teleosemantics, a property being represented is equivalent to that property being what accounts for the producer-consumer system's natural selection. So, if teleosemantics is true, the fact that a property is represented implies that in the past, that property accounted for the selection of the producer-consumer mechanism. Given how representation is defined within teleosemantics, it is warranted to conclude that if a property is represented, it was instantiated (at least at some point in the past) (2013, p. 275).

Artiga completes his objection by arguing that the idea that a property is instantiated whenever it is represented only seems unacceptable because of internalist intuitions about a priori access to representational contents (Artiga, 2013, p. 277). If we assume that we have a priori (e.g. introspective) access to the contents of our experience, then it would be problematic to extrapolate from contents to what is real; it would be unwarranted to conclude something about the world from a priori facts. But teleosemantics (along with other externalist theories of mental content) is not committed to this internalist idea about privileged a priori access to content. Teleosemantics

views content as something that can be discovered through empirical investigation: Content is, by (teleosemantic) definition, determined by the conditions that account for the evolutionary success of the development of a representational system (Artiga, 2013, p. 278). What those conditions were is a matter of empirical fact, and establishing the empirical facts is not "easy."

A consequence of the teleosemanticist's attitude towards content is that introspection could systematically yield wrong results about the contents of our experience. A teleosemanticist can consistently say that it is possible that we will discover that we do not, in fact, represent colour properties at all. Since teleosemantics allows that introspective, a priori access to what we represent could be systematically wrong, it can also allow that F being represented implies that Fs existed at some point. For a teleosemanticist, premise 2 is not obvious and its truth cannot be derived from introspective considerations. However, if we were to establish that premise 2 is true – through empirical investigation of the conditions under which the representation in question has been selected for – then the inference to realism about colour would be warranted.

Artiga's reply can be summed up as follows: The supposed negative implications of teleosemantics Mendelovici points to are not problematic. While he agrees with Mendelovici about teleosemantics implying that reliable misrepresentation cannot occur and that a property being represented implies realism about that property, he disagrees about these implications being an issue. He diagnoses Mendelovici's discomfort with these implications as stemming from an additional assumption she makes, which is that we can access representational contents introspectively and a priori.[56] This is the assumption that is incompatible with teleosemantics, and

[56] Artiga and Mendelovici disagree on the aprioricity of introspective knowledge, which leads to some confusion in their back-and-forth.

Artiga believes we can do without it. In other words, Artiga and Mendelovici agree about the implications of teleosemantics. Their disagreement lies elsewhere.

Mendelovici objects to several details in Artiga's interpretation of the dialectics of her argument. I will not continue recounting the details of this exchange here. What is important is the following: Mendelovici believes that the derivation of conclusion 3 (colour realism) from premises 1 and 2 is a sort of reductio ad absurdum of teleosemantics, while Artiga does not. Mendelovici says that we would not normally go about establishing colour realism by investigating whether colour properties are reliably represented ("without examining the world" for colour) (2016, p. 79). The essence of Mendelovici's argument is that this is simply the wrong way to go about determining realism about a property.

An ancillary part of Mendelovici's general argument against teleosemantics is that any theory of content should accommodate the fact that introspection does, at least some of the time, uncover what the contents of our representations are. Once we add this assumption to the argument, it clarifies why she is opposed to inferring realism about a property from that property being represented: If introspection delivers representational content, we could conclude that a variety of properties are real just by introspecting. Introspection is not usually the way realism about a property is established. While Mendelovici believes introspection is the primary way a subject goes about finding out what they are representing, she broadens her position to accommodate opponents who might be less convinced of introspection's import: There must be *some* way of pre-theoretically accessing contents (2016, pp. 82–85). Absent a way of pre-theoretically determining what content a mental state has, it's unclear why many of the problems plaguing theories of representation would be viewed as problems at all. Why is the disjunction problem a *problem*? The answer, Mendelovici argues, is that we have some way of pre-theoretically determining what contents are, and that pre-theoretical assessment tells us that contents are

determinate (2016, p. 88). This pre-theoretical intuition justifies the assumption that contents are not wildly disjunctive. In any case, whatever the method for pre-theoretically accessing representational contents is, she argues, it is unlikely that it allows us to automatically determine whether the thing represented is real. So, if we have pre-theoretic access to representational content, the move from a property being represented to it being instantiated out in the world is unwarranted. If the teleosemanticist rebuts by saying we don't have pre-theoretic access to representational content, Mendelovici argues that it is dubious that the object of their inquiry is representational content at all.

The issue here seems to stem from disagreement about the object of inquiry for a theory of representational content. This disagreement, in turn, depends on the methodological stance theorists of representation choose. Mendelovici states that a theory of representational content should explain the phenomenon we usually pre-theoretically associate with representation, such as the content we can access via introspection or through common-sense psychology. Teleosemanticists state that a theory of representational content should explain the evolutionarily selected-for capacity of organisms to detect features of their environment (and their capacity to use the results of detection to their advantage). These two explananda do not necessarily coincide.[57] One stance prioritises the pre-theoretical access we have to representational content, while the other does not. Which one, then, is the stance that a "real" theory of representation should take?

A strong reason to be sympathetic to Mendelovici's appeal to pre-theoretical considerations is that a lot of philosophising about representational content begins with appeals to pre-theoretical

[57] Someone might object that we could discover, a posteriori, that the two ways of individuating representational content coincide necessarily, which would mean that they are equivalent. However, if Mendelovici's argument about reliable misrepresentation is sound, it also demonstrates that the two approaches to representation cannot coincide, since one is able to accommodate the possibility of reliable misrepresentation while the other cannot.

knowledge about content, and even more of it assumes it implicitly. Whatever the flavour of a specific theory of representation, philosophers will begin their arguments by stating that mental states (or linguistic entities) are about or represent things by gesturing towards cases that are intuitively obvious, such as the fact that we can talk about tables, perceive strawberries, and think about cats. To list some notable examples,

> A ubiquitous feature of mental states is that they have content: that is, they represent features of the world. When I see a tree, my perceptual state represents the tree. When I believe that the Earth is round, my belief represents a state of the Earth. (Chalmers, 2002, p. 473)

> Karl experiences a tomato of a round and somewhat bulgy shape. He believes that rabbits are getting into his garden. He worries that democracy is in trouble. He is confident that 68 plus 57 will always make 125. […] By what constraints, and to what extent, do the non-intentional facts about Karl determine such intentional facts? (Pautz, 2021, p. 263)

> The experience of running downhill. The desire to drink coffee in the morning. The belief that everything is grounded in the physical. The intention that guides your hand as you reach out to grasp the door handle. The words 'all emeralds are green'. All these phenomena have the spooky feature of 'aboutness'. (Williams, 2020, p. 1)

We can contrast the above quotes that (apparently) rely on the pull of first-person experience with pre-theoretical considerations that also (apparently) rely on common-sense or "folk" psychology:[58]

> I have, as it happens, a strikingly intelligent cat. […] There's quite a lot of Greycat's behavior that I want to explain by adverting to the way that Greycat takes the world to be; how he represents things. For example: It's part of my story about why Greycat turns up in the kitchen in the morning that Greycat has a story about his bowl; and that, in Greycat's story, the bowl figures as – it's represented as being – a likely locus of food. (Fodor, 1987/1993, pp. ix–xi)

[58] The distinction between examples that rely on first-person experience and examples that rely on folk psychology is not clear-cut, as can be noticed; examples are often mixed and seem to rely on both kinds of pre-theoretical judgement. The important thing to note is that these pre-theoretical judgements are taken as a starting point of theorising about representation.

> When we describe any of our actions, or actions of other people, it is difficult to do so without talking about representations: I opened the window because I thought it was too hot and I did not want to turn on the air conditioner. (Nanay, 2022, p. 75)

> "Mental content" is a technical term for a very familiar phenomenon. Suppose that Groucho believes that there is money in the safe, Chico hopes that there is money in the safe and Harpo assumes (for the sake of the argument) that there is money in the safe. […] The fact that we have mental states with content is a crucial fact about our minds. (Schulte, 2023, p. 1)

These considerations are the starting points of theorising, both literally – these quotes are all found in introductory paragraphs – and in a more conceptual sense: The fact that many of our mental states have specific contents is taken as an obviously occurring phenomenon that requires further explanation. The theoretical framing is the following: *It is a fact* that we can think (or have other kinds of mental states) about, e.g., cats. *It is a fact* that we view others as thinking, perceiving beings who react to their environment accordingly. This is taken to mean that there is such a thing as contentful mental states. A theory of content should explain this fact; it is the starting point of inquiry.

Not only does the initial framing of the issue suggest that we have pre-theoretical access to the contents of our representation, but this kind of access is often assumed as constraining theories in other ways. For example, the issues related to the extensional requirement – that is, the disjunction and the distality problems – are partly seen *as* problems because they clash with our pre-theoretical intuitions about the contents of representations. Without taking pre-theoretical intuitions into account, it becomes unclear why "wildly disjunctive content" is "clearly unacceptable" (Schulte, 2023, p. 22). For example, why is it clearly unacceptable that a subject could have a representation with the content "cow or large horse" when in front of a cow?[59] It

[59] Of course, the disjunction problem is also linked to the issue of misrepresentation – if content is disjunctive enough, it ceases to allow for cases of misrepresentation. This is not the part of the issue that is relevant to the example I made here.

seems that the justification for this claim comes from pre-theoretical intuitions about what the contents of representations are and should be. The same goes for the distality problem: It is presupposed that it is an issue when a theory cannot individuate the "right" content among a long chain of more distal and more proximal causes. But based on what criteria do we determine what the right content should be in the first place? What allows us to select one element of the causal chain as the point of reference and label others as "too distal" or "too proximal"? It seems that such constraints are set based on pre-theoretical intuitions.

The naturalistically minded philosopher will have sympathies for the way the teleosemanticist individuates its object of inquiry. To repeat, the starting point of inquiry for teleosemantics is that a theory of representation should explain the capacity of organisms to detect features of their environment and react to them. Two motivating reasons for adopting this methodological stance stand out. First, a naturalistically minded philosopher will assume that whatever representation is, it must have a close relationship to the world; it must (in some way) present the representing subjects with features of the environment. For this to occur, the notion of representation must be explicated through a relation that obtains between the representing subject and features of the world. Teleosemantics (and other tracking theories that Mendelovici criticises) assumes that this relation obtains and that it has "tracking" properties, that is, it serves the purpose of tracking features of the environment.

