Basic-Acceptance Teleosemantics

Esteban Withrington

The Graduate Center, City University of New York

How does access to this work benefit you? Let us know!
Follow this and additional works at: https://academicworks.cuny.edu/gc_etds

Part of the Metaphysics Commons, Philosophy of Language Commons, and the Philosophy of Mind Commons

Recommended Citation
https://academicworks.cuny.edu/gc_etds/3080

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
Basic-Acceptance Teleosemantics

by

Esteban Withrington

This manuscript has been read and accepted for the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date

David Rosenthal

Chair of Examining Committee

Date

Nickolas Pappas

Executive Officer

Supervisory Committee:

Michael Devitt

Michael Levin

David Rosenthal

THE CITY UNIVERSITY OF NEW YORK
ABSTRACT

Basic-Acceptance Teleosemantics

by

Esteban Withrington

Advisor: Michael Devitt

I propose an approach to naturalize semantics that combines the use-theory of meaning with teleosemantics. More specifically, I combine Horwich’s claim that the meanings of words are engendered by the acceptance of basic sentences that govern their deployment with the teleosemantic model, developed by Millikan, Papineau and Neander, according to which the meanings of symbols are related to functions determined by the history of their use and of the underlying biological mechanisms responsible for it.

Horwich’s account is general enough to offer plausible explanations of the meanings of all kinds of words and provides a plausible explanation of how meanings govern the uses of words in inference. But, as Devitt shows, his claim that meanings are law-like regularities in the use of words does not make enough room for misuses due to ignorance or error, which may be regular. I argue that this problem can be overcome by adopting a teleonomic account of the functions of symbols, which allows for failures in performance. The teleonomic account characterizes functions as what items are supposed (but may fail) do to, based on their selective history. While this is a biological notion, Millikan and Papineau have proposed plausible ways to extend it to acquired representations. Available teleosemantic theories are truth-referential and are usually regarded as competing with use-theories that are motivated by deflationary views of
truth and reference. I argue that we need the basic-acceptance account independently of the fate of deflationism and that it can be articulated in truth-referentialist terms. Additionally, I argue that we need to combine it with teleosemantics. The resulting basic-acceptance teleosemantics claims that some basic sentences containing a word are supposed (but may fail) to govern its overall use. This account, unlike Horwich’s, makes plenty of room for words being misused due to ignorance or error.

Basic-acceptance semantics applies to symbols that play roles in inferential processes. For more basic animal representations that have direct perceptual causes and behavioral effects, I propose instead an account that combines Millikan’s effect-based teleosemantics with Neander’s cause-based teleosemantics. Millikan’s theory explains meanings in terms of the conditions in the world that representations are supposed to covary with in order to have the effects they have the function of producing. I argue that this theory has the advantage of making enough room for misrepresentation, due to its output-based character, but can ascribe meanings that are implausible because it ignores the causes of representations. Neander’s theory explains meanings in terms of the causes that the perceptual mechanisms are supposed to respond to. I argue that this theory has the advantage of ascribing meanings that are plausible given the perceptual capacities of organisms, but it does not make enough room for misrepresentation because it ignores the effects of representations. According to the hybrid account I propose, the meanings of basic representations are determined by what is supposed to cause their tokens in order for them to bring about the effects they have the function of producing. I argue that the hybrid account makes enough room for misrepresentation while ascribing meanings that are plausible given the capacities of the perceptual mechanisms that produce them.
For Cristina and Eduardo, who were always there for me, and for Gabriela, who inspired me to finish this work.
Acknowledgments

I want to express my deepest gratitude to my advisor, Michael Devitt, for his guidance, insightful comments, constructive criticism, patience, support and encouragement throughout the process of writing this dissertation. It was a real pleasure to have him as my advisor. I also want to thank the other members of my dissertation committee, David Rosenthal and Michael Levin, for their generosity with their time and their thoughtful remarks and questions.

I wrote this dissertation while teaching as an adjunct lecturer at The City College of New York. I want to thank everyone at the Department of Philosophy for their encouragement during this process, as well as their comradery and kindness, which made my time there one of the best experiences of my life.

Finally, I am very grateful to my family and friends for all their love and emotional support, without which I could not have finished this dissertation. Thank you for everything!
# Table of Contents

**Introduction** 1

1.1 Introduction 11

1.2 What are Propositions Alleged to Be? 13

1.3 Propositions and the “One over Many” Argument 23

1.4 Propositions and Attitude Ascriptions 32

1.4.1 Particular Content Ascriptions 34

1.4.2 General Content Attributions 40

1.4.3 Folkloric Opinion and Ontology 48

1.5 Against Meanings *qua* Abstract Objects 50

1.6 Naturalized Propositions? 57

1.7 Naturalized Meanings *plus* Abstract Propositions? 63

1.8 Conclusion 72

**Chapter 2: Truth, Reference and the Task of Semantics** 75

2.1 Introduction 76

2.2 The Nature of Truth and Reference 80

2.2.1 The Canonical Correspondence Theory 77

2.2.2 The Contemporary Correspondence Theory 80

2.2.3 Deflationism 84

2.3 Identifying and Explaining Meanings 90

2.3.1 How to Identify Meanings and Investigate their Nature? 96
<table>
<thead>
<tr>
<th>Chapter 3: Basic-Acceptance Semantics</th>
<th>109</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>109</td>
</tr>
<tr>
<td>3.2 The Place of Meaning in the World</td>
<td>110</td>
</tr>
<tr>
<td>3.3 The Range of Available Theories</td>
<td>113</td>
</tr>
<tr>
<td>3.4 A New Approach to Explain Meaning</td>
<td>117</td>
</tr>
<tr>
<td>3.5 Why Basic-Acceptance Semantics?</td>
<td>121</td>
</tr>
<tr>
<td>3.5.1 Horwich’s Use-Theory of Meaning</td>
<td>122</td>
</tr>
<tr>
<td>3.5.2 The Generality of a Basic-Acceptance Account</td>
<td>126</td>
</tr>
<tr>
<td>3.5.3 The Explanatory Power of a Basic-Acceptance Account</td>
<td>131</td>
</tr>
<tr>
<td>3.6 Problems with Horwich’s Use-Theory</td>
<td>138</td>
</tr>
<tr>
<td>3.6.1 Risk of Collapse</td>
<td>140</td>
</tr>
<tr>
<td>3.6.2 Problems of Ignorance and Error</td>
<td>143</td>
</tr>
<tr>
<td>3.7 Conclusion</td>
<td>152</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 4: Teleonomic Functions</th>
<th>153</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>153</td>
</tr>
<tr>
<td>4.2 What is Teleosemantics?</td>
<td>159</td>
</tr>
<tr>
<td>4.3 The Etiological Account of Biological Functions</td>
<td>165</td>
</tr>
<tr>
<td>4.4 Wright’s Defense of Teleology</td>
<td>168</td>
</tr>
<tr>
<td>4.5 Neander’s Proposal</td>
<td>171</td>
</tr>
<tr>
<td>4.6 Millikan’s Proposal</td>
<td>171</td>
</tr>
</tbody>
</table>
Introduction

In this dissertation, I propose a new approach to explain the nature of the meanings of human symbols. This approach is intended to cover the meanings of words and the meanings or contents of concepts—i.e., linguistic and mental symbols. The core idea is that a proper explanation of the nature of such meanings needs to combine insights from Paul Horwich’s “use” theory of meaning with insights from teleosemantic theories such as those proposed by Ruth Millikan, David Papineau and Karen Neander.

I borrow from Horwich’s theory the idea that the meanings of words are engendered by the acceptance of some basic sentences containing them which play the role of governing their use. Accepting a sentence involves “regarding it as true” and relying on it as a premise in practical and theoretical inferences. To put it picturesquely, to accept a sentence is to have it stored in the “belief-box”. I borrow from teleosemantic theories the idea that the meanings of symbols depend on etiological functions underlying their use or the mechanisms that govern it. Etiological functions are the functions that reproduced items have been naturally selected for performing. This may seem to constrain teleosemantics to innate representations. But Millikan and Papineau have proposed plausible accounts of how the approach can be extended to cover the meanings of acquired representations.

Horwich’s theory has two main advantages. The first advantage is that the theory is able to account for the meanings of different kinds of word. For example, Horwich proposes that the meaning of ‘bachelor’ is engendered by the acceptance of ‘The bachelors are the unmarried
that the meaning of ‘red’ is constituted by the acceptance of ‘This is red’ in response to the perception of red surfaces, and that the meaning of ‘and’ stems from the acceptance of the two-way argument schema, “p, q // p and q”. In these examples, we have one meaning constituted by links to other words, another meaning constituted by direct perceptual links to reality, and yet another meaning constituted by a rule of inference. It is worth pointing out that most theories of meaning do not have this generality. The classical “description theory”, for example, explains meanings in terms of the links of words to other words, which make it plausible only for terms like ‘bachelor’ and implausible for terms like ‘red’. The contemporary “indicator” theory, for example, explains meanings in terms of the (reliable) causal links of words to reality, which has some plausibility for terms like ‘red’, but none for terms like ‘bachelor’.

The second advantage of Horwich’s theory is that it is able to explain how the meanings of words govern their use, including their sophisticated roles in inference. The idea is that the use of a word in some basic sentences can account for its use in other sentences that result from inferential processes where the basic sentences work as premises. For instance, the sentence ‘The bachelors are the unmarried men’ can serve as an axiom in the inferential system, so that the use of ‘bachelor’ in other sentences can be derived from it. Similarly, stimulus-dependent tokens of ‘This is red’ can serve as premises for stimulus-independent uses of ‘red’. Many theories—like the “indicator” theory just mentioned—rely merely on links between isolated words and reality. A problem that often goes unnoticed is that such links cannot govern by themselves the use of words in inference or thinking processes. Only sentences can serve as premises in practical and theoretical inferences. So only properties related to the role of words in sentences can govern their overall deployment and, consequently, constitute their meanings.
The two advantages of Horwich’s theory just mentioned are lacking in available teleosemantic theories, which account for meanings based only on direct links between symbols and reality and are not articulated in terms of sentence acceptance. But Horwich’s theory has two problems pointed out by Michael Devitt. First, it is presented as a “use” theory that is compatible with deflationism: the view that reference and truth cannot play any role in the explanation of meaning. (Deflationists argue that the predicates ‘is true’ and ‘refers’ do not stand for any substantial properties that could play explanatory roles in semantics, because their meanings are fully explained by their roles in trivial statements such as the statement that ‘Snow is white’ is true if and only if snow is white and the statement that ‘Socrates’ refers to Socrates.) The problem with Horwich’s theory is that it risks collapsing into a form of truth-referentialism that is incompatible with deflationism, since it links words to reality in ways that non-deflationary theories utilize to explain truth and reference: these are precisely the sorts of links that we would expect if truth and reference did play an explanatory role in semantics. I think that a collapse of the “use” theory into a form of truth-referentialism is likely, but I am glad to embrace it, since I am not committed to deflationism.

The second and more serious problem is that Horwich’s theory, as it stands, is unable to regard misuses of words as misuses, because it makes meaning depend on whatever dispositions people have to deploy words, but people have dispositions to make mistakes—which the theory regrettably is forced to treat as correct uses. For example, if a person has a disposition to accept ‘This is a dog’ in response to the perception of dogs, coyotes and wolves, Horwich’s theory has to regard all such uses as correct, since they are among the dispositions that fix the meaning of the word. This problem motivates my proposal to combine Horwich’s basic-sentence acceptance
theory with teleosemantics.

Teleosemantic theories have the advantage of overcoming precisely the problem faced by Horwich’s theory and many other theories that similarly rely on dispositions. This is because teleosemantic theories deploy a concept of function that allows for characterizations of malfunctioning or failures in the performance of a function—and with the added benefit of allowing for such “normative” characterizations within the confines of naturalism. This is a notion of function that has its home in biology and is anchored on the selective history of items. The heart, for example, has the function of pumping blood (roughly) because hearts have been “selected” by Darwinian natural selection to pump blood. A heart that is not pumping blood can consequently be characterized as malfunctioning: as not doing what it is “supposed” to do. Teleosemantic theories deploy this notion of function (very roughly) to explain the misuses of symbols as situations where they are not deployed the way they are “supposed” to in order to perform their functions.

Teleosemantic theories apply straightforwardly to innately-determined representations. But Millikan and Papineau have offered reasonable accounts of how the approach can be extended to acquired symbols. Papineau suggests that the learning processes involved in the acquisition of novel concepts are selective processes. Millikan suggests that novel concepts have functions derived from the functions of the innate mechanisms responsible for their acquisition, which have been naturally selected for coining new concepts. Regarding linguistic words, Millikan suggests that they have direct functions because they are reproduced items subject to social selection. While these are very bold hypotheses, it is reasonable to expect acquired mental and linguistic representations to have etiological functions, since our linguistic and conceptual
capacities exhibit a complex “design” that is best explained as the result of natural selection and/or analogous selective processes.