A second reason one might prefer this stance is the hope that the philosophical notion of mental representation will one day be fully integrated with the notion of mental representation used in cognitive neuroscience. In cognitive neuroscience, the notion of mental representation is

used in a somewhat heterogeneous manner,[60] but it is generally accepted that representation enables information processing (see e.g. Gazzaniga et al., 2019, pp. 74–78). The goal of integrating the two homonymous uses of "mental representation" found in philosophy and the cognitive sciences will, for hopeful naturalists, have higher priority than other desiderata for a theory of representation. This leads to attitudes within cognitive neuroscience being adopted into philosophical theories of representation. For example, Neander is explicit about the fact that she adopts cognitive science's stance that "much of the information processing involved in perception, memory, learning, linguistic comprehension, decision-making, and the like is inaccessible to consciousness and yet is representational" (Neander, 2017, p. 4). If cognitive science has already severed the link between conscious access and representation, that is one reason to adopt a similar stance. However, it is less obvious that the cognitive sciences have given up on third-person pre-theoretical assessments of representational content such as the ones licensed by common-sense psychology.

The question of how much weight should be given to pre-theoretical judgements about content is complex. On one hand, the scientifically minded philosopher is warranted in replying that very little weight should be given to pre-theoretical judgements about content, since it is unclear that common sense has any bearing on scientific success. Empirical research into the nature of things rarely takes pre-theoretical judgements into account, and empirical findings can override such judgements. Humans can be and have been deeply mistaken about the nature of things they encounter in the world. However, it can be questioned whether complete severance

[60] For example, cognitive neuroscientists do not seem to be particularly sensitive to the philosophical issue of the indeterminacy of content when they discuss what organisms represent. Neander describes how researchers shift from, e.g., talk about toads representing prey to representing worm-like stimuli without acknowledging that those are seemingly different contents (2017, pp. 115–116).

between conscious access to representation and representation itself is methodologically warranted. Without settling these methodological questions, it is not clear how the theoretical target of theories of representation is individuated, and it is even less clear whether different theories are talking about the same phenomenon. The risk is that without paying close attention to methodology, theorists might be talking past each other. In any case, the peculiar nature of content warrants a careful questioning of the approach one should take when researching it.

4.4 EXTERNALISM, INTERNALISM, AND RATIONAL THOUGHT AND BEHAVIOUR

Much insight on the issues raised by the conversation between Mendelovici and Artiga can be borrowed from the long-running (but still ongoing) internalism-externalism debate about mental content. This section's aim is to show that teleosemantics suffers from the same issue that afflicts other externalist theories of mental content: the inability to adequately explain rational thought and behaviour. To do so, I will adapt an argument proposed by Boghossian (1994), who showed that subjects having access to the contents of their own mental states is a necessary precondition to our standard practice of explaining rational thought and behaviour.

There are two basic ways to think about content: as determined exclusively by properties of the individual who is undergoing the relevant mental state (internally), or as also being determined by properties of the environment (externally). These two supposed kinds of content are called narrow and broad content. The disagreement is over whether intentional mental states such as belief should be characterised broadly or narrowly (or both). The internalism-externalism

debate is often framed through the incompatibility between two plausible principles, which we may call "privileged access" and "external access":[61]

(**PRIV**): We have "armchair" access to the contents of our mental states.[62]

(**EXT**): The content of our mental states is (at least partly) determined by the environment.

(PRIV) and (EXT) are both plausible but are prima facie incompatible. The incompatibility arises because if (EXT) is true, access to contents must sometimes be mediated by empirical investigation of the environment and so ceases to be "armchair." (PRIV) is intuitively plausible from a first-person perspective. (EXT) is plausible because we can construct Putnam-style Twin-Earth scenarios regarding mental content that show that subjects who are exact physical copies of each other may have different mental contents due to differences in their environment. The connection between the incompatibility between (PRIV) and (EXT) and the question of what methodological stance we should choose with regards to representational content is obvious. Just like (PRIV) and (EXT) cannot both be true, a theorist of representation must choose between giving priority to either "armchair" access to representation – Mendelovici's way[63] – or to the idea that representational contents are determined by external factors and may contradict our armchair judgements.

---

[61] For examples, see Parent (2024) and Ebbs (2025).

[62] "Armchair" is an intentionally imprecise term that denotes various types of access; whatever that access is, it can be negatively characterised as *not* going out in the world and examining the environment. The various interpretations of the metaphorical armchair can mean a priori, first-person, or introspective access. The relationship between these different ways of understanding armchair access is subject to debate. I hope to sidestep this debate through the imprecision of the umbrella term "armchair."

[63] Mendelovici believes both introspection and common-sense psychology provide pre-theoretical access to contents. However, given her position about content, she must strongly prioritise introspective access. I propose an argument demonstrating that Mendelovici must make this methodological commitment in Section 5.3. As will become apparent soon, even though the internalism-externalism discussion is not a direct analogue to the discussion between Artiga and Mendelovici, it leads to very similar conclusions.

One can find an almost direct analogue to Mendelovici's argument against teleosemantics in Boghossian's (1997) argument against those who maintain that there is no tension between (PRIV) and (EXT). To summarise, Boghossian (painstakingly) argues that joint acceptance of (PRIV) and (EXT)[64] leads to an unpalatable consequence: it puts subjects in a position to conclude that something is *real* on the basis of armchair access to the contents of their mental states (Boghossian, 1997, pp. 165–166)[65] (notice that this is quite similar to one of Mendelovici's conclusions, which is that teleosemantics is bound to accept realism about colours if they are represented). The unpalatability of the consequence gives support to the idea that (PRIV) and (EXT) are incompatible and that one of these two principles should be abandoned. Boghossian's argument can help us sharpen the methodological assumptions that underlie the dispute between Mendelovici and Artiga. With her argument, Mendelovici highlights the existence of this tension and concludes that teleosemanticists are forced to abandon (PRIV) (or something similar to it). Teleosemanticists will probably not be persuaded by this kind of argument, as the back-and-forth between Mendelovici and Artiga demonstrates, because they do not feel compelled by (PRIV).

If one is not persuaded by (PRIV)'s intuitive force, there is a more pointed reason why abandoning it might be unappealing: it goes against certain basic assumptions about the role contents have in explaining behaviour and rational thought. In a different paper, Boghossian outlines two principles of transparency that are a consequence of (PRIV), after which he demonstrates that they are incompatible with (EXT):

---

[64] Boghossian puts (1) and (2) in different terms – (1) is "privileged self knowledge" and (2) is "externalism" (1997, pp. 161–162). He also emphasises the a priori status of the presumed privileged access to mental content, unlike Mendelovici.

[65] Boghossian speaks of both contents and concepts, taking concepts to be the relevant kind of mental content in his examples.

> ***Transparency of difference***: If a subject has two mental states with different contents, the subject should be able to know, via "armchair" access, that they are different.
>
> ***Transparency of sameness***: If a subject has two mental states with the same content, the subject should be able to know, via "armchair" access, that they are the same. (Boghossian, 1994, p. 36)

(I modified Boghossian's use of "a priori" into "armchair" to preserve consistency and to allow a somewhat more expansive understanding of (PRIV)). Externalism is easily shown to be incompatible with these two principles. The real problem with this, Boghossian claims, is that transparency is necessary for explaining rationality and rational behaviour. This is of interest to us because explaining rationality and rational behaviour are key goals of naturalistic theories of representational content.

Rejecting transparency blocks us from describing intuitively rational thought and behaviour as irrational and vice versa. The example Boghossian uses to illustrate this is simple. Imagine that we reject transparency and that content is externally individuated. Imagine an apple that has a hole on one side. Jane looks at the apple and has a thought with the broad content "$Apple_1$ has a hole." Now imagine that she later looks at that same apple from a different side, failing to recognise it; she has a thought with the broad content "$Apple_1$ has no holes." Jane has two blatantly inconsistent beliefs. However, we don't want to say that Jane is irrational; in the scenario we just described, she makes no *errors in reasoning*. Her beliefs are, intuitively, perfectly rational. Jane merely lacks empirical information that would allow her to discover that what she believed were two apples are one and the same apple (Boghossian, 1994, p. 41).

We can expand this example to see how lack of transparency also impacts explanations of behaviour. Psychological or behavioural explanation depends upon subjects following certain logical principles that mirror those of rational thought. Rational thought and rational behaviour

are interconnected phenomena. To see this, imagine that Jane only wants to eat apples with no holes; so, if an apple has a hole, she does not want to eat it. She approaches $apple_1$ facing its good side and takes a bite out of it. Given that she has a previously formed belief with the broad content "$Apple_1$ has a hole," and given her desire to only eat hole-less apples, her behaviour can now be described as irrational. But her behaviour is clearly not irrational. To be able to describe her behaviour as rational, we must postulate that she has immediate access to the contents of her thoughts, and that those contents are not broad.

We may apply these insights about the importance of transparency to the issue of whether representational content should be "armchair" accessible. Remember Fodor's comments about Greycat the cat:

> There's quite a lot of Greycat's behavior that I want to explain by adverting to the way that Greycat takes the world to be; how he represents things. For example: it's part of my story about why Greycat turns up in the kitchen in the morning that Greycat has a story about his bowl; and that, in Greycat's story, the bowl figures as – it's represented as being – a likely locus of food.

Much of what naturalistic theories of representation aim to do is provide a notion of representation that is useful in explaining behaviour. While Fodor's passage makes no mention of rationality, its role is implicit: Greycat's "stories" (representational states) only explain his behaviour (walking towards the bowl) if the stories can be interpreted as a series of rational steps taken towards the obtaining of a goal (eating). A simple such explanation goes as follows: Greycat desires food. Greycat believes that the bowl in the kitchen is a source of food. Greycat reasons that if he is to get the food he desires, he should move towards the bowl.[66] For this series of steps to explain

[66] The use of "desire," "belief," and "reason" is here somewhat metaphorical; I do not mean to imply that cats can believe or reason the way humans do. However, cats must be able to *somehow* connect two representational states that are roughly equivalent to "I desire food" and "Food is in the bowl" in order to be motivated to move towards the bowl (in other words, in order to have a reason to move towards the bowl). This is the sense in which I am stating "reasoning" is involved.

Greycat's behaviour, we must imagine that Greycat's representations are *accessible* to him in some manner; the series of steps must "make sense" from the cat's perspective. To say this in a less metaphorical way, the organism's perspective is an integral part of what makes it possible for us to adequately describe its behaviour. Perspective is important because it is a crucial part of understanding what organisms can reason with – what they have access to – to obtain their goals.

Teleosemantics identifies representational content with (very) broad content; whether an organism's representation has a certain content is determined by properties of the environment, including properties of a distant, past environment. As we've seen, this means that teleosemantics is incompatible with both (PRIV) and its implications regarding transparency, and subjects need not have armchair access to the contents of their mental states. The rejection of (PRIV) does not in and of itself bother teleosemanticists. Millikan, for example, is explicit about teleosemantics abandoning transparency (1984, pp. 13, 91). However, rejection of (PRIV) and transparency may mean that teleosemantics struggles with characterising the right behaviours and thought patterns as rational or irrational, which is a problematic consequence for any naturalistic theory of content. I will now attempt to argue for this by adapting Boghossian's example.

Assume teleosemantics is true. Imagine an organism O, whose environment contains apples, which are an important source of nutrition for Os. However, only non-holey (whole) apples contribute to Os' fitness, meaning that Os dislike and avoid eating holey apples. The ability to detect, represent, and consequently eat "good" apples led to O's ancestors proliferating; so, according to teleosemantics, O can and often does have beliefs with the content "this is a nutritious

thing."[67] Os can also recognise and avoid holey apples, which also contributed to their fitness. They often have beliefs with the content "This is not a nutritious thing." Imagine an O seeing a specific holey apple, which we can name $apple_1$, and forming a (true) belief with the wide content "$Apple_1$ is not a nutritious thing." After a while, O sees that same apple from an angle which obscures the hole and forms a (false) belief with the wide content "$Apple_1$ is a nutritious thing." Unbeknownst to O, what it thought were two different objects are one and the same object. But now we are forced to say that O holds two obviously contradictory beliefs at the same time: that $Apple_1$ is nutritious and that it is not nutritious. O may also attempt to eat the apple, going against its belief that it is not nutritious. O's approaching and eating of the apple *makes sense* given O's perspective but reads as irrational if we describe O's beliefs as only having wide contents. O's behaviour becomes difficult to explain and rationalise in a satisfactory way if we do not introduce a narrower facet of the content of O's thoughts, namely the kind of narrow content that accounts for O's perspective.

This inability to describe behaviour and thought processes adequately is an issue for teleosemantics, one that goes beyond the intuitive unpalatability of losing privileged access to mental contents. While one can plausibly argue that a naturalistic theory of representation does not have to account for first-person access to content, it is much more difficult to argue that it is unconstrained by the intuitive explanatory role content has in rational behaviour and thought. "Common sense" or "folk" psychology involves explanations of behaviour in terms of *reasons,* that is, in terms that allow for the rationalising of observed behaviour. If the notion of content defined

---

[67] The exact content is something *like* this – not "apple" – because of the way Millikanian, consumer-oriented teleosemantics individuates contents based on normal functions; this does not affect the point made in my argument. The argument can be modified to accommodate producer-oriented content ascriptions as well.

by teleosemantics does not serve to explain either intuitive first-person access to content nor intuitive third-person attribution of content based on common-sense psychology, it risks losing all connection to pre-theoretical notions of content and representation; it becomes unclear what its object of inquiry is.

One could argue that teleosemantics serves to explain the information-processing capacities of organisms, that it, the capacity organisms have to detect and then process information from their environment. However, this objection is unsatisfying because it bypasses the reason why representation is linked to information processing in the first place. Organisms' capacity to detect and process information serves a further purpose, that is, it is *useful* to representing organisms. (Consumer-oriented teleosemantics recognises this fact and builds it into the theory – an even stronger reason to reject this objection). It is impossible to describe the usefulness of information gathering without saying that the collected information presents organisms with *reasons for action.* I brought an umbrella *because* I believed it was raining; the cat ran away *because* it saw a dog. The use organisms have for information processing is intimately linked to behaviour that we want to describe as making sense, as rational. As we've seen, the content that figures in the intuitively accurate descriptions of rational thought and behaviour is not wide. So, not only does teleosemantics fail in accurately describing rational thought and behaviour, but it *should be able to* accurately describe rational thought and behaviour.

The tension between (PRIV) and (EXT) persists, even if one is uninterested in accounting for armchair access in and of itself. That's because (PRIV) has direct consequences regarding transparency that cannot be abandoned if we want to accurately describe representing subjects' rational thought inferences and behaviour. As I've argued, a naturalistic theory of representation such as teleosemantics cannot plausibly decouple itself from the goal of accurately describing representations' involvement in rational thought and behaviour.

CONCLUSION

Causal and teleosemantic theories of mental content face persistent challenges that put their viability into question. Two kinds of problems plague these theories: the issue of allowing for misrepresentation and the problem(s) of indeterminacy. These two kinds of problems mirror the normative and the extensional constraint posed by Kripke's sceptical challenge. I traced the historical development of causal and teleological theories through their attempts to resolve the issues they face.

Contemporary teleosemantic theories, which represent the most popular naturalistic theories of mental content at this moment, have not yet definitively resolved their issues with indeterminacy. Teleosemantic theories also face an additional problem, which they share with other externalist theories of mental content: They cannot adequately describe rational behaviour and thought. This issue is related to the inability of externalist theories to account for the first-person perspective that is crucial for our common-sense understanding of mental content and its role in thought and behaviour. Teleosemantic theories, then, end up being unable to match both common-sense third-person and first-person ascriptions of mental content, suggesting they may describe something other than genuine mental content. The accumulated objections to naturalistic theories over the decades make a compelling case for exploring non-naturalistic alternatives.

5. PHENOMENAL THEORIES OF CONTENT

INTRODUCTION

Naturalistic theories of both mental and linguistic content face serious difficulties, as has been suggested in Chapters 3 and 4. Because of these difficulties, optimism about the naturalisation of representational content has gradually waned. This situation has led to the recent rise of an alternative research project that focuses on the phenomenal aspect of representational mental states: the phenomenal intentionality project. In this work, I will refer to the disparate theories that loosely adhere to this research project as Phenomenal Intentionality Theories (PITs). The basic idea is that intentional (i.e. representational) properties of mental states are best understood as inextricably linked to phenomenal properties, that is, to the way it's like to be in a mental state. PITs state that the phenomenal aspects of experience are more fundamental and so explain and/or determine the representational properties of at least some kinds of mental states. In other words, PITs aim to answer the CDQ by proposing that mental states have representational properties in virtue of their phenomenal properties.

One significant feature of PITs is their internalist approach to content. Since PITs argue that phenomenal properties determine intentional content, and since phenomenal properties can only be accessed in first-person, an externalist version of PIT would be incoherent. Defenders of PIT are usually explicit about their defence of content as fundamentally *narrow* (see, e.g., Farkas, 2008, p. 274; Horgan & Tienson, 2002, pp. 526–527; Kriegel, 2013, p. 5); broad content, if it exists, is derived from narrow content.

PITs are not necessarily non-naturalistic, since phenomenal properties might very well turn out to be naturalisable. However, they are compatible with a non-naturalistic outlook, and unlike their predecessors, they do not explicitly adhere to the goal of naturalisation. This neutrality is a significant break with tradition. The underlying hope is that a theory of representation will have better chances of success – better than its predecessors, at least – if we *avoid* the explicit goal of naturalisation. The rest of this chapter serves to assess whether PITs have better chances of providing a plausible story about the foundations of representational content compared to their adversaries.

The chapter is divided into the following sections. Section 5.1 provides a minimal history and categorisation of theories that fall under the PIT umbrella. PITs are internalist theories of mental content, they are ontologically neutral, and they purport to succeed where their competitors have failed: they promise intuitively adequate, determinate mental contents. Section 5.2 considers how Kripkenstein's paradox affects PITs. The conclusion is that PIT-like theories do not possess the tools to resolve the sceptical challenge. In Section 5.3 I propose an argument that shows most versions of PIT have unacceptable consequences regarding the possibility of establishing identity between contents. The argument relies on the fact that PITs must grant that subjective reports about the intentional contents of one's perceptions are true. Given this assumption and given certain facts about the relationship between identity and indiscriminability, PITs lead to a contradiction.

5.1 THE HISTORY, PLACE, AND SCOPE OF PHENOMENAL INTENTIONALITY THEORIES

The phenomenal intentionality research programme has its official beginning in two papers that explicitly make the connection between the phenomenal and the intentional: Terry Horgan and John Tienson's "The Phenomenology of Intentionality and the Intentionality of Phenomenology" (2002) and Brian Loar's "Phenomenal Intentionality as the Basis of Mental Content" (2003). There have been many direct and indirect intellectual precursors to what the programme eventually developed into, but the relevance of this kind of history is limited here; I will restrict the scope of this section to the work that was kicked off by the two aforementioned papers. This section will summarise the basic goals, assumptions, motivations, and methodologies that prop up the programme. The summary will be brief and, consequently, it cannot do justice to the intricacies of the actual argumentation devised by supporters of PITs; the aim here is to get a sense of PITs' *place* in the space of theories of content. The ultimate purpose of this exercise is to frame the relevance of PITs to the aims of this thesis, that is, their relevance to the search for a successful and paradox-eschewing answer to the CDQ.

The goal of PITs is to provide an explanation of the intentional – i.e., of the representational – properties of the mental. PITs aim to provide this explanation either partly or fully in terms of the phenomenal properties of experience. In other words, our mental states represent *in virtue of the fact* that they have a certain phenomenology. This basic idea is supported by the intuitive pull of our first-person experience: When I see a strawberry, the phenomenology of this experience seems to be intimately linked to the way the strawberry is presented to me. The strawberry is presented to me – it is represented – *through* the phenomenology of the experience of seeing it. In other words, the representational content of this particular mental state seems to

be inextricably connected with the phenomenology of undergoing that mental state. Supporters of PITs believe this intuition should be given significant weight.

Not all naturalists and externalists about content deny the connection between phenomenology and intentionality; some believe, just as defenders of PITs do, that they are inextricably connected. However, they usually interpret the relationship between intentionality and phenomenology in the exact opposite way compared to defenders of PIT: They believe that phenomenal properties are determined by representational properties. This is a position called representationalism or intentionalism about phenomenal experience (Horgan & Tienson, 2002; Lycan, 2023; Pautz, 2013). Representationalists and defenders of PIT agree that there is a significant connection between the phenomenal and the intentional, but they disagree about which one is more fundamental. The representationalist stance, which takes the intentional to be more fundamental, may have been more appealing when philosophers were still optimistic about naturalising intentionality. However, as we've seen, this optimism has somewhat waned. Since the naturalisation of the intentional is not on the horizon, explaining the phenomenal via the intentional does not seem as demystifying anymore.