The main drawback of available teleosemantic theories is that they do not account for the diversity of symbols that Horwich’s theory does, and they do not explain how the meaning of a symbol is responsible for governing its overall use in inferential processes. Indeed, Millikan and Papineau offer accounts according to which the meanings of concepts and words depend on their direct links to reality: they do not provide plausible accounts for the meanings of words like ‘bachelor’ or ‘and’. Also, their accounts do not explain meanings in terms of sentence acceptance properties, so the meanings they ascribe are not able to govern the uses of words in inference. Even worse, the links postulated by these theories are not even causal in nature: they merely involve correlations or co-variations. So they are detached from the causal roles of representations in cognition.

The situation, then, is the following: basic-acceptance semantics accounts for the meanings of different kinds of words and for how meanings govern use, but Horwich’s version is unable to characterize misuses as misuses, while teleosemantics can characterize misuses as misuses, but Millikan’s and Papineau’s versions are able to account neither for the meanings of all kinds of words nor for how meanings govern use. To overcome this situation, I propose a combination of basic-acceptance semantics with teleosemantics. The resulting “basic-acceptance teleosemantics” promises to combine the main virtues of each of the approaches while overcoming their main deficiencies.

In a nutshell, my proposal is that the meanings of words are constituted by the basic
sentences that are “supposed” to govern their use. The meaning of ‘bachelor’, for example, is likely constituted by the fact that ‘The bachelors are the unmarried men’ is “supposed” to govern its use. And the meaning of ‘dog’ likely stems from the fact that accepting ‘This is a dog’ in response to the perception of dogs is “supposed” to govern its use. Like Horwich’s theory and unlike standard teleosemantics, basic-acceptance teleosemantics can account for words with different kinds of meaning and for how meanings are causally responsible for word’s uses. But basic-acceptance teleosemantics, unlike Horwich’s theory and like standard teleosemantics, can make enough room for misuses. Suppose that a person has a disposition to accept ‘This is a dog’ in response to the perception of dogs, coyotes and wolves. As long as the word has a function specifically related to its application only to dogs—e.g., the word proliferated because of uses governed by accepting ‘This is a dog’ in response to the perception of dogs—then its applications to wolves and coyotes can be properly characterized as misuses.

In the course of working on these ideas I encountered fundamental questions regarding the metaphysical status of meanings and the methodology of semantics. I present my resulting reflections in the first two chapters of this dissertation. I argue in Chapter 1 for a naturalized semantics, based on the fact that meanings play causal roles in the physical world, and object to the view that meanings are abstract entities. The idea that meanings are abstract entities is mainly motivated by certain view of attitude ascriptions. Consider ‘Mary believes that it will rain’ and ‘John said that it will rain’. According to a standard analysis, such ascriptions relate the person to an abstract object outside of space and time, which happens to be the same object in the case of these two examples: the “proposition” that it will rain. Despite being popular, this view is troubling because abstract objects cannot play causal roles in the physical world, yet the
meanings postulated in attitude ascriptions do play such roles. The content of the belief ascribed to Mary, for example, may explain why she picks up an umbrella. I argue in Chapter 1 that we need a naturalized semantics that explains the nature of meanings in terms of properties or phenomena that can play causal roles in the physical world. In the process, I defend a view of attitude ascriptions suggested by Devitt, which serves as an alternative to the “propositionalist” one: the suggestion is that ‘that’-clauses ascribe semantic properties to concrete representation tokens by means of synonymy with the sentences embedded in the clauses. For example, the ascription to Mary above may be paraphrased as ‘Some belief of Mary has the same meaning as ‘It will rain’.’ Under this analysis, attitude ascriptions quantify over concrete utterances and thoughts and ascribe worldly semantic properties to them. In Appendix 2, I offer another consideration in favor of this analysis: I argue that it avoids some serious problems faced by the alternative non-propositionalist analysis offered by Friederike Moltmann.

In Chapter 2, I discuss what is the task of semantics. While it is not unusual to assume that the task is to explain the truth-conditions of sentences and the referential properties of words, I argue, following Devitt, that this begs the question against deflationists, who argue that truth and reference cannot play any substantial role in semantics. I engage in this chapter on a brief discussion of the correspondence theory of truth and deflationism. I point out that the canonical view of the correspondence relation as one of structural resemblance is hopeless, but that a contemporary alternative based on the logico-syntactic structure of sentences and the referential properties of their words constitutes a promising project. I suggest that the debate between deflationism and the correspondence view is likely to be settled by whether we find out that we need or not truth and reference to play explanatory roles in semantics. This shows that
we cannot settle the debate before engaging in semantics, and that we should not identify meanings in truth-referential terms before attempting to explain their natures. Following Devitt, I suggest that we should identify meanings based on the causal roles of linguistic and mental representations. As “folk semanticists” we identify pre-theoretically meanings in attitude ascriptions in order to explain the behavior of others and use them as guides to reality. As noted above, we can explain Mary’s picking up an umbrella based on her belief that it is raining. Also, we can learn that it will rain by ascribing the right meaning to John’s utterance, assuming he is reliable. In general, we identify such meanings by the causes and effects of representations. The same considerations apply to words and concepts, which also have meanings that we identify by their causal roles, by the way they are used. Pre-theoretically we regard meanings as the properties that explain the uses of words and concepts. Identifying meanings by the causal roles of representations has the advantage of being “neutral” between various semantic theories and not begging the question against deflationists. It sets a common explanatory target for the various theories to compete against each other. But it also establishes a clear requirement for a proper semantic theory: it must ascribe meanings that explain the causal roles of representations.

The second part of this dissertation begins in Chapter 3, where I briefly survey available semantic theories, offer a sketch of the combination of basic-acceptance semantics and teleosemantics, and elaborate on the virtues and problems faced by Horwich’s version of basic-acceptance semantics—along the lines sketched above. In Chapter 4, I briefly present the basic ideas behind teleosemantics and discuss in some detail the teleological notion of function, arguing that it has crucial advantages over dispositionalist alternatives. In Chapter 5, I discuss standard teleosemantic theories in more detail, focusing on the case of simple animal
representations that have direct perceptual causes and behavioral effects. My original idea was to briefly survey these theories before moving on to discuss human representations. But while assessing the views, I came to the conclusion that the main teleosemantic proposals for basic representations all had serious drawbacks. I ended up proposing a hybrid account that combines Millikan’s effect-based teleosemantics with Karen Neander’s cause-based teleosemantics. Millikan’s theory explains meanings in terms of the conditions in the world that representations are supposed to covary with in order to bring about the effects they have the function of producing. I argue that this theory has the advantage of making enough room for misrepresentation, due to its output-based character, but can ascribe meanings that are implausible because it ignores the causes of representations. Neander’s theory explains meanings in terms of the causes that the perceptual mechanisms are supposed to respond to. I argue that this theory has the advantage of ascribing meanings that are plausible given the perceptual capacities of organisms, but it does not make enough room for misrepresentation because it ignores the effects of representations. According to the hybrid account I propose, the meanings of basic representations are determined by what is supposed to cause their tokens in order for them to bring about the effects they have the function of producing. I argue that the hybrid account makes enough room for misrepresentation while ascribing meanings that are plausible given the capacities of the perceptual mechanisms that produce them. It is worth pointing out that my argument for this hybrid account relies on the methodological considerations discussed in Chapter 2. Finally, in Chapter 6, I discuss how available teleosemantic theories extend the approach to human representations, but object that Papineau’s and Millikan’s accounts of the meanings of human concepts and words have the problems pointed out above. Again, my
objections rely in part on the methodological considerations from Chapter 2. Finally, I argue that a combination of basic-acceptance semantics with teleosemantics—along the lines briefly sketched above—is able to overcome the problems faced by Horwich’s theory, on one hand, and Papineau’s and Millikan’s theories, on the other.

Millikan has resurrected the view of correspondence as structural resemblance I reject in Chapter 2. I argue in Appendix 1 that, despite her talk of “picturing” relations and “isomorphisms”, structural resemblances play no role in her account. This, of course, does not protect the account from the serious objections raised above and discussed in more detail in Chapters 5 and 6.

I conclude this introduction with an important terminological warning. I use profusely the word ‘property’ throughout this work. However, it must be noted that in Chapter 1, while arguing against Platonism, I express sympathy for a form of nominalism. According to nominalism, there are no such entities as universals. And talk of “properties” is usually and justifiably understood as talk of universals. Some people, however, do not want to “give away” the word ‘property’ to realists about universals. Talk of properties may, for instance, be consistent with particularism about properties: according to which properties are real but particular, not universal. On the other hand, ‘property’ can be taken as a convenient manner of speaking that should be paraphrased away in an ultimate metaphysical analysis. I use profusely the word because I do find it extremely useful. But I want readers to understand that I expect my use not to commit me to realism about universals. If pressed, I would say that I am using the word merely as a useful but figurative manner of speaking.
Chapter 1:  

The Metaphysics of Meaning

1.1 Introduction

There is a commonly held view that threatens any form of a naturalized semantics. Many in the field of philosophical semantics regard the contents of utterances and thoughts as *abstract entities*, rather than natural worldly phenomena. The main motivation for this view comes from an analysis of attitude ascriptions. Instead of analyzing attitude ascriptions as identifying phenomena that play causal roles in the physical world, the investigation of attitude ascriptions—as Michael Devitt points out—is “dominated by philosophers who talk of ‘propositions,’ Platonic objects that are separate from the concrete spatio-temporal world of meaningful tokens” (Devitt 1996: 83). These are the entities that W.V.O. Quine dismisses as “creatures of darkness” (Quine 1956: 180). The aim of this chapter is to argue *against* the view that the meanings or contents of concrete utterances and thoughts involve any relations to such entities. I will argue that there are compelling metaphysical reasons for regarding contents as natural worldly phenomena—to be explained by a naturalized semantics—and rejecting any appeal to propositions *qua* abstract objects. To defend naturalized semantics from the threat of Platonism, we have to delve into the metaphysics of meaning.

One of the main motivations for a naturalized semantics comes from *Metaphysical Naturalism*, which in its strongest form is the same doctrine as *physicalism*: the view that “all entities are physical entities and that the laws they obey are in some way dependent on physical laws” (Devitt 2010: 254). *Metaphysical Naturalism* should be distinguished from the doctrine of
*Epistemological Naturalism*, according to which philosophical knowledge is continuous with scientific knowledge and does not have the privileged status it was traditionally thought to have. Quine defends this doctrine with his rejection of *a priori* knowledge. Since all knowledge is empirical, Quine argues, there is no place for a “first philosophy” that investigates reality independently of science: “it is within science itself, and not in some prior philosophy, that reality is to be identified and described” (Quine 1981a: 21). (See Devitt 2010: 254-255.)

But there is a weaker *metaphysical* doctrine that is also often regarded as a form of “Naturalism”: the view that only physical forces can act upon the natural world. Most contemporary philosophers accept that physical events can only have physical causes—since this follows from well-established and fundamental scientific principles like the “conservation of energy”. Accepting the causal closure of the physical world, however, leaves room for admitting the existence of non-physical entities that do not have any physical effects. (Many contemporary dualists about the mental, for example, have retreated to an epiphenomenalist version of dualism according to which mental states do not cause intentional behavior.) Some philosophers seek to naturalize semantics because they accept this weaker form of naturalism and think that meanings do have effects in the physical world, but they are not physicalists. This is the case of Paul Horwich (1998a & 1998b), who offers a naturalistic theory of *meaning*, but also supports the view that the *propositions* expressed by linguistic and mental sentences are abstract objects. I will address Horwich’s way of combining propositionalism with naturalized semantics and argue that, while there is one interpretation under which the abstract entities he posits are innocuous, there is another interpretation under which his view does conflict with the causal closure of the physical world. While I sympathize with the stronger naturalist doctrine—physicalism—I will
argue that the weaker and widely accepted doctrine—the causal closure of the physical world—is enough to rule out abstract objects playing the role of meanings or contents.

I will proceed as follows: I will clarify what propositions are alleged to be (Section 1.2), I will criticize the main arguments offered in support of the existence of such entities (Sections 1.3 & 1.4), and I will argue that meanings must be worldly phenomena, rather than abstract objects, in order to play the explanatory roles they ought to play (Section 1.5). Afterwards, I will discuss whether there are grounds for naturalizing propositions, rather than rejecting them altogether (Section 1.6). And, finally, I will argue against Horwich’s account of propositions \textit{qua} abstract objects—or, more specifically, one interpretation of it—which he combines with a naturalistic account of meanings (Section 1.7).

1.2 What are Propositions Alleged to Be?

The way talk of propositions is usually introduced does not seem \textit{ostensibly} controversial or obscure, but serious complications arise when we consider the ontological commitments such talk is often interpreted as having and whether the postulated entities can play the explanatory roles they are supposed to play. In this section, I shall discuss what propositions are often alleged to be: abstract objects that constitute the contents of utterances and thoughts and that are the primary bearers of truth. There are alternative views of the nature of propositions. For example, there is the view that propositions are identical to the \textit{facts} that make sentences and beliefs true, the view that they are \textit{worldly properties} of utterances and thoughts, and the view that they are merely convenient \textit{fictions} we talk about figuratively to formulate generalizations about the meaningful utterances and thoughts. My concern here is with the view that propositions are
abstract objects. In the following sections, I will critically assess the main arguments for this view and argue that the meanings or contents of utterances and thoughts should not be regarded as propositions *qua* abstract objects, since meanings play causal-explanatory roles that abstract objects cannot play. In this section I will discuss the main features of the view I am opposing.