While all PITs argue that intentional properties depend on phenomenal properties in at least some cases, they differ in two main respects: how the dependence relation is spelled out and what subset of intentional properties is dependent on phenomenal properties. The dependence relation can be cashed out in terms of supervenience, grounding, or identity; and the theories can purport to explain the intentional properties of either some or all mental states. All defenders of PITs agree that perceptual states have their intentional properties in virtue of their phenomenal properties. Visual perception of colour is usually taken as the good paradigmatic case for the dependence of intentionality on phenomenology.

The principal motivation for adopting PITs is the presumed failure of alternative theories of content. In particular, many cite the issues that alternative, naturalistic theories have with indeterminacy (Horgan & Graham, 2012; Pautz, 2013). Theories that posit that phenomenal properties determine intentional properties appear to circumvent the issue: From a first-person perspective, we seem to have no issue with individuating the phenomenal aspects of our experience in a determinate manner, so intentional properties should inherit this determinacy.

Another important motivation for adopting PITs is the idea that a theory of content should be able to account for the way content is presented to the subject, a task which externalist theories have struggled with. A consequence of externalism is that subjects do not have access to the intentional contents of their mental states. In contrast with externalist theories, PITs present a story that renders intentional contents accessible to subjects. Given their belief in the phenomenological basis of intentionality, and given the immediate subjective access subjects have to the phenomenal properties of their experience, it is easy to see that intentional contents will also be accessible (this is a double-edged sword; PITs must deal with the issue of non-conscious but still intuitively representational mental states). The first-person accessibility of intentional content is the key methodological assumption that underlies PITs and that starkly distinguishes them from alternative theories of intentionality. The clash between theories that do and do not assume subjects have conscious access to content can be seen in the conversation between Mendelovici and Artiga, which was recounted and analysed in Sections 4.3 and 4.4.

There are several standard arguments for PITs which flow from these initial motivations. The first one, which has already been mentioned, is based on the introspectively supported intuition that there is something about phenomenal experience that seems to be inherently representational and vice versa. Conscious phenomenal experiences always present the world as being a certain way – the phenomenal is intentional; and representational mental states are always

accompanied by a phenomenal experience (or so the story goes) – the intentional is phenomenal. All PITs are committed to the first of these two directions, that is, to the fact that the phenomenal is inherently intentional.[68]

The most vivid argument that supports the idea that the phenomenal character of experience always presents the world as being a certain way relies on our examples of visual perception. Imagine an experience of blue-ness (e.g., what happens when you look up at the sky during a sunny day). The way it feels like to undergo that experience seems to immediately present us with something, that is, with the blue quality of the sky. We have reason to believe that what is presented is an intentional content because our mental state is assessable for accuracy: It makes sense to wonder whether the sky *is* blue, and my perception of the sky as blue can be mistaken. In other words, my perception of the sky as blue seems to happen in virtue of the phenomenal experience I undergo, and the entirety of the experience is representational because it can misrepresent. Predictably, these kinds of arguments work best for simple perceptual experiences and are not as intuitively plausible when it comes to more complex mental states such as beliefs or thoughts, which do not exhibit as vivid a phenomenology.

A second argument for PITs is that it resolves the previously mentioned indeterminacy issues alternative theories of content struggle with. If phenomenal properties determine intentional properties, and phenomenal properties are sufficiently determinate, it seems that intentional properties could inherit that determinacy as well. It is usually assumed that phenomenal properties are *given* to the person having the experience, and that they are given in a determinate manner: We

[68] PITs can accommodate the existence of intentional states that do not have their intentionality in virtue of their phenomenology via, for example, the notion of derived intentionality.

do not usually have doubts about the qualitative feel of our experiences. Terry Horgan and George Graham have recently spelled out their support for this idea:

> "Not only is the phenomenal character of experience inherently intentional, but it is also inherently determinately intentional… You know what you are thinking and what you mean by your utterance, and there is a determinate fact of the matter about what you are thinking and what you mean by your utterance, because there is something it is like to think a determinate thought and to make an utterance that expresses that thought." (Horgan & Graham, 2022, pp. 152–153)

The argument is that PITs have an edge over competitor theories of intentional content due to their ability to deal with content indeterminacy.

Another standard argument for PITs flows from the debate between internalists and externalists about mental content. As has been covered in more detail in Section 4.4, there is a tension between two plausible principles:

(**PRIV**): We have "armchair" access to the contents of our mental states.

(**EXT**): The content of our mental states is (at least partly) determined by the environment.

(PRIV) is subjectively intuitive, while (EXT) is justified by appeal to Twin-Earth style thought experiments. The two principles are prima facie incompatible.[69] Faced with this dilemma, internalists reject (EXT), while externalists reject (PRIV). Both sides see this incompatibility as good reason to reject one of the principles. PITs, of course, embrace (PRIV). Some defenders of PIT argue that the broad contents that underlie our intuitions about (EXT) are derived from the more fundamental narrow contents which underlie our intuitions about (PRIV); others deny there are any broad contents at all (see sec. 6.3 of Bourget & Mendelovici, 2019 for an overview of the debate about broad contents within PITs). PITs inherit the standard arguments in favour of

[69] Some have tried a compatibilist route – e.g. (Burge, 1988); it is unclear whether such attempts have succeeded.

internalism, which are usually based on implausible consequences of rejecting (PRIV), including difficulties in explaining the role of mental content in subjects' perspectives and reasoning.

To summarise, the reasons for accepting PITs are somewhat familiar if one is acquainted with other debates in the philosophy of mental content. PITs have three main advantages: They can accommodate first-person intuitions about the contents of our own mental states, they can make sense of the presumed perspective-laden nature of contents, and they do not immediately run into the indeterminacy issues that older theories struggled with. The real reason for the rise of PITs lies in the vacuum created by the undelivered promises of their naturalistic rivals. However, the existence of this vacuum does not guarantee that PIT-style alternatives can do what naturalistic theories failed to achieve.

5.2 PITS AND KRIPKENSTEIN

Before continuing our inquiry into PITs' ability to provide an adequate theory of mental content, it is prudent to consider how Kripke's sceptical paradox affects PITs. While *Wittgenstein on Rules and Private Language* (1982/1995) predates the phenomenal intentionality research programme by several decades, it contains a discussion of potential theories that fundamentally resemble PITs. As has been noted in Chapter 2, Kripke focuses on linguistic meaning in his book, but his argumentation is equally applicable to "mental meaning" – that is, to mental representation. The sceptical argument casts doubt on the possibility of a coherent theory of representational or intentional content of any kind.

Compared to his discussion of dispositional theories of meaning, Kripke pays relatively little attention to the idea that our inner experience could determine meaning. However, he does briefly discuss this possibility. The theory Kripke sketches is the following:

> Why not argue that "meaning addition by 'plus'" denotes an irreducible experience, with its own special quale, known directly to each of us by introspection? […] Presumably the experience of meaning addition has its own irreducible quality, as does that of feeling a headache. The fact that I mean addition by 'plus' is to be identified with my possession of an experience of this quality. (Kripke, 1982/1995, p. 41)

This outline of a theory – which we may call proto-PIT – has a lot in common with PITs. First, it proposes that "meaning" (representing, in the terminology used in this thesis) should be identified with an inner quality of experience. Second, it specifies that the quality in question is introspectable, that is, accessible in the first person. Third, the quality is compared to other mental states that have a distinct *feel*, like headaches or itches; and phenomenal properties simply are what makes up the *feel* of undergoing a mental state. Keeping in mind that this is clearly a rudimentary sketch of a theory, we can assume that Kripke's discussion will be relevant to PITs.

Kripke dismisses this approach to meaning as unable to meet the normativity requirement. In the interpretation of the normativity requirement used in this thesis, this means that proto-PIT cannot account for the existence of correctness conditions. In Kripke's words, "suppose I do in fact feel a certain headache with a very special quality whenever I think of the '+' sign. How on earth would this headache help me figure out whether I ought to answer '125' or '5' when asked about '68 + 57'?" (1982/1995, pp. 41–42). In other words, how could phenomenal properties by themselves determine the (in)correctness of a representation?

Presaging that PITs will focus more closely on perceptual representation, Kripke grants that in the case of "visual words" proto-PITs may initially seem more plausible; but this initial plausibility does not shield them from the same criticism levelled at the meaning of "+".

Phenomenal feel by itself does not seem to present us with correctness conditions – or at least, that is the gist of Kripke's conclusion. The argument against proto-PIT relies on an analogy between images and the phenomenal feel of an experience. Kripke says that even if every use of the word "cube" were associated with a mental image of a cube, that mental image would do nothing to determine the correctness conditions of the word "cube." A picture[70] does not represent anything prior to an intended interpretation. To appreciate this point, consider Hilary Putnam's notorious example of an ant crawling in the sand and tracing a line that turns out to look like a faithful depiction of Winston Churchill (1981/2004, pp. 1–2). Does the traced line *represent* Winston Churchill? I believe that the most common intuition is that it does not. The intuition that the line represents Winston Churchill can only be based on some kind of resemblance the line bears to Churchill. However, similarity does not seem to be either necessary or sufficient for representation. This is at least in part due to inherent differences between these two relations: similarity is symmetric and reflexive, while representation is intuitively not symmetric nor reflexive. Assuming that similarity is sufficient for representation also leads to an immediate and unpalatable proliferation of representation, since every single object would end up representing sufficiently similar objects. If we apply these arguments to the analogous case of phenomenal character, we can see that the association of certain "mental images" (in Kripke's terminology) with the representation of a cube (or of a colour, or of anything else) does not seem to help determine the representation's correctness conditions. So, the phenomenal feel of an experience does not determine its meaning, its representational properties.

---

[70] The word "picture" might be misleading here, because it already implies something has been de*picted*, i.e. represented. The more precise way of putting this is that a configuration of visual elements – lines, shapes, colours, etc. arranged in some manner – do not in and of themselves represent. Since this is an unwieldy formulation, I have decided to replace it with "picture."