Propositions are often said to be “what is common to a set of synonymous declarative sentences”, so that two sentences “express the same proposition if they have the same meaning” (Haack 1978: 76-77). For example, ‘Snow is white’ in English, ‘La nieve es blanca’ in Spanish and ‘Schnee ist weiß’ in German have the same meaning, so they are all alleged to express the same proposition—namely, *that snow is white*. A proposition is supposed to be something different from each of the declarative sentences that expresses it. It is what Alonzo Church calls a “proposition in the abstract sense”, which he distinguishes from a “proposition in the traditional sense” (Church 1956a: 3). The word ‘proposition’ was traditionally used to refer to a “declarative sentence taken together with its meaning” (Church 1956a: 3); but in its contemporary use the word refers *only* to the “content of meaning” of a declarative sentence (Church 1956a: 5) which, according to Church, is an “abstract object” (Church 1956b: 26). According to this view, the meaning of a declarative sentence *is* a proposition *qua* abstract object and what is common to various synonymous sentences is that they are all related to a single abstract object. A variant of the view is that the meaning of synonymous declarative sentences is a property they have in virtue of a relation to a proposition *qua* abstract object.

The claim that propositions are what synonymous declarative sentences have in common faces complications in the case of context-sensitive sentences. Presumably when Mary and John each say ‘I am hungry’, they are expressing different propositions, even though their utterances
have, in a sense, the same meaning.¹ So, as David Kaplan argues, a distinction needs to be drawn between two kinds of meaning. One kind is the unvarying meaning of an expression—e.g., utterances of ‘I am hungry’ always mean that the speaker is hungry—which Kaplan calls its “character” (Kaplan 1989: 505-507). The other kind is the varying meaning of an expression that changes from context to context—e.g., ‘I am hungry’ means that Mary is hungry in one context and that John is hungry in another—which Kaplan calls its “content” (Kaplan 1989: 500-505). Kaplan convincingly argues that, while the character of an expression depends on linguistic conventions, the content of a particular utterance of the expression depends on its character together with the context of the utterance: “The character of an expression is set by linguistic conventions and, in turn, determines the content of the expression in every context” (Kaplan 1989: 505). For example, the content of Mary’s utterance of ‘I am hungry’ is that Mary is hungry, because the character or linguistic convention for ‘I’ is that it refers to the speaker and the speaker in this context of utterance is Mary. Additionally, Kaplan argues that, when uttering a sentence, the content of the utterance—rather than its character—is what is said and what has a truth-value (Kaplan 1989: 500). So, what is said and what is true or false when Mary utters ‘I am hungry’ is that Mary is hungry. Armed with Kaplan’s distinction, proponents of propositions can handle context-sensitivity by specifying that two utterances express the same proposition when they have the same content, rather than the same character. Kaplan himself identifies propositions with the contents of sentence utterances: “The content of a sentence in a given context is... a proposition.” (Kaplan 1989: 500). To simplify exposition, I will mostly ignore in what follows the complexities introduced by context-sensitivity. But Kaplan’s distinction is

¹ In this case, the utterances are tokens of the same English sentence type. But the same considerations apply to tokens of synonymous context-sensitive sentences from different languages. When Carlos says ‘Tengo hambre’ in Spanish, he is presumably expressing a different proposition than Mary and John, even though ‘Tengo hambre’ is synonymous with ‘I am hungry’.
important regardless of whether the contents of declarative sentences are regarded as abstract objects or not. I will return to this issue after arguing against the account of propositions qua abstract objects.

Propositions are said to also be the contents of beliefs, identified in reports of the form ‘S believes that $p$’, where the ‘that’-clause contains a sentence expressing the content of the belief. Bertrand Russell, for example, said: “The content of a belief, when expressed in words, is the same thing... as what in logic is called a «proposition»... «That all men are mortal», «that Columbus discovered America»... are propositions.” (Russell 1921: 240-241). Russell characterized believing as a particular kind of psychological attitude that may be taken towards different propositions. Someone may believe that $p$, expect that $p$, remember that $p$, hope that $p$, fear that $p$, etc. Conversely, he pointed out that “there are various different attitudes that may be taken towards the same content.” (Russell 1921: 243). Someone may believe that $p$, expect that $p$, remember that $p$, hope that $p$, fear that $p$, etc. However, he warned: “These attitudes are not equally ultimate.” (Russell 1921: 244). Since Russell’s work, beliefs, desires and intentions have come to be regarded as the most fundamental “propositional attitudes”, because intentional behaviors are presumably caused by intentions formed as a result of practical inferences from beliefs and desires. While desires and intentions have satisfaction-conditions, only beliefs have truth-conditions. I will focus on beliefs and discuss other psychological attitudes only when relevant.

It is worth pointing out that Russell rejected the view that propositions are entities that exist separately from concrete representations when he characterized beliefs as propositional attitudes: he held instead that a meaningful declarative sentence is a “word-proposition”, while a contentful mental image is an “image-proposition” (Russell 1921: 241). Notice that what today is
regarded as the “Russellian” conception of propositions is the view held by the early Russell (1903)—I briefly discuss it below—which he later rejected. Russell’s view during this later period was closely related to Ludwig Wittgenstein’s: “the proposition is the propositional sign in its projective relation to the world” (Wittgenstein 1922: 3.12). Russell and Wittgenstein reverted to the traditional use of the word ‘proposition’.

Contemporary accounts of propositional attitudes, however, often adopt the view that they are relations between agents and abstract objects. According to this view, what the belief that snow is white, the sentence ‘Snow is white’ and the sentence ‘Schnee ist weiß’ have in common is that they are all related to a single abstract object: the proposition that snow is white. Regarding a proposition as an abstract object raises difficult metaphysical questions. Stephen Schiffer, for example, characterizes the nature of a proposition as follows:

Well: (i) it is abstract, in that it has no spatial location... (ii) It is mind- and language-independent in that it exists in possible worlds in which there are neither thinkers nor speakers... [It] is also language-independent in that, while it may be the content of a sentence of any language, it itself belongs to no language; it is not a linguistic entity... [Propositions are] abstract, mind- and language-independent objects... (Schiffer 1992: 506)

Basically, the claim is that propositions are Platonic objects that exist outside of the physical world of space and time and independently of any concrete utterances and thoughts. The main source of this view is Gottlob Frege, who argued that the proposition expressed by a sentence has

---

2 Locke, for example, said: “The signs we chiefly use are either ideas or words; wherewith we make either mental or verbal propositions.” (Essay: Book II, Ch. XXXII, §19). Similarly, Ockham claimed that the parts of “mental propositions” are “called concepts” and that “a mental proposition is composed... in the same way as a spoken proposition is composed of spoken words” (Summa Logicae: Part I C12).
only the *senses* of the words as constituents (Frege 1956: 292), which are “neither things of the
outer world nor ideas”, but abstract entities that exist in a Platonic “third realm” (Frege 1956:
302). These Platonic objects are extremely mysterious. Where do they exist, if not in space and
time? What are they made of? How can we know they exist if we cannot access the other-
worldly realm they inhabit? And, more importantly, how can they be the content of concrete
utterances and thoughts if—given the causal closure of the physical world—they cannot interact
with them? Since there are no satisfactory answers to these crucial questions—I will focus on the
last one, which is particularly pressing—there is a very strong case for disregarding propositions
qua abstract objects as “creatures of darkness”.

Finally, propositions are claimed to be the *primary* bearers of truth. This specific claim
has to be distinguished from the claim that propositions are truth-bearers—a broader claim that is
compatible with a number of differing views about the nature of propositions and their relation to
declarative sentences and beliefs. The claim that “propositions” are truth-bearers, without further
qualification, applies equally to “propositions” regarded as *meaningful sentences or contentful
beliefs* and to propositions regarded exclusively as the *meanings or contents* that sentences or
beliefs may have. In contrast, the claim that propositions are the *primary* truth-bearers is clearly
committed to the latter of these views. The idea is that particular sentences and beliefs are only
*derivatively* true: a belief is true only insofar as it has a true proposition as its content and a
sentence is true only when it expresses a true proposition. Talk of sentences or beliefs being true
is considered at best as parasitic. Frege, for example, claimed: “when we call a sentence true we
really mean its sense is” (Frege 1956: 292). The view is popular among contemporary
philosophers. Some argue even that it is implicit in our ordinary way of talking. Paul Horwich,
for example, claims: “Ordinary language suggests that propositional truth is fundamental and that the notion of an utterance ‘expressing a true proposition’ and the notion of a belief ‘being directed at a true proposition’ are understood in terms of it.” (Horwich 1998b: 133). I will challenge this view and the underlying assumption that ordinary language considerations can settle metaphysical controversies.

The particular view on the nature of propositions I oppose—and will argue against in the remainder of this chapter—is the one that endorses all of the three following claims:

(1) Propositions are the common meanings of synonymous sentences and beliefs.

(2) Propositions are the primary truth-bearers.

(3) Propositions are mind- and language-independent abstract objects.

Whether propositions so conceived really exist depends on whether synonymous declarative sentences and the beliefs identified by ‘that’-clauses containing them indeed have something in common, whether what they have in common is a legitimate and primary truth-bearer and whether what they have in common is a relation to an abstract entity.

The view that combines claims (1), (2) and (3) should be clearly distinguished from alternative views. I will focus for a moment only on sentences to simplify exposition. Consider the view according to which propositions are identical to the concrete facts represented by true sentences (e.g., Russell 1904). This view rejects (3), should also reject (2)—since it cannot draw a distinction between truth-makers and truth-bearers—and endorses a qualified version of (1): synonymous true sentences express the same proposition. Another view regards propositions as worldly semantic properties shared by synonymous sentences and holds that these properties are
the primary truth-bearers (e.g., Armstrong 2004). This view endorses (1) and (2) but rejects (3). Finally, there is the view that the primary truth-bearers are meaningful sentence tokens that are true or false partly in virtue of their meanings (and partly in virtue of how the world is), where such meanings are worldly phenomena (e.g., Devitt 1996). This view rejects (3), endorses (2) only if ‘proposition’ is used in its traditional sense and endorses a qualified version of (1): synonymous sentences have common meanings, but these meanings should not be regarded by themselves as propositions, since what is true or false are meaningful tokens.

While it is clear that various sentences and beliefs do have common semantic features that are responsible for their truth or falsity, it is far from clear that these common features involve abstract entities.\(^3\) Consider the various utterances and written inscriptions of ‘Snow is white’ in English, of ‘La nieve es blanca’ in Spanish, of ‘Schnee ist weiß’ in German, and the various instances of the belief that snow is white in the minds of different people. Surely all these concrete sentence and belief tokens have “something” in common: they all mean or have the content that snow is white. Moreover, these sentence and belief tokens surely are true in virtue of their meanings or contents (together with snow being white). Unless we are skeptical about the very notions of meaning or sameness of meaning—i.e., synonymy—there is no reason to disregard them as “creatures of darkness”.\(^4\) But why should we regard meanings as abstract entities rather than worldly features of sentence and belief tokens?

---

3 I will focus mainly on sentence and belief tokens. Different but synonymous (non-context-sensitive) sentence types are said to express the same proposition. In turn, different concrete tokens of the same (non-context-sensitive) sentence type are said to express the same proposition. The major contrast is between concrete sentence and belief tokens, on one hand, and the alleged abstract objects called “propositions”, on the other.

4 Quine’s rejection of propositions stems, in part, from his skepticism about synonymy: “If there were propositions, they would induce a certain relation of synonymy or equivalence between sentences themselves: those sentences would be equivalent that expressed the same proposition. Now my objection is... that the appropriate equivalence relation makes no objective sense at the level of sentences.” (Quine 1970: 3) I will ignore this skepticism here and assume that a naturalized semantics can account for different sentence and belief tokens having the same meaning or content.
The view that meanings are abstract objects is usually motivated by the metaphysical conviction that, when various sentence and belief tokens have the same representational content, what is common to them is not just a property but, rather, a relation to a single object: the same *thing* that is said or believed on many occasions and which has the property of being true or false.\(^5\) This postulated object is assumed to be something different both from the concrete sentence and belief tokens related to it and from the reality that would make it true. A proposition, on this view, is an *intermediary* between linguistic and mental entities, on one side, and the represented world, on the other. For example, it is alleged that a token of ‘Snow is white’ is (derivatively) true because it expresses the proposition that snow is white, which in turn is (primarily) true because snow is white. This contrasts with the simpler view that a token of ‘Snow is white’ is true because snow is white, with no additional entity mediating between the sentence and the world. The simpler view has the advantages of being more economical—i.e., postulating fewer entities—and relying only on things whose reality is already uncontroversial (Devitt 1996: 212).