We can see now that this kind of argument will not convince defenders of PITs, who believe that phenomenal properties *do* represent in some capacity, and that they do so without the need of any additional interpretation. Defenders of PITs sometimes simply assume that this is the case, but there are also arguments that defend the idea that phenomenal states are always assessable for accuracy. Charles Siewert developed such an argument defending phenomenal states' assessability for accuracy. Siewert's argument relies on the assumption that experiences with phenomenal properties present the world as *seeming* a certain way in virtue of those phenomenal properties (Siewert, 1998, p. 220). If having a phenomenal experience does lead to the world seeming a certain way to me, the way the world seems to me could be accurate or inaccurate, depending on whether the world actually is or is not as it seems to me. So, phenomenal properties automatically present us with accuracy conditions (Siewert, 1998, p. 221). Siewert gathers that this means that phenomenal properties are inherently intentional (1998, p. 227). We can notice, however, that the argument is borderline question begging: Assuming that phenomenal states always present the world as seeming a certain way is an assumption that is a little too close to the desired conclusion. If phenomenal properties inherently present us with accuracy conditions, that is simply another way of saying that they present us with correctness conditions: if a phenomenal experience is accurate, it means it is a correct representation of something; if it is inaccurate, it is an incorrect representation of something. And the crux of the issue, to cite Boghossian once again, is that "Having a meaning is essentially a matter of possessing a correctness condition. And the skeptical challenge is to explain how anything could possess that" (2002, p. 149).

The obvious Kripkensteinian objection to Siewert's argument – one that Siewert anticipates – is that phenomenal states are assessable for accuracy only once they are supplemented with something, e.g., an interpretation. What this objection presupposes is that there is a sense in which "pure" phenomenal properties do not yet present the world as seeming a certain way – a sense in which they do not inherently possess correctness conditions. And there is an

understanding of the phenomenal that adheres to this point of view: the one that views phenomenal properties as properties of our experience we can infallibly access. There is a sense in which I cannot be wrong about the way it feels to undergo a certain experience; the thing I cannot be wrong about is the phenomenal character of my experience.

There seem to be two distinct features of our experience that get called "phenomenal": the ones that present the world as seeming a certain way, and the ones that present my own experience a certain way. We can call them pure phenomenal properties and impure phenomenal properties:

> **Pure PP:** The pure phenomenal property of experience E = the "what it's like" aspect of undergoing experience E.
>
> **Impure PP:** The impure phenomenal property of experience E = what it seems like the world is while undergoing experience E

The two do not coincide because an experience E's having impure PPs means that E is assessable for accuracy in virtue of its impure PPs while the same does not hold for pure PPs. To illustrate this, we may consider the phenomenon of perceptual constancy. The simplest example concerns colour constancy. We may experience a wall *as* white throughout the day, even though the shade of the wall changes due to variations in lighting conditions. What can be gathered here is that based on my experience, the wall *seems* white; but the "what it's like" aspect of my experience changes considerably throughout the day. The impure PP of the colour of the wall stays the same in the morning and in the afternoon, since in both cases, it seems to me that the wall is white. On the other hand, the pure PP of my experience differs in the morning and in the afternoon, since the "what it's like" aspect of my experience has changed. I can be wrong about the feature of my experience that makes it the case that the wall seems white in a constant manner, but I cannot be wrong about the feature of my experience that we capture with the "what it's like" dictum. Impure PP and pure PP describe different aspects of subjective experience.

Siewert believes that phenomenal features *are* the ones that present the world as seeming a certain way; they are identical to impure PPs. A critic could object that phenomenal features do not, by themselves, present anything to the subject, and that Siewert is misidentifying impure phenomenal features as pure phenomenal features. Which of these two features of experience should be named "phenomenal" is not going to be at issue here. What is at issue, though, is whether either of these features of experience can be used to construct a paradox-proof theory of representational content. I will now show that they cannot.

Neither pure PPs nor impure PPs can provide a basis for a theory of representational content. Explaining why impure PPs are inadequate for this purpose is relatively simple. Impure PPs are already straightforwardly representational: they present the world as being a certain way. Because of this, explaining representational content through impure PPs goes against the non-circularity requirement posed by Kripke's sceptical challenge. We can immediately ask: In virtue of what do impure PP present the world in a certain manner? Until this question is answered, we have not gained any deeper understanding of representational contents. The question was: In virtue of what do representations have the contents that they do? Pointing towards phenomena that are themselves representational cannot provide an illuminating and comprehensive answer. This kind of answer is either viciously circular, leads to regress, or is incomplete.

Can pure PPs provide a basis for a theory of representational content? One argument in favour of pure PPs giving rise to accuracy conditions relies on the following idea: when I have an experience that is, for example, qualitatively red and spherical, then my experience is *accurate* only if it is caused by something that is itself red and spherical.[71] However, this kind of reasoning seems

[71] I am grateful to Anandi Hattiangadi for pointing out this kind of argument to me.

to again rely on the idea that representation arises from similarity, and specifically from the assumed similarity between experience and the thing that is being experienced. However, even assuming that a red sphere and my experience of that red sphere resemble each other, as has been argued previously, similarity does not seem to be necessary or sufficient for representation. So, the idea that a phenomenal experience is accurate if it is caused by something qualitatively similar to the experience seems fundamentally misguided.[72] As Kripke suggested, it is in principle unclear how pure PPs could explain the contents of representation. Pure PPs are uninterpreted qualitative feels, and as such, they do not have accuracy conditions. Without accuracy conditions, there is nothing in PPs that could potentially explain the emergence of correctness conditions. Because of this, a theory of representation based on pure PPs violates the normativity requirement of the sceptical challenge. I do not discount the possibility that there could be some feature of pure PPs that explains the emergence of correctness conditions; however, it is not clear to me what that feature could be. Until such a feature is found, I see no reason to be hopeful that any theory could explain representational properties via (purely) phenomenal properties.

[72] At this point, someone might object that there is an important role played by causation here: an accurate phenomenal experience is not only qualitatively similar to the thing it represents, but it is *caused* by the thing it represents. However, having argued for the irrelevance of similarity for the question of whether something exhibits representational properties, we are left with causation accounting for how phenomenal properties represent. To see some arguments for why causation is insufficient for explaining representational contents, see Chapters 3 and 4.

5.3 THE INCONSISTENCY OF PITS

In this section, I will argue that even on a very minimal characterisation of PITs, they lead to a contradiction. From PIT and some very intuitive assumptions about the possibility of identity between intentional contents, one can derive an absurd conclusion.

As has been described at the beginning of this chapter, the basic idea of PITs is that the contents of at least some intentional states[73] are constituted or determined by their phenomenal character – that is, by the "what it's like"-ness of a relevant experience. The first step PIT defenders take is to argue that some intentional states have their content in virtue of an experience *feeling* a certain way, in the sense in which seeing a colour has a specific phenomenal character. A second step is arguing that these intentional states – the ones whose content is (somehow) constituted by their phenomenal character – are more fundamental than all other intentional states, which should be seen as derived from these "original" instances of intentionality (e.g., see Mendelovici 2018, p. 22).

One way of spelling out the metaphysical constraints on PITs is by formulating them in the familiar language of supervenience. Since the assumption is that some contents of intentional states are constituted by their phenomenal character, we can state that:

> **PIT Supervenience:** Two original intentional states cannot differ in their content without differing in phenomenal character.

Or, equivalently:

---

[73] This hedge – in which only some intentional states are said to be constituted by their phenomenal character – is necessary to leave space for the possibility that some intentional states are not fully constituted by their phenomenal character, e.g. thoughts and beliefs.

PIT Supervenience: If two original intentional states have the same phenomenal character, then they have the same content.

There are two things to note here: First, this supervenience constraint is quite weak, which means it should be taken as a minimal constraint on PITs. Second, the formulation of *PIT Supervenience* requires us to refer to sameness or difference of both phenomenal character and content. This will become relevant later.

In *The Phenomenal Basis of Intentionality,* Mendelovici argues for a strong version of these theories called identity PIT. She defends the idea that every instance of original intentionality is identical to a phenomenal state. Notice that this is a stronger constraint than *PIT Supervenience*.[74] To make her case, Mendelovici argues against extant theories of intentionality – that is, against competitor theories that purport to explain "what intentionality really is" and where it arises from (2018, p. xv). The targeted theories include causal and teleosemantic theories of intentionality, which she calls "tracking theories,"[75] but her arguments, if effective, have wide-reaching consequences for many possible kinds of theories.

Mendelovici's "mismatch argument" against tracking theories is based on the following intuition: In the case of visual perception, it seems to us that we represent something like *primitive*

---

[74] This kind of formulation might raise questions as to whether identity PITs are theories that truly take the phenomenal aspect of experience to be more fundamental or prior to representation. After all, certain representationalist approaches advocate a similar position but reduce the phenomenal to the representational/intentional instead of the other way around. Mendelovici anticipates this question and answers as follows: "Advocates of PIT hold that intentionality/phenomenal consciousness has more of the characteristics we might have previously attributed to phenomenal consciousness than the characteristics we might have previously attributed to intentionality, while, presumably, advocates of representationalism hold that intentionality/phenomenal consciousness has more of the characteristics we might have previously attributed to intentionality than the characteristics we might have previously attributed to phenomenal consciousness." (2018, pp. 111–112).

[75] Mendelovici uses this notion to refer to theories which state that representations track certain kinds of (natural) properties in the environment.

colours. In these simple cases we are introspectively aware of something that corresponds to a primitivist view of colour (see Maund, 2024, sec. 2.1. for a brief overview of colour primitivism in the literature). We represent "qualitative, primitive, simple, sui generis, non-dispositional, non-relational" colour properties (Mendelovici, 2018, p. 88). As an example, looking at the sky (supposedly) results in a mental state that has primitive sky-blue as its content. However, tracking theories do not predict that primitive colours are the content of our visual representations; therefore, tracking theories should be abandoned (Mendelovici, 2018, pp. 33–44).

Mismatch argument from omission:

(P1) If tracking theories are true, then intentional state S does not have the content "primitive sky-blue."

(P2) Intentional state S does have content "primitive sky-blue." (justified by introspection)

(C) Tracking theories are false.

Mendelovici also argues for PITs via the positive counterpart of the mismatch argument: Unlike tracking theories, PIT's predictions *do* match our pre-theoretical intuitions about the contents of our intentional states (Mendelovici, 2018, pp. 88–89). This gives PIT a comparative edge over tracking theories; so, we should accept PIT.