I end this section with a clarification of the notion that propositions are mind- and language-independent entities. While I am focusing on the version of this view that can be traced back to Frege, there is another equally popular version that is not (necessarily) committed to propositions *qua* abstract objects: the so called “Russellian” account of propositions. The early Russell held that propositions do not contain symbols: “*Words* all have meaning... they are symbols which stand for something other than themselves. But a proposition... does not itself

---

\(^{5}\) The meaning of a sentence or belief token, on this view, may be regarded as an object or as the property of being related to an object. However, regarding the meaning as a property is also compatible with it *not* being a relation to an object: “[Whereas] if the meaning of a token is an object then the token obviously has the property of being related to that object; it is not the case that if the meaning is a property then that property is obviously a relation to some object.” (Devitt 1996: 57).
contain words: it contains the entities indicated by words” (Russell 1903: 48). On this view, the sentence ‘Mary loves John’ has the words ‘Mary’, ‘loves’ and ‘John’ as constituents, while the proposition it expresses is a mind- and language-independent entity that only has the referents of these words in the world as constituents: Mary, love and John. The early Russell and Frege agree that only the meanings of words are constituents of the proposition expressed by a sentence. Their disagreement is about the nature of those meanings. Frege regards senses while the early Russell regards referents as the constituents of propositions.

The “Russellian” account of the parts of a proposition needs also an account of what binds them together into a single entity. Russell claimed that “every proposition has a unity which renders it distinct from the sum of its constituents” (Russell 1903: 53). In the case of the sentence ‘Mary loves John’ the unity is provided by the syntactic arrangement of its words. But the proposition that Mary loves John allegedly contains no words. Russell proposed that what provides the unity of such a proposition is that the loving relation “actually relates” Mary and John (Russell 1903: 50). As Jeffrey King notices, on this view there is “a collapse of true propositions into the facts that make them true” (King 2007: 23). Russell for a time embraced the collapse and suggested that a proposition is true if and only if it is identical to a fact (Russell: 1904). But the collapse creates a serious problem. It entails, as King points out, that “there are no false propositions” (King 2007: 23). If it is not a fact that Mary loves John, then there is no unity that we can call the “false proposition” that Mary loves John. This simple point constitutes a reductio ad absurdum of the doctrine.

The Russellian account of the parts of propositions may be combined with other accounts of the unity of the proposition. It may be claimed the unity is provided by an abstract
that is not identical to the structure of the fact it represents. This proposal avoids the collapse of propositions into facts, but propositions are regarded again as abstract structured entities—even if their parts are concrete. The case against propositions *qua* abstract entities I make below applies also to this kind of view. King (2007) proposes a different modification of the Russellian view, according to which the unity of a proposition is provided by the syntactic structure of the sentence expressing it. This proposal also avoids the collapse of propositions into facts. But it clearly departs from the Platonic view I am discussing: King’s “propositions” are not language-independent.

Many philosophers, however, still subscribe to propositions *qua* abstract, mind- and language-independent objects. In the next two sections I will discuss the main arguments offered in support of this view. I will argue that none of them succeeds and that the explanation of the common contents of various declarative sentences and beliefs—in virtue of which they are true or false—should rely on worldly phenomena rather than relations to alleged abstract objects.

### 1.3 Propositions and the “One over Many” Argument

One consideration that may be offered in support of regarding the common contents of various sentence and belief tokens as abstract objects relies on the classical “One over Many” or “Universals” problem. In this section, I shall present the problem, discuss how it may be applied

---

6 King’s ingenious proposal, however, adopts a “direct reference” view of meaning along the lines proposed by Nathan Salmon (1986) and Scott Soames (2002). This view faces well-known problems: it cannot account for differences in meaning between—or in the “propositions” expressed by—sentences like ‘Lois Lane loves Superman’ and ‘Lois Lane loves Clark Kent’. Frege’s solution to this problem was to regard “modes of presentation” or senses as part of the meanings of words (Frege 1948). Frege problematically regarded the senses of names as associated descriptions: a view that has been debunked by Saul Kripke (1982). He also regarded the senses of sentences as abstract entities: the target of this chapter. But a broadly Fregean solution can be articulated without any of these problematic commitments (*see* Devitt 2001b & 2012).
to meanings and argue that it fails to provide support for regarding them as abstract objects.

How can many things have one single feature in common? Multiple objects, for example, are red. This mundane situation may seem, as David Armstrong puts it, “profoundly puzzling”:

The same property can belong to different things... Apparently, there can be something identical in things which are not identical. Things are one at the same time as they are many. How is this possible? (Armstrong: 1978: 11)

Plato proposed an uncanny solution to this problem: when many concrete and particular things share a common feature, there must be one abstract and universal thing—a “form”—that exists independently of any of them but that they all are related to. In the Republic Plato says: “[Take] any set of many things you like. For example, there are... many couches and tables... But the forms connected to these manufactured items are surely just two, one of a couch and one of a table.” (Rep.: 596a-b) Plato says of “all the forms”: “each of them is itself one thing, but because they appear all over the place in partnership with actions and bodies, and with one another, each of them appears to be many things.” (Rep.: 476a) Needless to say, Plato’s solution seems even more puzzling than the problem. According to Plato, if many things are red, it must be because they are all related to a single abstract object: redness itself. Similarly, it may be argued, we need to explain how numerically different sentence and belief tokens can have a single meaning or content in common that is responsible for their truth or falsity, and the Platonic solution is to postulate an abstract object that all of these tokens are related to. It is worth addressing this instance of the “One over Many” argument, even if it is seldom explicitly made nowadays by those who regard propositions as abstract objects. “One over Many” thinking is common and
may well have some influence on the view that meanings or propositions are abstract entities. Consider, for example, the following passage by George Pitcher:

If one person says “It is raining,” another “Il pleut,” and a third “Es regnet,” a correct answer to the question “What did he say?” would in each case be “He said that it is raining”—for each would have said the same thing. And it is this element which all three utterances have in common—this same thing that is said in all three cases—that is the real bearer of truth, not the different sentences which the speakers happen to utter. (Pitcher 1964: 5)

Since different synonymous sentence tokens all say the same thing, the passage suggests, this common feature must itself be another thing that they all are related to and that is the “real” truth-bearer. But this argument, as Richard Kirkham points out, is a “non sequitur” (Kirkham 1992: 64). It simply does not follow from different utterances saying the same thing that each utterance does not say it by itself or that it is not a truth-bearer in its own right: “The [different] tokens say the same thing, but each one says it independently. Why could not each of the sentence tokens be true because of what it says?” (Kirkham 1992: 64). The metaphysical principle that the common features of worldly things must involve a common relation to an abstract thing that “really” has that feature is, as Kirkham points out, itself problematic:

Consider the property of being my sister: there are two women in the world who possess this property, and it is not something they have in common that “really” possesses the property, they each have it. There is not one thing that is my sister,

---

7 In fairness, Pitcher seems to be merely presenting here what he takes to be a commonly held consideration, since he eventually rejects the view that meanings are Platonic objects (Pitcher 1964: 9).
there are two. (Kirkham 1992: 65)

Similarly, when various utterances say the same thing, there is no reason to suppose that what is said is another “thing”, in addition to the utterances, that is the “real” truth-bearer. The Platonic solution faces serious difficulties (which I discuss later in this section) and is incompatible with a naturalized semantics that expects to explain the meanings of utterances in terms phenomena that play causal roles in the world.

If we accept that there is a genuine “One over Many” problem, there is an alternative and more reasonable solution. Instead of endorsing a Platonic transcendent realism about universals—the view that “universals exist separated from particulars” (Armstrong: 1978: 140)—we may adopt an Aristotelian immanent realism—which “admits universals but denies that they are transcendent” (Armstrong: 1978: 137)—. Aristotle rejected the view that the common features of particular objects are themselves objects with a separate existence and proposed that, while universal features—or “forms”—are real, they only exist in particular objects. In the *Metaphysics*, Aristotle says: “in general nothing that is common is substance” since “that which is one cannot be in many things at the same time, but that which is common is present in many things at the same time; so that clearly no universal exists apart from the individuals” (*Met.*: Z XVI 1040b). According to an immanent realism, redness is a universal but concrete property that exists in—rather than separately from—particular red things, while being a sister of Richard Kirkham is a worldly property possessed by two particular individuals, not a single abstract

---

8 Kirkham’s reply works well against the kind of Platonism implicit in Pitcher’s passage: one that holds that F-ness itself must be F. While Plato apparently held this view, a contemporary Platonist may be better off rejecting it to avoid the problems associated with it (Armstrong 1978: 71-72). The difficulties I discuss below, however, apply to any kind of Platonism.
entity that they both are related to. Similarly, it may be argued, the common semantic properties of various sentence and belief tokens are universal but worldly properties that exist in these tokens. Such immanent realism avoids abstract objects altogether and is compatible with a naturalized semantics, since worldly meanings can play causal roles in mind and language (Armstrong 2004: 12-14). However, one may reasonably be skeptical about the need to posit universals to begin with. Is there really any mystery regarding how different tokens have the same meaning or how different things can be red?

Those who think that the “One over Many” problem is genuine find mundane situations like various things being red deeply puzzling and in need of explanation. Yet, Quine reasonably argues that this is a pseudo-problem and no explanation is needed:

One may admit that there are red houses, roses and sunsets, but deny, except as a popular and misleading manner of speaking, that they have anything in common...

That the houses and roses and sunsets are all of them red may be taken as ultimate and irreducible, and it may be held that [the realist about universals] is no better off, in point of real explanatory power, for all the occult entities which he posits under such names as ‘redness’. (Quine 1961: 10)

Quine’s suggestion is that particular things being red can be taken as a basic fact that needs no explanation and, additionally, that positing redness as a universal does not explain anything not already accounted for by acknowledging the existence of red things (Quine 1961: 10). His point is that there is nothing puzzling to begin with and the alleged explanations are idle. Now, suppose that I say ‘My car and my bike have the same property: redness’. Devitt argues that a

9 Armstrong (1978) offers a contemporary defense of immanent realism which I will discuss later in this section.
claim like this can be paraphrased into a claim that serves the same purpose but is not committed to a universal: I can say instead ‘My car and my bike are both red’, which trivially follows from ‘My car is red’ and ‘My bike is red’ (Devitt 2010: 14-15). The realist about universals, however, is puzzled even about how a single thing can have a property: how can my car be red? But one may wonder whether this is really puzzling. The fact that my car is red can be said to merely involve the existence of an object, my car, that really is red (Devitt 2010: 16). A further explanation in terms of universal redness seems to be neither required nor informative.

Devitt argues, however, that there are legitimate explanations we may seek for something having a property, which are not related to the problem of universals: we may wonder what caused or what is the purpose of something having certain property, and, in the case of a non-fundamental property, we may wonder what constitutes something having that property (Devitt 2010: 16).\textsuperscript{10} Suppose that we have a satisfying reductive explanation of what constitutes my car and my bike being red in terms of the wavelengths of light they reflect and absorb, as well as a satisfying causal explanation of why they do reflect certain wavelengths of light and absorb others. The problem of universals concerns even how objects can have these fundamental physical properties. But, if it is a pseudo-problem, no further explanation is required: “The explanation must stop somewhere. What better place to stop than with a fundamental physical fact of our world?” (Devitt 2010: 17).

\textsuperscript{10} Devitt (2010: 13-30) does not use the word ‘property’ in making this point, since he equates—in this context—posing properties with positing universals. However, he regards property talk as an acceptable manner of speaking to the extent that we are prepared to paraphrase it in a way that makes no commitment to “properties” (i.e., universals). While the transcendent realist regards properties as universal and abstract, the immanent realist regards properties as universal but concrete. Alternatively, properties can be regarded as both concrete (not abstract) and particular (not universal). [A detailed discussion of particularism about properties can be found in Armstrong (1978: 77-87).] To reject universals is not to reject concrete particular properties. I use here and elsewhere the word ‘property’ in a metaphysically non-committal way that is compatible with properties being universal or particular and with property talk being just a manner of speaking to be paraphrased away when addressing ultimate metaphysical questions.
These considerations can be applied to the question of how different sentence and belief tokens can have the same meanings. A token of ‘Il pleut’ and a token of ‘Es regnet’ having the same semantic property, it may be argued, is just a matter of both of them meaning that it rains: neither transcendent nor immanent universals are required. A naturalized semantics, of course, does not regard a token meaning that it rains as a basic fact, but rather as a fact in need of reductive explanation. The legitimate search for such an explanation, however, is not related to the problem of universals. Explanations must stop somewhere and no further explanation is required once we have a satisfactory reductive explanation of what fundamental properties constitute meanings.

I have argued that there are alternatives to the Platonic solution to the “One over Many” problem for synonymous sentence and belief tokens. One option is an immanent realism according to which meanings are universal but worldly properties that exist in the tokens. Another option is to reject the “One over Many” as a pseudo-problem. Neither of these alternatives posits abstract objects to account for the common meanings of synonymous tokens and both are compatible with a naturalized semantics. But why should we prefer either of these alternatives over Platonism? A compelling reason is that abstract objects can play no causal roles in mind and language (Devitt 1996: 210; Armstrong 2004: 12). This consideration undermines the Platonic treatment of meanings as abstract objects not only to account for synonymy, but for any reason.\footnote{I will elaborate this point in more detail after addressing other arguments for propositions qua Platonic objects.} There is also a compelling reason to reject abstract objects specifically as what explains how numerically different things can be identical in nature. As Armstrong convincingly shows, the appeal to transcendent abstract objects fails because it generates a vicious relation regress: it attempts to explain how things have a property in terms of their relation to an abstract
object, but then these things must also have the property of being related to the abstract object, which either is different from the first relation—in which case it needs to be explained by yet another relation, leading to an infinite regress—or is the same as the first relation—in which case the analysis is circular (Armstrong: 1978: 70-71). It follows that positing relations to abstract meanings fails to explain how various sentence and belief tokens can be synonymous: it generates a vicious relation regress.

Any attempt to solve the problem of universals that appeals to a relational analysis is undermined by a relation regress, since “the relations appealed to by such analyses themselves require the same analysis” (Armstrong: 1978: 139). This applies also to relational versions of immanent realism: if an object having an immanent property is a relation it has to this property, then it must also have the immanent property of having this relation, generating a vicious regress (Armstrong: 1978: 106-107). So, a relational immanent realism about meaning properties fails to account for how various tokens can be synonymous: the required relation between the tokens and their immanent meaning properties also generates a vicious relation regress.

After convincingly arguing that every other attempt fails, Armstrong proposes a non-relational form of immanent realism, according to which objects and their properties have a “more intimate union” than a mere relation: the properties of a thing, he suggests, are not separate constituents of the thing” (Armstrong: 1978: 107-108). But how can there be something identical in things which are not identical? If having a property is not a relation and a property is not a separate constituent of a thing, how can numerically different things have this

12 Armstrong (1978) also shows how vicious relational regresses undermine various forms of nominalism that attempt to solve the “One over Many” problem avoiding any appeal to universals. The views of Quine (1961) and Devitt (2010) just discussed are also versions of nominalism: they reject realism about universals. But there is a crucial difference. Quine and Devitt do not attempt to solve the problem but to dissolve it, so their proposal is immune to any relational regress objections: they argue that no explanation, relational or not, is required.
Same property? Armstrong answer is, basically, that this is inexplicable:

Different particulars may be... identical in nature. Such identity in nature is literally inexplicable, in the sense that it cannot be further explained...

We simply have to accept that different particulars may have the same property or be related by the same relation. (Armstrong: 1978: 109)

A non-relational immanent realism about meaning properties, consequently, would claim that various synonymous tokens have a meaning which is universal but inseparable from each of them and that their synonymy is an inexplicable fact that we simply have to accept.

Armstrong’s analysis shows that every relational solution to the “One over Many” problem fails. It seems initially reasonable to expect, then, a non-relational solution—if we take the problem to be genuine. Perhaps a non-relational account of properties can be worked out. But Armstrong’s own appeal to properties that are inseparable from objects makes inexplicable how they can be shared. And this problem may be unavoidable for any non-relational account, just like vicious regresses are unavoidable for relational accounts. If the alleged best solution to the “profoundly puzzling” fact that different objects can have the same property is itself a mystery, it does not seem a genuine solution (Devitt 2010: 18). Rather, it seems that the “One over Many” is a pseudo-problem after all (Devitt 2010: 19) and there is no mystery to begin with.

My concern in this section has been whether the “One over Many” problem provides a reason for positing propositions qua abstract objects. Does the fact that different sentence and belief tokens have the same meaning or content require a Platonic account of the metaphysics of meaning? I have argued that the “One over Many” fails to provide support for the claim that...
meanings are abstract objects. Suppose that the problem is genuine. Then a Platonic solution fails because it generates a vicious regress. Perhaps some version of immanent realism can solve the problem. But relational versions of immanent realism are undermined by the same problem as transcendental realism and Armstrong’s non-relational version does not seem to be a genuine solution either, since it leaves unexplained how different things can have the same property. It is likely that the problem is not genuine, and we simply need no explanation of how different tokens can be synonymous. A token of ‘Snow is white’ and a token of ‘La nieve es blanca’ being synonymous may merely be a matter of both tokens meaning that snow is white: a fact that requires a reductive explanation in terms of more basic meaning constituting facts but no further explanation in terms of universals. Either way, the “One over Many” does not undermine a naturalized semantics according to which meanings are constituted by worldly properties that play causal roles in mind and language. In the next section, I discuss other arguments for propositions \textit{qua} abstract objects that rely on the alleged ontological commitments implicit in our ordinary attitude ascriptions.

1.4 Propositions and Attitude Ascriptions

The metaphysical conviction that contents are abstract objects is often defended nowadays on the basis of ordinary language considerations. People use expressions like ‘what he said is true’, ‘what she believes is true’, or even ‘some things he said are true’, which may be interpreted as referring to things other than concrete utterances and thoughts or the reality they represent and which are the primary bearers of truth. Even ascriptions like ‘Mary believes that the Earth is flat’ are said to refer to mind- and language-independent propositions. The way people talk, advocates
of propositions qua abstract entities claim, is ontologically committed to the reality of such things. Sure enough, people sometimes talk about the “things” that are said or believed. But should this way of talking be taken literally? Is such talk actually committed to the existence of abstract objects that play the role of intermediaries between utterances, thoughts and the reality they represent? And, if so, should we regard this folkloric opinion as correct? These are all controversial matters. Perhaps advocates of propositions are misinterpreting ordinary language.\footnote{Wittgenstein (1953) claims that “intermediaries” between sentences and the world are mere \textit{chimeras} which philosophers posit because they misunderstand the workings of ordinary language: “‘A proposition is a queer thing!’ Here we have in germ the subliming of our whole account of logic. The tendency to assume a pure intermediary between the propositional \textit{signs} and the facts. Or even to try to purify, to sublime, the signs themselves.—For our forms of expression prevent us in all sorts of ways from seeing that nothing out of the ordinary is involved, by sending us in pursuit of chimeras.” (Wittgenstein 1953: § 94).} Maybe talk about the “things” that are said or believed is just a figurative way of talking about the meanings of concrete utterances and thoughts. Notice that truth is predicated of concrete utterances in ordinary language. We may “genuinely say ‘His closing words were very true’ or ‘The third sentence on page 5 of his speech is quite false’” (Austin 1950: 113). Likewise, truth is predicated of concrete thoughts. We may say ‘Her belief is true’, which refers to a concrete psychological state. So why should we regard propositional talk as more fundamental? Maybe ordinary language is simply non-committal about the ontology of contents and truth-bearers. More importantly, even if the way people talk were clearly committed to propositions qua abstract objects, it would remain to be seen whether these objects really exist, since folkloric opinion is not the ultimate guide to reality.

In what follows, I will present two influential arguments for propositions specifically based on the alleged commitments of attitude ascriptions in ordinary language: the first one is based on particular content attributions, while the second is based on general content attributions
and their inferential relations to particular attributions. I will argue that there are alternative and more plausible accounts of such attributions and that, even if the propositionalist accounts were correct about the commitments of attitude ascriptions, it would not follow that propositions exist.

1.4.1 Particular Content Ascriptions

An influential consideration offered for regarding contents as abstract objects relies on the alleged logical structure of particular content attributions of the form ‘S believes that p’ or ‘S said that p’. It is claimed that these attributions assert a relation between S and the object that p, because the ‘that’-clause works as a name or a “referential singular term” (Schiffer 1992: 506). Basically, particular content attributions are taken to have the same form as ‘A loves B’, which asserts a relation between A and the object B. The logical form of ‘A loves B’ is ‘Lab’, where ‘a’ and ‘b’ are names and ‘L’ is a two-place predicate. Analogously, the forms of ‘S believes that p’ and ‘S said that p’ are alleged to be ‘Bs<p>’ and ‘Ss<p>’, where the two-place predicates ‘Bxy’ and ‘Sxy’ symbolize the believing and saying relations, ‘s’ is the name of the believing or saying subject and ‘<p>’ is the name of the proposition: the thing that is believed or said.

However, there are alternative accounts of the logical form of particular content attributions. A plausible alternative, suggested by Devitt, is that they have the same form as ‘A loves some B’, which asserts a relation between A and some object that has the property B (Devitt 1996: 56-57). If the ‘that’-clause functions as an “indefinite singular term” instead of a referential one, Devitt points out, then what the content attribution asserts is a relation between S and some object which has the semantic property specified by ‘that p’—where the object is “a token thought” in the case of ‘S believes that p’ and “a token utterance” in the case of ‘S said
that p’ (Devitt 1996: 56). The logical form of ‘A loves some B’ is the existentially quantified
‘∃x(Lax & Bx)—something that A loves has the property of being a B—, where ‘B’ functions as
a predicate. Similarly, the forms of ‘S believes that p’ and ‘S said that p’ can be said to be
‘∃x(Bxs & Px)—some belief/utterance of S has the property P—, where the
one-place predicate ‘Px’ symbolizes the semantic property specified by ‘that p’, the existential
quantifier ‘∃x’ ranges over sentence and belief tokens, and the two-place predicates ‘Bxy’ and
‘Uxy’ symbolize the relations “x is a token belief of y” and “x is a token utterance of y”. If we
additionally grant that the sentence embedded in the ‘that’-clause (the ‘p’ in ‘that p’) specifies a
semantic property by means of synonymy or sameness of content, then what the content
attribution asserts is, more precisely, that some belief or utterance of S has the same content
as ‘p’. The logical forms of ‘S believes that p’ and ‘S said that p’ would then be ‘∃x(Bxs & Sx‘p’)
and ‘∃x(Uxs & Sx‘p’)—there is something such that it is a token belief/utterance of S and it has
the same content as the sentence token ‘p’—where the two-place predicate ‘Sxy’ symbolizes the
sameness of content or synonymy relation and its second place is occupied by a name for the
embedded sentence ‘p’ (not for a proposition) formed by enclosing it within quotation marks.
There are good reasons to believe that synonymy does play a role in content ascriptions. As
Devitt points out:

The expression ‘that p’, used to ascribe a putative meaning, seems to ascribe that
property in virtue of the fact that its content sentence, ‘p’, has the property or one
very like it. So there always seems to be an “intimate link” between the two
properties. However... we can abstract from this link, treating... ‘that p’ like any
arbitrary expression for a property. (Devitt 1996: 82)
Synonymy seems to play a crucial role in the way a ‘that’-clause specifies a meaning. But the specified meaning is not itself a relation between some belief or utterance and the sentence embedded in the ‘that’-clause. Rather, it is a property that both the belief or utterance and the embedded sentence have independently. It is important to distinguish the specification of a property from a reductive explanation of its nature. For example, I can specify the color of my car by pointing out that it has the same color as the Statue of Liberty. The specified color is a property that both my car and the Statue have. The reductive explanation of what constitutes my car’s color involves neither its relation to the Statue, nor its relation to an abstract object, but rather what wavelengths of light are reflected and absorbed by the surface of my car. Similarly, the content specified by ‘S believes that p’ is a property that both some belief of S and ‘p’ have. But the reductive explanation of what constitutes the content of the belief involves neither its relation to ‘p’, nor its relation to an abstract object, but rather the fundamental properties that explain the causal role of the belief. For example, Nancy’s belief that it will rain plays a causal role in explaining why she picks up an umbrella, so an explanation of what underlying properties constitute the content of her belief must account for its causing Nancy’s behavior—just as a reductive explanation of water as constituted by H$_2$O molecules explains why it boils at 100 degrees Celsius (at sea level).

Donald Davidson (1968) offers a similar analysis how ‘that’-clauses specify meanings. He sympathizes with the idea that ‘Galileo said that the earth moves’ can be paraphrased as “Galileo uttered a sentence that meant in his mouth what ‘The earth moves’ means now in mine”—which asserts “synonymy between utterances” or “samesaying” (Davidson 1968: 140). But he prefers not to have the embedded sentence “sealed in quotation marks” because it is then
“mentioned and not used” (Davidson 1968: 141). So he proposes a modification according to which ‘that’ functions as a “demonstrative singular term” which refers to the used utterance that follows it: ‘S said that p.’, he suggests, is equivalent to ‘S said that. P.’ (Davidson 1968: 142). Are quotation marks really problematic? Consider the following two non-attitudinal meaning reports: (a) ‘La nieve es blanca’ means that snow is white, and (b) ‘La nieve es blanca’ means the same as ‘Snow is white’. The mentioned English sentence in (b) serves the same purpose as the ‘that’-clause in (a): they both specify the same meaning. Just as (b) can paraphrase (a), a mentioned sentence can be part of the correct paraphrase of a ‘that’-clause in an attitude report.¹⁴ I will not press this issue further. The crucial point is that synonymy seems to play a role in the way ‘that’-clauses specify meanings and that it does not require the postulation of propositions qua abstract objects. Some philosophers may be tempted to object that an account of different tokens can have the same meaning requires the postulation of abstract objects, so the proposed analysis of particular content attributions does not truly avoid them. But this objection assumes an implausible Platonic account of synonymy, which I have argued against in the previous section. Additionally, as I argued following Devitt, synonymy only plays the role of specifying a semantic property, it is not constitutive of the semantic property.