Introspection plays a very important role in these initial arguments for PIT. Without reliance on introspection, one cannot justify the assumption that we represent something like primitive colours in basic perception cases. Mendelovici makes explicit a methodological assumption that is sometimes left implicit in alternative defences of PIT: Our theories of content must respect certain kinds of pre-theoretical intuitions over others; specifically, our *introspective judgements* about the content of (at least) our perceptual experiences must take priority over other ways of determining representational content (2018, pp. 27–28).

Prioritising introspection when it comes to the contents of our perceptual experiences leads to insurmountable difficulties for any theory of content. In particular, it can be shown that this methodological setup leads to a contradiction, which in turn leaves us with no way of accounting for sameness (or difference) of representational content. Being able to account for identity or difference between representational contents is a very basic desideratum for any theory of content. One of the most basic cases in which we want to be able to theoretically account for sameness of content is when we have, within one subject, the same type of intentional state with the same object, but numerically distinct states (e.g., over time or within a larger visual field). I will argue that PITs are ill-equipped to explain this.

To avoid unnecessary confusion related to the distinction between derived and original intentionality, we can use what Mendelovici takes to be good paradigmatic examples that support PIT: introspectively accessible perceptual states. Consider the following scenario:

Continuum: Imagine a simple continuous gradient.

The gradient above contains eight different colour shades specified in hexadecimal code, which we may call (from lightest to darkest) $s_1 \ldots s_8$. Most of these adjacent shades are indiscriminable to the human eye. The exceptions are s_1 and s_8 (the lightest and the darkest shade respectively), which are clearly discriminable, as shown by the "crease" in the lower right quadrant of the circle. What

this tells us is that, if s_1 and s_2 are indiscriminable, and s_2 and s_3 are indiscriminable, that does not mean s_1 and s_3 are indiscriminable; indiscriminability is not transitive.

This kind of example is reminiscent of Timothy Williamson's work on vagueness and his observations about the nature of (in)discriminability, and in particular his remarks on the relationship between indiscriminability and identity (1994, pp. 178–179). It is well-known that there are cases in which subjects cannot discriminate between different stimuli on a continuum, even though the differences between them "add up" to a discriminable difference (1994, p. 69). This tells us that the relation of indiscriminability is not transitive, as has been demonstrated by the example above, and since identity *is* transitive, indiscriminability is simply not sufficient for establishing identity.

Bringing this back to PIT and sameness of content, it seems as if PIT defenders have nothing *but* indiscriminability to try to establish identity between contents; and that is not good enough for our usual standards of what makes for identity. The consequence is that we cannot speak of identity (or difference) *proper* if the only criteria we can use is discriminability. In other words, if we want to talk about sameness of content proper, we need an ulterior criterion beyond discriminability. PIT defenders cannot provide this ulterior criterion if they prioritise introspection, as they are bound to do.

To clarify this further, let's return to **Continuum**. The subject will report, based on introspection and the phenomenal character of their perceptual experience, that the content of their experience of s_1 and s_2 is identical; they cannot even individuate that they are two different shades. They will also report that the content of their experience of s_2 and s_3 is identical, too, and so on for any two adjacent shades with the exception s_1 and s_8. Since identity is transitive, however, if the subject's reports are true, we would need to accept the consequence that the content of the subject's experience of s_1 and s_8 is identical. However, we can safely assume that the subject would

report that the content of their experience of s_1 and s_8 is not identical. Since PIT defenders need to take introspective reports related to the content of one's experience as true, we have run into a contradiction.

1. The content of the visual experience of s_1 is identical to the content of the visual experience of s_2. (subject's report)
2. The content of the visual experience of s_2 is identical to the content of the visual experience of s_3. (subject's report)
3. …
4. The content of the visual experience of s_7 is identical to the content of the visual experience of s_8. (subject's report)
5. The content of the visual experience of s_1 is identical to the content of the visual experience of s_8. (1-4, transitivity of identity)
6. The content of the visual experience of s_1 is not identical to the content of the visual experience of s_8 (subject's report)
7. $\perp$ (5, 6)

The choice is between two routes at this point. The first one is abandoning prioritising introspection, which would allow us to reject at least one of the subject's reports as false; but PIT defenders cannot do this because of how central introspection is to their approach and methodology. The second route is denying that there is such a thing as identity proper between contents, which would allow one to discard 5.

Do we want to be able to talk of sameness of content proper, that is, identity between contents? It seems to me as if the answer is yes. First, it is very intuitive to believe two distinct intentional states can have the same content; it seems worthwhile to honour this intuition, in particular within a theory of content. In addition to this, PIT is often minimally or partially defined in terms of the supervenience relation supposedly subsisting between phenomenal and representational properties. To do so, there needs to be an intra-theoretic way of spelling out what it means for two intentional states to have the same representational properties – that is, what it means for two intentional states' contents to be identical.

To summarise, the **Continuum** scenario is a case in which defenders of PIT run into a contradiction. To avoid the contradiction presented by the argument, one needs to either abandon prioritising introspection when it comes to intentional contents of perceptual experiences or give up on the notion of identity between contents altogether. It seems clear to me that the first of these options is preferable, especially if the theory we are trying to develop is a theory of intentional content; biting the bullet on identity between contents cannot plausibly be taken as collateral damage but is a deep and fundamental issue. Methodological reliance on introspection is an ill-advised move.

An immediate counterargument to the points just made is that PITs might not need to rely on introspection as fully as has just been laid out. However, I believe there is a relatively simple argument to be made that blocks this kind of response. The basic idea behind PITs is that phenomenal character determines the content of intentional states in at least some cases; and phenomenal character is exclusively introspectively accessible. This makes it very hard to imagine that anything other than introspection can be relied on when it comes to uncovering the contents of experience in these fundamental cases, such as perceptual experience.

Mendelovici likely predicted that excessive reliance on introspection could be an issue, particularly in interpersonal content individuation, as she includes *considerations of psychological role* as an additional way of pre-theoretically "figuring out" what the contents of our intentional states are (2018, pp. 27–28). Considerations of psychological role as Mendelovici understands them are meant to be something akin to the functional roles mental states play in the psychology and behaviour of conscious beings, such as their role in inferences and behavioural output. However, this fundamentally functional and dispositional reading of the contents of intentional states is clearly secondary in PIT because it does not support PIT over its adversaries.

Given a conflict between what introspection tells us about the content of our intentional states and what considerations of psychological role tell us about the content of a perceptual intentional state, the defender of PIT will need to take the introspective judgement to have primacy. To demonstrate this, we can construct an argument that shows that considerations based on psychological role do not reliably provide us with the contents of intentional states that match those revealed to us via introspection. This argument closely mirrors Mendelovici's mismatch argument for tracking theories, in which she concludes that tracking theories should be abandoned because the predictions they make regarding the contents of our experience do not match our pre-theoretical intuitions.

Mismatch argument from omission v.2.0 (general):

(P1) If predictions of content according to psychological role are true, then intentional state S does not have content *c*.

(P2) Intentional state S does have content *c*. (Justified only by introspection)

(C) Predictions of content based on psychological role are false.

The argument works as long as it is *possible* for predictions from psychological role to come apart from predictions from introspection; this seems to me to be obviously true, though I will provide an example shortly for the less-than-convinced reader.

A hidden assumption in this argument is that, given a mismatch between the content predicted by considerations based on psychological role and the content given to us via introspection, we should choose introspection.[76] This is something the defender of PIT is bound to do; otherwise, they'd have no way of dismissing alternative theories of content which already

[76] A similar (though not identical) "hidden assumption" is also needed for the original mismatch argument to go through, as can be seen in Mendelovici 2018, pp. 41-43. Mendelovici is aware of this, which is why she discloses her methodological assumptions right before presenting the argument.

provide a functional, dispositional, or causal explanation of content. In other words, an attempt to overcome this argument by denying P2 is akin to dismissing a crucial part of PIT: that introspection reliably provides us with the contents of our experience via the access it provides to the phenomenal character of that experience. Accepting that different kinds of factors (such as the psychological role of an intentional state) trump introspective access when it comes to determining content boils down to admitting that phenomenal character is not always crucial in determining content; and that is not consistent with strong versions of PIT.

For a concrete example of a possible mismatch between the content predicted by considerations based on psychological role and the content given through introspection, we can use a simple inverted spectrum case. Imagine a subject who can perceive colour with the same sensitivity, reliability, and in a way that is behaviourally indistinguishable from a normal subject while having "inverted" colour qualia. In this case, there should be no difference in the psychological role of the inverted subject's perception of a blue sky and anyone else's perception of a blue sky; the intentional state of perceiving the colour of the sky would play the same inferential role, lead to the same behavioural outputs, and so on. On the other hand, based on the definition of inverted spectrum cases, the phenomenal experience of the two subjects differs. For example, the inverted subject's phenomenal experience when gazing at the sky is the same as the normal subject's phenomenal experience when looking at a well-lit orange and vice versa.

Mismatch argument from omission v.2.0 (specific):

(P1) If predictions of content according to psychological role are true, then S's intentional state of perceiving sky-blue does not have the content "edenic orange."[77]

(P2) S's intentional state of perceiving sky-blue does have the content "edenic orange." (Justified by the inverted subject's introspection)

(C) Considerations based on psychological role are false.

The only way this alternative version of the mismatch argument could fail is if it were impossible for predictions based on psychological role, which include predictions made from the third person, to mismatch what is apparent to subjects introspectively. We have no reason to believe this to be the case; if someone disagreed with this being a possibility, the burden of proof – specifically of proving that introspection and considerations based on psychological role necessarily yield the same predictions regarding the contents of intentional states – would fall on them. Having to choose between introspection and considerations based on psychological role, proponents of PIT are forced to choose introspection; otherwise, the theory would stray from its fundamental assumptions. PIT defenders must prioritise introspection over other ways of determining the contents of intentional states, and that leads the contradiction presented by the ***Continuum*** case.

I argued that PITs lead to a contradiction, even assuming a fairly weak version of the theory. To demonstrate this, I've argued that in cases of perceptual colour continua, PITs must choose between either discarding introspective judgements about perceptual experiences or forgoing identity between contents altogether. Since the second option is unacceptable for any theory of intentional content, we are left with the first one. I've then argued that PIT methodologically relies on introspection in a way that does not allow defenders of the theory to

[77] I use the notion of "edenic colours," which are supposed to stand for primitive, simple colours, because it is the same one used by Mendelovici in her mismatch argument. Mendelovici suggests that our pre-theoretical considerations about colour perception tell us that we represent edenic colours when having visual experiences, and that this is an intuition that should be honoured by any adequate theory of content.

discard subjective reports about perceptual experiences. This leads me to conclude that PITs are inconsistent and should be abandoned.