The proposal that ‘that’-clauses function as indefinite singular terms in particular content ascriptions is not only a plausible alternative to the proposal that they function as names or referential singular terms. There is reason to believe that it is a more plausible account. If ‘that p’

¹⁴ I am ignoring issues that may arise from context-sensitivity. Suppose I report: ‘Mary said that I am rich’. According to Davidson, I am using the embedded sentence ‘I am reach’. My use of the sentence in this context would fix its content: I am talking about me. Does the view according to which the embedded sentence is mentioned have a problem with context-sensitivity? Perhaps. But a slight modification of the analysis may avoid this problem. What I am reporting may be that some utterance of Mary has the same meaning that ‘I am reach’ would have if I asserted it. I may not be inclined to actually assert it, since I may believe I am not rich.
functions as a definite singular term referring to the proposition \( p \), we should be able to replace it with \textit{the proposition that }\( p \), which is a definite singular term referring to the same alleged abstract object. In some contexts, there seems to be no problem: ‘Nancy believes \textit{that it will rain}’ may be paraphrased as ‘Nancy believes \textit{the proposition that it will rain}’. However, the substitution fails in other contexts: ‘Nancy fears \textit{that it will rain}’ cannot be paraphrased as ‘Nancy fears \textit{the proposition that it will rain}’. Fearing a proposition obviously is not the same as fearing that it will rain. So the propositionalist account faces what Friederike Moltmann calls the “Substitution Problem”: “Even though \textit{believe} allows for a replacement of a \textit{that}-clause by \textit{the proposition that }\( S \), many other attitude verbs don’t” (Moltmann 2003: 82). In contrast, I suggest, the proposal that ‘that’-clauses function as indefinite singular terms avoids the substitution problem altogether, since it does not regard them as referring to propositions: ‘Nancy fears \textit{that it will rain}’ should be \textit{not} be paraphrased as ‘Nancy fears \textit{the proposition that it will rain}’, but rather as (1) ‘Some fear of Nancy has the same content as ‘It will rain’’ or (2) ‘Some fear of Nancy has the content specified by ‘that it will rain’’, which are not problematic. The form of (1) is \( \exists x(Fnx \ & \ Sx'r') \)—where \( Fxy \) symbolizes ‘\( x \) is a fear of \( y \)’, \( n \) names Nancy, \( Sxy \) symbolizes ‘\( x \) has the same content as \( y \)’ and \( r' \)—enclosed within quotation marks—stands for the sentence ‘It will rain’ (not a proposition): there is something such that it is a fear of Nancy and it has the same content as ‘it will rain’. The form of (2) is \( \exists x(Fnx \ & \ Rx) \)—where \( R \) is a one-place predicate symbolizing the monadic semantic property specified by ‘that it will rain’: there is something such that it is a fear of Nancy and it has the content specified by ‘that it will rain’. Both (1) and (2) have the same logical forms as the accounts I have offered for belief ascriptions. Neither involve any reference to propositions, so the substitution problem does not
arise. While the propositionalist analysis is undermined by the substitution problem, the account of ‘that’-clauses as indefinite singular terms provides a uniform and plausible account of content attributions involving all attitude verbs that avoids the problem altogether.

A propositionalist, however, may object that regarding particular attributions as quantifying over token sentences and beliefs is also compatible with the contents of such tokens being specified as abstract objects (or relations to abstract objects). The correct paraphrase of ‘Nancy fears that it will rain’ may be claimed to be (3) ‘Some fear of Nancy has the proposition that it will rain as its content’. The form of (3) is $\exists x(Fxn \& Cx<r>)$, where ‘$Cxy$’ stands for “$x$ has $y$ as its content” and ‘$<r>$’ is a singular term referring to a proposition qua abstract object: there is something such that it is a fear of Nancy and its content is the proposition that it will rain. Notice that the form of (3) differs from the form of the non-propositionalist (1) above only in its second conjunct. Basically, (3) makes use of an analysis like (1) and flips it into a propositionalist one. While the standard propositionalist analysis is undermined by the substitution problem, this variant seems to avoid it: the paraphrase of ‘Nancy fears that it will rain’ as (3) does not seem problematic. On this regard, (1) and (3) seem equally plausible. So the substitution problem does not settle the issue of what is ascribed in ordinary attitude ascriptions. However, the modified propositionalist analyses has a serious problem. If the content of Nancy’s fear is an abstract object, it cannot play a causal role in the explanation why she is constantly looking at the sky. I will get back to this issue. But notice that attitude ascriptions are successfully used by ordinary people to explain each other’s behavior. This provides a strong reason for preferring a non-propositionalist analysis like the one I am suggesting over any propositionalist variant.

15 A similar concern is expressed by Devitt (2006: 143-144).
1.4.2 General Content Attributions

So far I have focused on isolated *particular* content attributions. But another influential and seemingly stronger case for regarding contents as abstract objects relies on the logical form of *general* content attributions and the validity—and presumed soundness—of certain *inferences* linking them to particular attributions. General content attributions contain phrases of the form ‘something S said’ or ‘everything S believes’ which quantify over what is said or believed. Consider the following inference patterns:

(1) S believes *that* p and R believes *that* p. So, *there is something* that both S and R believe.

(2) S believes *everything* that R says. R says *that* p. So, S believes *that* p.\(^\text{16}\)

As Stephen Schiffer and Paul Horwich argue, at *face value* the existential quantifier ‘there is something’ and the universal quantifier ‘everything’ in the general content attributions seem to range over a domain of *objects*, which would require—for the inferences to be valid—the ‘that’-clauses in the particular content attributions to function as *referential* singular terms picking out objects in the same domain (Schiffer 1992: 504-506; Horwich 1998b: 86-90). In other words, the logical form of (1) seems to be the valid existential generalization ‘\(Bs<p> \& Br<p> // ∃x(Bsx \& Brx)\)’ and the logical form of (2) seems to be the valid universal instantiation ‘\(∀x(Srx \rightarrow Bsx) / Sr<p> // Bs<p>\)’—where ‘∃x’ and ‘∀x’ are standard objectual quantifiers ranging over propositions.\(^\text{17}\) If the face value interpretation is correct, what the ‘that’-clauses purport to name and what the quantifiers purport to range over are propositions. If we *additionally* grant that some instances of these valid patterns are sound, it follows that propositions exist. For example,

---

\(^{16}\) These are the patterns underlying a couple of the examples offered by Schiffer (1992: 505). Horwich (1998b: 86-87) offers similar examples.

\(^{17}\) I am following here the *standard* analysis of particular attributions adopted by Schiffer and Horwich.
the soundness of ‘Copernicus and Galileo believed that the Earth moves, so there is something they both believed’ would establish that the proposition believed by Copernicus and Galileo is a real “entity” or object (Horwich 1998b: 90). Nevertheless, there are alternatives to the face value interpretation and, even if the face value interpretation of the validity of patterns like (1) and (2) were correct, it would remain to be seen whether they have any literally sound instances.

One alternative to the face value interpretation is the proposal that the quantifiers used in general content attributions are substitutional rather than objectual (Schiffer 1987). It may be argued, for example, that quantifiers have two distinct uses or meanings in ordinary language: one is the standard ontologically committal use, where they function as objectual quantifiers that range over entities, while the other is a metaphysically innocent use, where they function as substitutional quantifiers that do not range over a “language-independent domain of entities” but instead “generalize over the instances” of expressions in the language (Hofweber 2006: 166). An example of the first kind of use would be ‘There is something that hit the window’, which clearly asserts that certain object exists. In contrast, ‘There is something that both S and R believe’ would be, on this view, an example of the second kind of use, asserting that there is some true substitution instance of ‘S and R believe that x’ (in which ‘x’ is replaced by a declarative sentence). Instead of asserting that an object exists, the quantified sentence is equivalent to an infinite disjunction of its substitution instances: either S and R believe that p or S and R believe that q or S and R believe that r, etc. (Hofweber 2006: 167-168). Thus, it may be claimed that propositions are “shadows of sentences” rather than objects (Hofweber 2006: 194). Notice that if the quantifiers in (1) and (2) are non-objectual, the ‘that’-clauses need not be referential singular terms for the inferences to be valid, since non-objectual quantifiers can generalize over the
instances of non-referential expressions.\footnote{Consider a singular term ‘t’ related to a predicate ‘F’ in the sentence ‘t is F’. Using objectual quantification to infer ‘something is F’ from ‘t is F’ requires ‘t’ to be referential. But any singular term in ‘t is F’ can be generalized into ‘something is F’ using substitutional quantification: “In the particular quantifier case: ‘F(t)’ implies ‘Something is F.’... [It] does not matter what ‘t’ is as long as it is grammatically a singular term. In particular, whether or not ‘t’ is a referring expression and whether or not it succeeds in referring, even if it tries, is irrelevant...” (Hofweber 2009: 277). So valid inferences like ‘7 is a prime, so something is a prime’, ‘Red is a color, so something is a color’ and ‘It is true that snow is white, so something is true’, can be regarded as not committed to numbers, properties and propositions being objects—as long as the expressions ‘7’, ‘red’ and ‘that snow is white’ in the premises are non-referential and the quantifiers in the conclusions are non-objectual (Hofweber 2005, 2006 & 2009).} ‘That’-clauses may be, as suggested above, indefinite singular terms specifying the semantic properties of concrete utterances and thoughts.

Whether there are non-objectual uses of quantifiers in ordinary language is controversial, but there is another alternative that relies only on standard quantification. The proposal is that, while the quantifiers in general content attributions do purport to range over objects, there are no real objects in this domain, because propositions are merely fictional entities. What the proponent of this view needs to explain is how a person uttering ‘There is something that both Copernicus and Galileo believed’, quantifying over propositional objects, can convey a truth even when there are no such objects and the sentence is literally false. Notice that, on this view, the sentence uttered is as false as ‘There is something that possessed Mary Magdalene’, which quantifies over demons. Yet quantifying over propositions, unlike quantifying over demons, seems to be “explanatorily useful, generally successful, and needed” (Devitt 1996: 212-213). Are people who engage in propositional talk systematically in error or superstitiously committed to non-existing objects if such talk is indeed literally false?

An explanation of how a literally false sentence can be used to convey a truth comes from the very proponent of the view that a sentence is ontologically committed to the objects it quantifies over (Quine 1961: 13-14). Quine suggests that people are not committed to the objects
presupposed by a sentence when they use it as an “avoidable manner of speaking” (Quine 1961: 13)—that is, when they would be prepared to replace it by a paraphrase that serves the same purpose but is “innocent of such commitments”, so that “the seemingly presupposed objects may justly be said to have been explained away as convenient fictions” (Quine 1961: 103-104).19

Basically, a false sentence that quantifies over non-existing objects can convey a truth when it is used as a manner of speaking which can be paraphrased into a true sentence that does not quantify over such objects. If propositional talk were an avoidable manner of speaking, people who engage in it would not be committed to the existence of propositional objects: they would be quantifying over convenient fictions in literally false sentences that (sometimes) serve the purpose of conveying truths—a purpose that can also be served by paraphrases that do not quantify over propositions but express the same truths literally.

Stephen Yablo elaborates further on how fictional talk can be useful. He proposes that figurative or metaphorical uses of language are make-believe games in which the participants pretend or imagine that certain fictions are true, not for the game’s own sake, but to “talk about game-independent reality”: the fictions serve the purpose of describing features of the real world (Yablo 1998: 246). For example, by pretending or imagining that Italy is a boot, we can accurately describe the real location of the town of Crotone using the sentence ‘Crotone is on the arch of the Italian boot’ (Yablo 1998: 252-253).20 Yablo proposes that one of the hints that a sentence is used metaphorically is its paraphrasability: metaphors are “oftentimes paraphrasable

19 The original sentence and its paraphrase must “serve the same purpose” but cannot be synonymous: “if they were synonymous, it is hard to see how they could differ in their ontological commitments” (Devitt 2010: 25).

20 Obviously, figurative uses of language can also fail to accurately describe reality (e.g., ‘Crotone is on the heel of the Italian boot’). A metaphorical sentence conveys a truth only when the fictions pretended to be true successfully track features of the real world: “the reason for the pretence [is] to portray the world as holding up its end of the bargain, by being in a condition to make a pretence like that appropriate” (Yablo 1998: 248).
away with no felt loss of subject matter” (Yablo 1998: 259n). The metaphorical content of a sentence, he suggests, coincides with the literal content of its paraphrase: “a paraphrase of $S$ expresses in literal terms what $S$ says metaphorically” (Yablo 1998: 248n). If propositional talk were figurative rather than literal, people quantifying over propositions would be pretending or imagining that these fictional objects exist and thereby describing features of the real world.

A very plausible case can be made that propositional talk should be regarded as a figurative and avoidable manner of speaking. As Devitt points out, paraphrases for general content attributions that quantify over propositions are readily available (Devitt 1996: 213). For example, (a) and (b) can be paraphrased as (a') and (b'), respectively:

(a) There is something that both Copernicus and Galileo believed.

(a') Some belief of Copernicus had the same content as some belief of Galileo.

(b) Nancy believes everything Mary says.

(b') When Mary assertively utters any sentence, Nancy forms a belief with the same content.

The paraphrases quantify over sentence or belief tokens, rather than propositions, but they serve

---

21 Yablo proposes several other “hints” of metaphoricality and does not take paraphrasability to be a necessary condition, since he argues that some metaphors are “representationally essential and so not paraphrasable away” (Yablo 1998: 255). Yablo also argues that there is no sharp demarcation between the literal and the metaphorical—“the boundaries of the literal are about as blurry as they could be, the clear cases on either side enclosing a vast interior region of indeterminacy” (Yablo 1998: 233)—and, furthermore, that we may not be able to completely eliminate metaphors from science—“our best theory may well include metaphorical sentences (whose literal contents are) not meant to be believed” (Yablo 1998: 255). If correct, these radical theses would undermine the Quinean program of finding answers to existence questions in the ontological commitments of our best scientific theories, literally formulated. This is not the place to discuss these radical theses. The Quinean philosopher, however, can go along with Yablo’s characterization of paraphrasable figurative talk, which offers an interesting elaboration of the Quinean theme that manners of speaking make use of convenient fictions.

22 Yablo claims that this explains “rather neatly” how a “paraphrase ‘says the same’ as what it paraphrases”, even though the literal meanings of the two sentences do not coincide (Yablo 1998: 248n).

23 This paraphrase is based on one suggested by Devitt (1996: 13).
the same purposes as the original sentences they paraphrase (Devitt 1996: 213). First, (a') and (b') serve the same communicative purposes as (a) and (b). Second, they serve the same explanatory purposes. For example, (b') serves the same purpose as (b) in the explanation of Nancy’s picking up an umbrella after hearing Mary utter ‘It will rain’. Notice that the suggested paraphrases assert that pairs of sentence or belief tokens have the same content. This neatly matches the view that particular content attributions assert that some sentence or belief token has the same content as the sentence embedded in the ‘that’-clause.\footnote{I already argued in support of this view. However, if I am wrong and either the standard or the modified propositionalist analysis of particular content attributions is right, then particular attributions can also be treated as mere \textit{manners of speaking} (Devitt 1996: 214). The analysis of particular attributions I proposed would then offer a literal paraphrase of what is metaphorically conveyed in ordinary language.} Furthermore, the validity of inferences linking general and particular content ascriptions is preserved by their paraphrases. Consider the following paraphrased instances of patterns (1) and (2):

(i) Some belief of Copernicus and some belief of Galileo had the same content as ‘The Earth moves’. So, some belief of Copernicus had the same content as some belief of Galileo.

(ii) When Mary assertively utters any sentence, Nancy forms a belief with the same content. Some sentence assertively uttered by Mary has the same content as ‘It will rain’. So, some belief of Nancy has the same content as ‘It will rain’.

These inferences are obviously valid because having the same content is an \textit{equivalence relation}: it is symmetric, transitive and reflexive. The valid logical form of these two inferences can be represented in standard first-order logic by adding meaning postulates defining the two-place predicate \(S_{xy}\) (“\(x\) has the same content as \(y\)”) as symmetric = \(∀x∀y(S_{xy}→S_{yx})\), transitive = \(∀x∀y∀z((S_{xy} & S_{yz}) → S_{xz})\) and reflexive = \(∀x(S_{xx})\). A shorter definition of \(S_{xy}\) as an equivalence relation is that it is Euclidean = \(∀x∀y∀z((S_{xy} & S_{xz}) → S_{yz})\) and reflexive. So the
two inferences can be represented, I suggest, as having the following valid forms:

(i) \[ \forall x \forall y \forall z [(S_{xy} \& S_{xz}) \rightarrow S_{yz}] \& \forall x (S_{xx}) \]
\[ \exists x \exists y [(B_{xc} \& S_{x'e'}) \& (B_{yg} \& S_{y'e'})] \]
\[ \therefore \exists x \exists y [(B_{xc} \& B_{yg}) \& S_{xy}] \]

(ii) \[ \forall x \forall y \forall z [(S_{xy} \& S_{xz}) \rightarrow S_{yz}] \& \forall x (S_{xx}) \]
\[ \forall x (U_{xm} \rightarrow [\exists y (B_{yn} \& S_{xy})]) \]
\[ \exists x (U_{xm} \& S_{x'r'}) \]
\[ \therefore \exists x (B_{xn} \& S_{x'r'}) \]

Where ‘c’ = ‘Copernicus’, ‘g’ = ‘Galileo’, ‘m’ = ‘Mary’, ‘n’ = ‘Nancy’, ‘e’ = the sentence ‘The earth moves’, ‘r’ = the sentence ‘It will rain’, ‘B_{xy}’ = ‘x is a belief of y’, ‘U_{xy}’ = ‘x is an utterance of y’, ‘S_{xy}’ = ‘x has the same content as y’, and ‘\forall x’ and ‘\exists x’ range over sentence and belief tokens. Notice that the first premise in both inferences defines ‘S_{xy}’ as an equivalence relation. Both inferences are formally invalid if this premise is omitted.

Alternatively, we can paraphrase the particular content ascriptions in a way that does not explicitly assert synonymy or sameness of content:

(i’) Some belief of Copernicus and some belief of Galileo had the content specified by ‘that the earth moves’. So, some belief of Copernicus had the same content as some belief of Galileo.

(ii’) When Mary assertively utters any sentence, Nancy forms a belief with the same content. Some sentence assertively uttered by Mary has the content specified by ‘that it will rain’. So, some belief of Nancy has the content specified by ‘that it will rain’.
These inferences are also intuitively valid and may be interpreted as having the same logical forms as (i) and (ii). If we interpret instead the specified contents as monadic properties, the inferences are also formally valid as long as we regard them as enthymemes assuming the obvious premise that two tokens are synonymous (a dyadic relation) if and only if whatever contents each of them has (their monadic properties) are the same. Symbolizing this unstated premise requires substitutional quantification (or the ontologically problematic quantification over properties), since it generalizes over (monadic) content properties. But we can represent in standard first-order logic instances of the connection between synonymy and particular specified contents. So (i′) and (ii′) can be represented, I suggest, by the following valid forms:

(i′) \[∀x∀y[(Ex & Ey) → Sxy]\]
\[∃x∃y[(Bxc & Ex) & (Byg & Ey)]\]
\[∴ ∃x∃y[(Bxc & Byg) & Sxy]\]

(ii′) \[∀x∀y[(Sxy & Rx) → Ry]\]
\[∀x\{Uxm → [∃y(Byn & Sxy)]\}\]
\[∃x(Uxm & Rx)\]
\[∴ ∃x(Bxn & Rx)\]

Where ‘E’ = ‘has the content specified by ‘that the earth moves’’, ‘R’ = ‘has the content specified by ‘that it will rain’’, and the rest of the symbols are the same as in (i) and (ii) above. The first premises of (i′) and (ii′) assert intuitively obvious connections between synonymy and particular specified content properties. Without these added premises, both inferences are formally invalid.

I proposed before a non-propositional analysis of particular content attributions. I have
suggested now that general attributions can be treated as figures of speech whose metaphorical content does not quantify over propositions and is literally expressed by suitable paraphrases. I have also shown that the validity of inference patterns like (1) and (2) can be accounted for by combining non-propositionalist accounts of particular and general content attributions.\textsuperscript{25}

1.4.3 Folkloric Opinion and Ontology

From a strictly Quinean point of view, the availability of paraphrases for content ascriptions that quantify over propositions does not, by itself, free a person who engages in propositional talk from an ontological commitment to abstract objects. According to Quine, this depends on whether the person agrees to replace the problematic sentences by their paraphrases in order to “free himself from ontological commitments of his discourse” (Quine 1961: 103). It is an interesting empirical question whether ordinary people would be prepared in general to accept the suggested paraphrases. I suspect they would, since the paraphrases clearly serve the same purposes as what they paraphrase. But establishing that people are genuinely committed to propositions qua abstract objects perhaps requires more than a rejection of these paraphrases: what would we conclude if they also reject alternative paraphrases that are more explicitly Platonic, such as ‘There is an abstract object that exists outside of the world of space and time and which is a proposition believed by Copernicus and Galileo’? I doubt that ordinary people would in general be prepared to accept such explicitly Platonic paraphrases. But, of course, this is an empirical matter. In any case, Quine points out that ordinary people’s ontological

\textsuperscript{25} I also suggested that the quantifiers in general content ascriptions may generalize over their substitutional instances without ranging over propositions. This alternative also preserves the validity of the inferences. However, it is controversial whether substitutional quantifiers are used in ordinary language. The fictionalist account is simpler, since it relies only on standard first-order logic.
commitments are often unclear when they talk about “things” beyond well-defined objects: “The common man’s ontology is vague and untidy... It takes in many purported objects that are vaguely or inadequately defined... [We] cannot even tell in general which of these vague things... to count him as assuming.” Quine concludes: “We must recognize... that a fenced ontology is just not implicit in ordinary language.” (Quine 1981a: 9) Similarly, Yablo points out that people often talk figuratively without being fully conscious of it (Yablo 1998: 246) and, furthermore, that they often “fall back” on the metaphorical content of areas of discourse, with “no conscious equivocation”, when they find out that the objects talked about do not exist or are problematic (Yablo 1998: 258n). Perhaps people are simply non-committal about the ultimate ontology of representational contents.

Suppose, for the sake of argument, that people were seriously committed to propositions qua abstract objects: that they preferred explicitly Platonic paraphrases of general attributions over the non-propositional paraphrases. This interesting anthropological fact would not establish that propositions qua abstract objects exist, just like the real commitment that many people have to immaterial souls does not establish that these entities exist. People talk not only about immaterial souls, but also about witches, ghosts, gods and whatnot. Non-metaphorical and sincere superstitious talk is ontologically committed to the existence of all sorts of things that clearly do not exist and this may also be the case for talk that quantifies over propositions (Devitt 1996: 212-213). If propositions are mere fictions—and this cannot be ruled out merely by an analysis of folkloric opinion—people actually committed to them are in error. In that case, instead of adopting a hermeneutic fictionalism—according to which propositional discourse is used figuratively—we should defend a revolutionary fictionalism—according to which
propositional discourse *ought to* be used only figuratively—. The claim would be that people *ought to* accept paraphrases like those suggested above for general content ascriptions. Anyway, as I pointed out above, I doubt people are genuinely committed to propositions *qua* abstract objects, since they may accept the proposed paraphrases or simply be non-committal; so this is only a “worst case scenario” that I have granted for the sake of argument.

In this section, I have presented arguments for propositions *qua* abstract objects based on the alleged ontological commitments of ordinary attitude ascriptions. I have argued that there are plausible alternative accounts of such ascriptions that avoid any commitment to abstract objects and that, anyway, the issue of whether meanings are abstract objects cannot be settled merely by an account of folkloric opinion. In the next section, I will argue that there is a compelling reason to reject the view that meanings are abstract objects.

### 1.5 Against Meanings *qua* Abstract Objects

In this section, I argue that the meanings of utterances and thoughts are *not* Platonic abstract objects. The argument is not based on an analysis of the semantic *status quo*, on what people *do ascribe* when they ascribe meanings, but rather on what we *ought to ascribe* in order to explain the causal roles of meanings (Devitt 1996: 61-62). I do point out, however, that the general success of explanations of behavior in folk-psychology suggests that at least part of what people do ascribe is, indeed, what they ought to ascribe. Once meanings *qua* abstract objects have been rejected, I suggest, we can either reject propositions altogether—except as “convenient fictions”—or, perhaps, attempt to naturalize them. Devitt (1996) proposes the former of these

---

26 The distinction between hermeneutic and revolutionary fictionalism comes from Stanley (2001).
alternatives, while Armstrong (2004) suggests the latter. I will discuss in the following section whether there is a need for “naturalized propositions” once propositions qua abstract objects have been rejected.

There is a compelling metaphysical reason to suppose that semantic properties of sentence and belief tokens we ought to ascribe are natural worldly properties of each of these tokens. Abstract objects cannot causally act upon any concrete things in the world, since “we have fairly good scientific reasons to believe that Nature, the spatio-temporal system, is a causally self-enclosed system” (Armstrong: 1978: 129). The causal closure of the physical world undermines, for example, the hypotheses that there is a God that acts upon the world or that we have non-physical souls that interact with our bodies. According to such hypotheses, some physical events have non-physical causes. But that violates the causal closure principle. Cartesian substance dualism has been largely abandoned for this reason. But the same consideration undermines the hypothesis that abstract objects act upon concrete objects: “If the case for God and the soul acting upon particulars is weak, it is far weaker in the case of the realm of numbers, possible words, propositions and so on.” (Armstrong: 1978: 130).

Those who nowadays postulate the existence of abstract objects often accept that they do not causally interact with the spatio-temporal world. However, this raises the question of why should we postulate them to begin with:

If any entities outside this realm are postulated, but it is stipulated further that they have no manner of causal action upon the particulars in this realm, then there is no compelling reason to postulate them. Occam’s razor then enjoins us not to

Basically, no explanatory gain comes from the postulation of abstract objects and standard simplicity considerations count against postulating explanatorily idle entities.