CONCLUSION

This chapter's aim was to determine whether PITs can answer the CDQ while avoiding Kripkenstein's paradox. Even after considering two different ways in which one could understand phenomenal properties, I concluded that they cannot in any interpretation form the basis for representational content. There is nothing about the phenomenal properties of an experience that can produce correctness conditions, as Kripke already suggested in *WRPL*. The alternative, interpreting phenomenal properties as immediately presenting subjects with accuracy conditions, is another case of "sneaking in semantic elements" to our foundational explanation of content. In other words, an answer to the CDQ based on phenomenal properties violates either the normativity requirement or the non-circularity requirement.

In addition to this, even relatively weak versions of PITs suffer from internal issues. In section 5.3, I demonstrated that a contradiction can be derived from PITs as long as they allow that there are cases in which phenomenal properties determine the intentional properties of a mental state. The only additional assumption that must be made to derive the contradiction is that there can be identity between contents. The root of the problem lies in PITs' inevitable methodological reliance on introspection. I conclude that PITs do not present a promising avenue for adequately answering the CDQ.

6. CONCLUSION

I began this work by wondering whether there have been any recent advances in explaining how things like beliefs, perceptions, and sentences have determinate contents. Kripkenstein's sceptical paradox proved useful in isolating both the main desiderata for theories of content and the obstacles that they might run into when attempting to answer the CDQ. The challenge turned out to be the following: constructing a theory that yields *the right contents*, and that explains how *correctness conditions* arise, without sneaking in any semantic elements into the explanans. My general conclusion, beyond the individual concluding remarks made at the end of each chapter contained in the thesis, is that none of the surveyed theories manage to meet this challenge.

Where does this leave us? I do not believe I have shown that answering the CDQ is impossible. However, the issues faced by recent attempts at explaining how mental states and linguistic expressions have the meanings that they do are discouraging. The results of this thesis add to the mounting evidence that there may be no way of accounting for determinate content, at least not in a reductive manner.

There are three possible ways to proceed when it comes to future theorising about content. The first is to continue the pursuit of a reductionist account of content; I believe that the most promising direction to take here is a version of teleosemantics that allows for some degree of content indeterminacy. The challenge then becomes adequately justifying and defending the idea that representational content is indeterminate. It is important to note that this solution amounts to (at least partially) giving up on the part of the challenge that requires theories to account for *the right* contents.

A second strategy consists in adopting a primitivist approach to content wherein semantic properties are not explained in virtue of more fundamental, non-semantic properties. This amounts to saying that there is no underlying aspect of reality that systematically gives rise to contents. While this solution might preserve content-determinacy, it gives up on the part of the challenge that requires theories to give an account of content that is non-circular, that is, that does not explain semantic properties in virtue of semantic properties. Contents are left, to some degree, unexplained.

The third and last strategy, and the one I am most hopeful about, consists in positing that the longevity of Kripkenstein's sceptical challenge points towards the possibility that the notion of content is inconsistent. As has become apparent throughout the research presented in this thesis, "content" is a theoretical term, but it is bound up in many contrasting intuitions about what its role is. Contents simply have too many jobs to do. To illustrate, here is a (non-comprehensive) list of the roles content is supposed to play: it explains how mental states and expressions are about other things; it's something we have privileged introspective access to; it allows for semantic evaluation and the possibility of misrepresentation; it's what figures in adequate explanations of behaviour and reasoning. Contents are also widely assumed to be determinate and causally relevant to behaviour. It might be the case that one thing just cannot play all these roles.

The reader might protest that this third strategy seems to embrace the sceptical conclusion a little too closely. However, I do not believe this to be the case. The idea is not to eliminate the notion of meaning or of representational content altogether, but to reduce its workload to some degree. While this reduction in workload does mean that we need to revise the notion of representational content to some degree, it does not lead to full-blown scepticism about attributions of meaning. In other words, we might still be able preserve the intuition that there is a fact of the matter about whether someone means something by a sign. The way in which this

supposed division of labour should be handled – which tasks are going to stay under content's purview, and which will be relegated to some other theoretical notion – provides an interesting avenue for further research.

BIBLIOGRAPHY

Adams, F., & Aizawa, K. (2021). Causal Theories of Mental Content. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). https://plato.stanford.edu/archives/fall2021/entries/content-causal/

Artiga, M. (2013). Reliable misrepresentation and teleosemantics. *Disputatio*, *5*(37), 265–281. https://doi.org/10.2478/disp-2013-0020

Artiga, M. (2021). Beyond black dots and nutritious things: A solution to the indeterminacy problem. *Mind & Language*, *36*(3), 471–490. https://doi.org/10.1111/mila.12284

Bergman, K. G. (2023). Should the teleosemanticist be afraid of semantic indeterminacy? *Mind & Language*, *38*(1), 296–314. https://doi.org/10.1111/mila.12395

Bergman, K. G. (2025). Living with semantic indeterminacy: The teleosemanticist's guide. *Mind & Language*, *40*(1), 53–73. https://doi.org/10.1111/mila.12514

Boghossian, P. A. (1989). The rule-following considerations. *Mind*, *98*(392), 507–549. https://doi.org/10.1093/mind/xcviii.392.507

Boghossian, P. A. (1994). The transparency of mental content. *Philosophical Perspectives*, *8*, 33–50. https://doi.org/10.2307/2214162

Boghossian, P. A. (1997). What the externalist can know a priori. *Proceedings of the Aristotelian Society*, *97*(1), 161–176. https://doi.org/10.1111/1467-9264.00011

Bourget, D., & Mendelovici, A. (2019). Phenomenal Intentionality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019). https://plato.stanford.edu/archives/fall2019/entries/phenomenal-intentionality/

Braun, D. (1995). Causally relevant properties. *Philosophical Perspectives*, *9*, 447. https://doi.org/10.2307/2214230

Brentano, F. (2009). *Psychology from an Empirical Standpoint* (O. Kraus & L. L. McAlister, Eds.; A. C. Rancurello, D. B. Terrell, & L. L. McAlister, Trans.). Routledge. (Original work published 1874)

Buleandra, A. (2008). Normativity and correctness: A reply to Hattiangadi. *Acta Analytica*, *23*(2), 177–186. https://doi.org/10.1007/s12136-008-0028-y

Burge, T. (1988). Individualism and self-knowledge. *Journal of Philosophy*, *85*(11), 649–663. https://doi.org/10.5840/jphil1988851112

Chalmers, D. (2002). *Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press.

Copp, D., & Morton, J. (2022). Normativity in Metaethics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022). https://plato.stanford.edu/archives/fall2022/entries/normativity-metaethics/

Cummins, R. (1997). The lot of the causal theory of mental content. *The Journal of Philosophy*, *94*(10), 535–542. https://doi.org/10.2307/2564550

Davidson, D. (1967). Causal relations. *The Journal of Philosophy*, *64*(21), 691–703. https://doi.org/10.2307/2023853

Davies, B., & Evans, G. R. (Eds.). (1998). *Anselm of Canterbury: The Major Works*. Oxford University Press.

Dennett, D. C. (1996). *Content and Consciousness* (2nd ed.). Routledge. (Original work published 1969)

Deutsch, M. (2023). Is there a "qua problem" for a purely causal account of reference grounding? *Erkenntnis*, *88*(5), 1807–1824. https://doi.org/10.1007/s10670-021-00428-3

Devitt, M. (1998). *Reference*. In The Routledge Encyclopedia of Philosophy. Taylor and Francis. Retrieved 21 Dec. 2025, from https://www.rep.routledge.com/articles/thematic/reference/v-1. doi:10.4324/9780415249126-U034-1

Devitt, M., & Sterelny, K. (1999). *Language and Reality: Introduction to the Philosophy of Language* (2nd ed.). Blackwell. (Original work published 1987)

Di Lucia, P. (2011). Anselmo d'Aosta/Anselm of Canterbury: La duplice duplicità del vero. In S. Colloca (Ed.), *The Value of Truth, The Truth of Value*.

Douglas, S. P. (2018). The qua problem and meaning scepticism. *Linguistic and Philosophical Investigations*, *17*, 71–78. https://doi.org/10.22381/LPI1720184

Dretske, F. (1977). Laws of Nature. *Philosophy of Science*, *44*(2), 248–268. https://doi.org/10.1086/288741

Dretske, F. (1982). *Knowledge and the Flow of Information*. MIT Press. (Original work published 1981)

Dretske, F. (1986). Misrepresentation. In R. J. Bogdan (Ed.), *Belief: Form, Content, and Function* (pp. 17–36). Oxford University Press.

Ebbs, G. (2025). Content Externalism and Skepticism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2025). https://plato.stanford.edu/archives/spr2025/entries/skepticism-content-externalism/

Enoch, D. (2006). Agency, shmagency: Why normativity won't come from what is constitutive of action. *The Philosophical Review*, *115*(2), 169–198. https://doi.org/10.1215/00318108-2005-014

Farkas, K. (2008). Phenomenal intentionality without compromise. *Monist*, *91*(2), 273–293. https://doi.org/10.5840/monist20089125

Fodor, J. A. (1984). Semantics, Wisconsin style. *Synthese*, *59*(3), 231–250. https://doi.org/10.1007/bf00869335

Fodor, J. A. (1990). *A Theory of Content and Other Essays*. MIT Press.

Fodor, J. A. (1993). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (3rd print). MIT Press. (Original work published 1987)

Frisch, M. (2023). Causation in Physics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2023). https://plato.stanford.edu/archives/win2023/entries/causation-physics/

Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition and Logical Form*. Academic Press.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2019). *Cognitive Neuroscience: The Biology of the Mind* (Fifth edition). W.W. Norton & Company.

Geach, P. T. (1954). *Mental Acts: Their Content and Their Objects*. Routledge.

Gibbard, A. (2012). *Meaning and Normativity*. Oxford University Press.

Ginsborg, H. (2022). Going on as one ought: Kripke and Wittgenstein on the normativity of meaning. *Mind & Language*, *37*(5), 876–892. https://doi.org/10.1111/mila.12342

Glüer, K., & Wikforss, Å. (2015). Meaning normativism: Against the simple argument. *Organon F*, *22*(1), 63–73.

Glüer, K., Wikforss, Å., & Ganapini, M. (2024). The Normativity of Meaning and Content. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024). https://plato.stanford.edu/archives/fall2024/entries/meaning-normativity/

Goodman, N. (1983). *Fact, Fiction, and Forecast* (4th ed.). Harvard University Press. (Original work published 1954)

Gow, L. (2024). Intentionality and representation: Two kinds of aboutness. *Australasian Philosophical Review*, *8*(1), 20–30. https://doi.org/10.1080/24740500.2024.2485179

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388. https://doi.org/10.2307/2182440

Guardo, A. (2009). The argument from normativity against dispositional analyses of meaning. In A. M. Volker, K. Puhl, & J. Wang (Eds.), *Language and World – Papers of the XXXII International Wittgenstein Symposium* (pp. 163–165). Austrian Ludwig Wittgenstein Society.

Gutzmann, D. (2015). *Use-Conditional Meaning: Studies in Multidimensional Semantics* (First edition). Oxford University Press.

Hattiangadi, A. (2007). *Oughts and Thoughts: Rule-following and the Normativity of Content*. Clarendon Press.

Hattiangadi, A. (2009). Some more thoughts on semantic oughts: A reply to Daniel Whiting. *Analysis*, *69*(1), 54–63. https://doi.org/10.1093/analys/ann009

Hattiangadi, A. (2024a). Physicalism, intentionality, and normativity: The essential explanatory gap. In G. N. Kemp, A. Hossein Khani, H. Sheykh Rezaee, & H. Amiriara (Eds.), *Naturalism and Its Challenges* (1st ed., pp. 69–88). Routledge. https://doi.org/10.4324/9781003430568

Hattiangadi, A. (2024b). Quadders and zombies: A Kripkean argument against physicalism. In C. Verheggen (Ed.), *Kripke's* Wittgenstein on Rules and Private Language *at 40* (1st ed., pp. 181–200). Cambridge University Press. https://doi.org/10.1017/9781009099103.011

Horgan, T., & Graham, G. (2012). Phenomenal intentionality and content determinacy. In R. Schantz (Ed.), *Prospects for Meaning* (pp. 321–344). De Gruyter. https://doi.org/10.1515/9783110216882.321

Horgan, T., & Graham, G. (2022). Content-determinacy skepticism and phenomenal intentionality. In S. C. Hetherington & D. Macarthur (Eds.), *Living Skepticism. Essays in Epistemology and Beyond* (pp. 139–160). Brill.

Horgan, T., & Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In D. Chalmers (Ed.), *Philosophy of Mind: Classical and Contemporary Readings* (pp. 520–533). Oxford University Press.

Kaplan, D. (1999). The meaning of ouch and oops. *[Lecture Transcript]*. Available at https://eecoppock.info/PragmaticsSoSe2012/kaplan.pdf

Kim, J. (1976). Events as Property Exemplifications. In M. Brand & D. Walton (Eds.), *Action Theory*. Springer Netherlands. https://doi.org/10.1007/978-94-010-9074-2

Korta, K., & Perry, J. (2024). Pragmatics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024). https://plato.stanford.edu/archives/fall2024/entries/pragmatics/

Kriegel, U. (2013). The phenomenal intentionality research program. In U. Kriegel (Ed.), *Phenomenal Intentionality* (3rd ed, pp. 1–26). Oxford University Press USA - OSO.

Kripke, S. (1977). Speaker's reference and semantic reference. *Midwest Studies in Philosophy*, *2*(1), 255–276. https://doi.org/10.1111/j.1475-4975.1977.tb00045.x

Kripke, S. (1995). *Wittgenstein on Rules and Private Language: An Elementary Exposition* (8th reprint). Harvard University Press. (Original work published 1982)

Kripke, S. (2001). *Naming and Necessity* (12th reprint). Harvard University Press. (Original work published 1980)

Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*(17), 556–567.

Loar, B. (2003). Phenomenal intentionality as the basis of mental content. In T. Burge, M. Hahn, & B. T. Ramberg (Eds.), *Reflections and replies: Essays on the philosophy of Tyler Burge* (pp. 229–257). MIT Press.

Lycan, W. (2023). Representational Theories of Consciousness. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2023). https://plato.stanford.edu/archives/win2023/entries/consciousness-representational/

Mackie, J. L. (1990). *Ethics: Inventing Right and Wrong* (Reprinted). Penguin Books. (Original work published 1977)

Maddy, P. (1984). How the causal theorist follows a rule. *Midwest Studies in Philosophy*, *9*(1), 457–477. https://doi.org/10.1111/j.1475-4975.1984.tb00072.x

Martínez, M. (2013). Teleosemantics and indeterminacy. *Dialectica*, *67*(4), 427–453. https://doi.org/10.1111/1746-8361.12039

Maund, B. (2024). Color. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024). https://plato.stanford.edu/archives/fall2024/entries/color/

Mendelovici, A. (2013). Reliable misrepresentation and tracking theories of mental representation. *Philosophical Studies*, *165*(2), 421–443. https://doi.org/10.1007/s11098-012-9966-8

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 University Press.

Millar, A. (2002). The normativity of meaning. *Royal Institute of Philosophy Supplement*, *51*, 57–73. https://doi.org/10.1017/S1358246100008080

Miller, A., & Wright, C. (Eds.). (2002). *Rule-following and Meaning*. Acumen.

Miller, R. B. (1992). A purely causal solution to one of the qua problems. *Australasian Journal of Philosophy*, *70*(4), 425–434. https://doi.org/10.1080/00048409212345301

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories*. MIT Press.

Millikan, R. G. (1995). *White Queen Psychology and Other Essays for Alice* (1st paperback ed., 2nd print). MIT Press. (Original work published 1993)

Moore, G. E. (1922). *Principia Ethica* (1st reprint). Cambridge University Press. (Original work published 1903)

Nanay, B. (2022). Entity realism about mental representations. *Erkenntnis*, *87*(1), 75–91. https://doi.org/10.1007/s10670-019-00185-4

Neander, K. (2017). *A Mark of the Mental*. MIT Press.

Papic, S. (2023). Can Linguistic Correctness Provide Us with Categorical Semantic Norms? *Phenomenology and Mind*, *24*, 182–191. https://doi.org/10.17454/pam-2413

Papic, S. (2024). Why the qua problem has not been dissolved: Reply to Deutsch. *Erkenntnis*. https://doi.org/10.1007/s10670-024-00802-x

Papineau, D. (1984). Representation and explanation. *Philosophy of Science*, *51*(4), 550–572. https://doi.org/10.1086/289205

Parent, T. (2024). Externalism and Self-Knowledge. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024). https://plato.stanford.edu/archives/fall2024/entries/self-knowledge-externalism/

Pautz, A. (2013). Does phenomenology ground mental content? In U. Kriegel (Ed.), *Phenomenal Intentionality* (3rd ed, pp. 194–234). Oxford University Press.

Pautz, A. (2021). Consciousness meets lewisian interpretation theory: A multistage account of intentionality. In U. Kriegel (Ed.), *Oxford Studies in Philosophy of Mind, Vol. 1*. (pp. 194–234). Oxford University Press.

Piccinini, G. (2004). Functionalism, computationalism, and mental contents. *Canadian Journal of Philosophy*, *34*(3), 375–410. https://doi.org/10.1016/j.shpsa.2004.02.003

Potts, C. (2004). *The Logic of Conventional Implicatures*. Oxford University Press.

Predelli, S. (2013). *Meaning Without Truth*. Oxford University Press.

Putnam, H. (2004). *Reason, Truth and History* (Digital reprint). Cambridge University Press. (Original work published 1981)

Quine, W. V. O. (1960). *Word and Object*. MIT Press.

Reiland, I. (2023). Linguistic mistakes. *Erkenntnis*, *88*(5), 2191–2206. https://doi.org/10.1007/s10670-021-00449-y

Rupert, R. D. (2001). Coining terms in the language of thought: Innateness, emergence, and the lot of Cummins's argument against the causal theory of mental content. *The Journal of Philosophy*, *98*(10), 499–530. https://doi.org/10.2307/3649467

Schaffer, J. (2005). Contrastive causation. *The Philosophical Review*, *114*(3), 327–358. https://doi.org/10.1215/00318108-114-3-327

Schulte, P. (2023). *Mental Content*. Cambridge University Press. https://doi.org/10.1017/9781009217286

Schulte, P., & Neander, K. (2022). Teleological Theories of Mental Content. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022). https://plato.stanford.edu/archives/sum2022/entries/content-teleological/

Searle, J. R. (2004). *Intentionality: An Essay in the Philosophy of Mind* (15th reprint). Cambridge University Press. (Original work published 1983)

Searle, J. R. (2011). *Speech Acts: An Essay in the Philosophy of Language* (34th reprint). Cambridge University Press. (Original work published 1969)

Siewert, C. (1998). *The Significance of Consciousness*. Princeton University Press.

Stampe, D. W. (1977). Toward a causal theory of linguistic representation. *Midwest Studies in Philosophy*, *2*, 42–63. https://doi.org/10.1111/j.1475-4975.1977.tb00027.x

Sterelny, K. (1983). Natural kind terms. *Pacific Philosophical Quarterly*, *64*(2), 110–125. https://doi.org/10.1111/j.1468-0114.1983.tb00188.x

Verheggen, C. (Ed.). (2024). *Kripke's Wittgenstein on Rules and Private Language at 40*. Cambridge University Press.

Warren, J. (2024). Reference magnetism does not exist. *Erkenntnis*, *89*(7), 2825–2833. https://doi.org/10.1007/s10670-022-00654-3

Whiting, D. (2016). What is the normativity of meaning? *Inquiry*, *59*(3), 219–238. https://doi.org/10.1080/0020174X.2013.852132

Williams, J. R. G. (2020). *The Metaphysics of Representation*. Oxford University Press. https://doi.org/10.1093/oso/9780198850205.001.0001

Williamson, T. (1994). *Vagueness*. Routledge.

Wittgenstein, L. (2009). *Philosophical Investigations* (P. M. S. Hacker, J. Schulte, & G. E. M. Anscombe, Trans.; 4th ed.). Wiley Blackwell. (Original work published 1953)

Wright, C. (1984). Kripke's account of the argument against private language. *The Journal of Philosophy*, *81*(12), 759–778. https://doi.org/10.2307/2026031

Zalabardo, J. L. (1997). Kripke's normativity argument. *Canadian Journal of Philosophy*, *27*(4), 467–488. https://doi.org/10.1080/00455091.1997.10717482