The same considerations apply to the postulation of propositions \textit{qua} abstract objects that constitute the meanings of sentence and belief tokens and that are the primary truth-bearers. Abstract objects cannot causally interact with concrete—datable and placeable—utterances and thoughts, so they cannot determine their semantic properties. As Devitt points out, Platonic objects “can play no causal role in mind and language”, so “we have the best of reasons for thinking that they are not part of mental and linguistic reality.” (Devitt 1996: 210). Armstrong rejects the postulation of propositions \textit{qua} Platonic objects for the same reason:

There are metaphysicians who are prepared to postulate a realm of propositions over and above the space-time world. But, presumably, we could not stand in any causal or nomic relation to such a realm. And if we cannot stand in such relations to propositions it is unclear that such a postulation is of any explanatory value. (Armstrong 2004: 12).

That propositions \textit{qua} abstract objects cannot act upon the world is rarely disputed nowadays. But the Platonic propositionalist may be willing to accept that propositions do not causally interact with concrete utterances and thoughts, which amounts to acknowledging that they cannot play any explanatory role in mind and language. If no explanatory gain comes from postulating propositions, however, there is no compelling reason to postulate them in the first place.

Accepting that propositions do not causally interact with utterances and thoughts, while
claiming that they constitute their meanings, amounts to regarding meanings as causally inert epiphenomena. This is deeply troubling, since we postulate meanings or contents in order to explain people’s behavior (Devitt 1996: 57-60). Ascribing to Mary a belief that it is raining, a belief that using an umbrella can help her not to get wet, and a desire not to get wet, plays a crucial role in the explanation of why she picks up an umbrella. Similarly, ascribing to Mary’s utterance of ‘It is raining’ the meaning that it is raining plays a crucial role in ascribing to her the belief that it is raining, which in turn plays a crucial role in explaining her behavior: why she picks up an umbrella. Had the contents of her beliefs, her desire and her utterance been different, we would reasonably expect her behavior to vary accordingly.

The causal closure of the physical world provides a compelling reason for regarding the semantic properties we ought to ascribe to explain behavior as natural worldly properties, rather than relations to abstract objects. Moreover, ordinary people do ascribe contents to explain other people’s behavior and their explanations are often successful, which provides evidence that the meanings ascribed in folk-psychology do play a causal role in the physical world. In other words, there is strong evidence that at least some of what we do ascribe is what we ought to ascribe (Devitt 1996: 81). While there is plenty of controversy in the field of naturalized semantics about which is the best reductive explanation of what constitutes meanings, it is clear that meanings must play a causal role in the psychological explanation of behavior. Consequently, the view that meanings are abstract, mind and language independent objects, must be rejected. Does this mean that we should reject propositions or that we should attempt to provide a non-Platonic account of propositions?

27 As Devitt (1996: 57-60) points out, another crucial purpose for which we ascribe meanings is to learn about the world. When a person’s thoughts or utterances are reliable, we can learn about the world by ascribing the right meanings or contents to them. It is unclear how this would be possible if meanings were abstract objects.
In Quinean fashion, Devitt prefers to avoid talk of propositions altogether, and talks instead of meaningful sentence and belief tokens:

By talking of meaningful tokens, we can avoid talk of propositions. This is an advantage not simply because propositions are creatures of darkness but also because talk of propositions... is explanatorily unhelpful. (Devitt 1996: 13)

Devitt admits that propositional talk may be legitimate to the extent that such talk can be “paraphrased away, when the ontological chips are down, into talk about the properties of concrete thoughts and utterances” (Devitt 1996: 210). Although he warns that, even then, propositional talk is “unnecessary and misleading” (Devitt 1996: 210). Presumably, it is unnecessary because it is an avoidable manner of speaking and it is potentially misleading because there is the risk of taking literally what is only useful as a figure of speech.

Armstrong, in contrast, does not want to reject propositions altogether. Instead, he proclaims: “as a naturalist, I want to look for a this-worldly account of propositions” (Armstrong 2004: 12). His proposal is that propositions are indeed the contents of beliefs and the meanings of sentences, but that they are properties to be explained by a naturalized semantics, rather than Platonic objects: “I would be hoping for a naturalistic theory of content and meaning, and so a naturalistic theory of... propositions.” (Armstrong 2004: 14) The difference between the approaches of Armstrong and Devitt may seem merely verbal. After all, they both regard meanings as natural properties. Yet, there is a genuine difference. Armstrong, unlike Devitt, argues that propositions—not sentence and belief tokens—are the primary truth-bearers. (Armstrong 2004: 12). Devitt’s view is best characterized as treating meaningful tokens as the
primary truth-bearers—tokens that are true or false *in virtue of* their semantic properties. Armstrong even proposes that propositions are *abstractions* but, he insists, “not in any other-worldly sense” (Armstrong 2004: 13). He gives the following reason:

That... content or meaning is an abstraction becomes clear when we notice that contents and meanings are types rather than tokens. Beliefs in different minds may have the very same content, numerically different statements may have the very same meaning. (Armstrong 2004: 13)

Basically, Armstrong thinks there is a “One over Many” problem for synonymy, so he suggests an Aristotelian immanent realism, according to which propositions are worldly but universal semantic property types that exist in sentence and belief tokens.

The disagreement between Armstrong and Devitt on the nature of meanings/propositions stems from their disagreement on property realism. If there is a genuine “One over Many” problem for synonymy, an immanent realism about meaning is preferable to a transcendent realism. If the “One over Many” is instead a pseudo-problem, there is no reason for regarding meanings as worldly but universal properties: we can simply regard them as the particular properties of concrete tokens. In the latter case, there is no need for propositions—except as “convenient fictions” that we may talk about figuratively to convey semantic facts about concrete utterances and thoughts. But is there a need for *propositions* in the former case? Why not just talk about *meanings* (*qua* immanent properties)? Armstrong offers the following reason for positing *propositions*: “We can certainly apply the truth predicates very widely, but I am inclined to think that all other suggested truthbearers besides propositions are called truthbearers...
on account of their relationship to certain propositions.” (Armstrong 2004: 12). In other words, he agrees with the need for “something” that is not only the common meaning of synonymous tokens, but also the primary bearer of truth. Such a “thing” is a universal property, rather than an abstract object, but plays the same roles the alleged abstract object was expected to play.

I have argued that abstract objects cannot play the causal roles meanings ought to play, so meanings are not abstract objects. I pointed out that we may then reject the existence of propositions, following Devitt, or attempt to naturalize them, following Armstrong. But do naturalized meanings play the same roles Platonic propositions where expected to play? Are they also the primary truth-bearers? I discuss this topic in the next section.

I end this section with a brief clarification regarding the status of abstract entities in general. There is an influential argument that abstract mathematical objects exist because our best scientific theories quantify over them. This is known as the “Quine-Putnam indispensability argument” (Quine 1961; Putnam 1979). It may be objected that my case against propositions qua abstract objects would equally apply to mathematical objects, yet the indispensability argument shows that mathematical objects exist—so there must be something wrong with my case. This is an extremely difficult issue. However, the objection is far from compelling. First, notice that the indispensability argument, while popular, is not the last word on the issue. There is Hartry Field’s (1980) ambitious though incomplete project of nominalizing physics. According to Field, numbers are merely useful fictions, reference to which can in principle be eliminated without loss from our best scientific theories. But suppose that reference to numbers cannot be eliminated. This alone does not show that they are not fictions. Stephen Yablo (1998) argues that abstract entities are fictions even if reference to them cannot be eliminated from science: we may
need to learn to live with some in ineliminable metaphors. There is plenty of debate on this issue, so it is far from established that non-fictional numbers are indispensable. Second, and more importantly, I have argued that proper paraphrases for propositional talk are readily available. So the situation of propositional talk is very different from that of mathematical talk. Even if some abstract entities such as numbers must be accepted, this provides no reason for regarding propositions as real abstract entities. Finally, numbers in physics do not themselves play causal roles. No serious contemporary Platonist about numbers claims that they have physical effects. As Mark Balaguer points out: “there are no causally efficacious mathematical objects... if there exist any... then they are causally inert... [since] they exist outside of spacetime...” (Balaguer 1998: 132). Whatever role reference to numbers has in scientific explanations, it is very different from that of reference to electrons and quarks. The fact that numbers are causally inert is what makes Platonism about them a plausible position. But the situation is very different with meanings, which clearly have causal roles in the physical world. In contrast with Platonism about numbers, Platonism about meanings clashes with the causal closure of the physical world. Consequently, it is simply not a plausible position.

1.6 Naturalized Propositions?

Armstrong attempts to naturalize propositions, regarding them as worldly but universal semantic properties that also play the role of being the primary bearers of truth. I shall argue in this section that, even if Armstrong were right about meanings being immanent universals, it would not follow from this alone that they must also be the primary truth-bearers. I will not argue, however, that meanings are not the primary truth-bearers. Rather, I will only attempt to clarify what we
would need to establish in order to determine whether they are the primary truth-bearers.

It is far from clear that meanings should play the role of primary truth-bearers once we have rejected the account of meanings qua abstract objects and replaced it with an account of meanings as worldly properties—whether universal or not. The semantic properties of sentence and belief tokens surely are partly responsible for their truth. A Spanish speaker’s utterance of ‘La nieve es blanca’ and someone’s belief that snow is white are true because they both mean that snow is white—and because snow indeed is white. But it does not follow from this fact alone that the real bearers of truth are meanings themselves, rather than meaningful sentence and belief tokens—that is to say, concrete utterances and thoughts with their semantic (and syntactic) properties. Meanings or contents, abstracted from utterances and thoughts, may not even be legitimate—let alone primary—truth-bearers.28 An analogy may help to clarify this point:

A boat has the property of positive buoyancy—of weighting less than the water it displaces—and, due to this, it has the property of floating in water. But what floats is the boat, not the property that allows it to float. While other objects also float because they are positively buoyant, it does not follow that the property they have in common floats. Positive buoyancy does not float, positively buoyant objects float.

28 J. L. Austin argues, based on ordinary language considerations, that propositions—regarded as sentence meanings—cannot be truth-bearers: “In philosophy, indeed, ‘proposition’ is sometimes used... for ‘the meaning or sense of a sentence or family of sentences’ : but... a proposition in this sense cannot, at any rate, be what we say is true or false. For we never say ‘The meaning (or sense) of this sentence (or of these words) is true’: what we do say is... that ‘The words taken in this sense, or if we assign to them such and such a meaning, or so interpreted or understood, are true.’ ” (Austin 1950: 113). According to Austin, the primary truth-bearers are neither propositions nor sentence types (which he calls simply “sentences”), but rather the datable and placeable events he calls “statements”: “A statement is made and its making is a historic event, the utterance by a certain speaker or writer of certain words (a sentence)...” (Austin 1950: 113-114). I do not think that ordinary language considerations settle the issue. Yet, Austin’s take is worth mentioning because many authors take ordinary language to be committed to propositions—qua sentence meanings—being the primary truth-bearers.
Similarly, it may be reasonably argued:

A token of ‘La nieve es blanca’ in Spanish has the property of being true because it has the property of meaning that snow is white—and snow indeed is white. But what is true is the token, not the property that allows it to be true. While there are other sentence and belief tokens that are true also because they mean or have the content that snow is white, it does not follow that the semantic property they have in common is true. Meanings or contents are not true or false, *meaningful sentence and belief tokens* are true or false.

My point here is not that meanings *cannot* have the property of being true or false simply because they are properties. After all, some properties may themselves have properties. For example, the property of being red may have the property of being a color. What the analogy with positive buoyancy shows, rather, is that an object $O$ having property $X$ in virtue of having property $Y$ does not entail that $X$ must be a property of $Y$ itself in order to be a property of $O$. So, truth-values being properties of meanings and, consequently, sentence and belief tokens being only derivatively true or false, does *not* follow *merely* from meanings being properties in virtue of which sentence and belief tokens can be true or false.

To establish that meanings are the primary truth-bearers, we would need to show that the relation between meaning and truth is more like the relation between being red and being a color.  

29 This is, of course, controversial. Armstrong argues that the truth of statements like ‘Red is a color’ requires properties to be real: nominalists who reject properties cannot account for such truths (1978: 58-63). While he prefers an account in terms of the universal properties, he points out that a *particularist* account—according to which “the properties of particulars are themselves particular” (Armstrong 1978: 59)—does not face this problem. ‘Red is a color’ may be analyzed by the particularist as asserting the following: if any particular thing has the particular property of being red, then it has the particular property of being colored, and being colored is a second-order particular property of the first-order particular property of being red (Armstrong 1978: 61). Devitt regards the issue of accounting for the truth of statements like ‘Red is a color’ as a genuine problem, but points out that it is different from the “One over Many” problem (2010: 19-20).
than like the relation between positive buoyancy and floating. While this may be the case, Armstrong does not provide any explicit reasons to think so. Why should we regard being true as a second-order property of first-order meaning properties, rather than as a first-order property of meaningful tokens? Armstrong seems to simply assume that meaning properties are the real bearers of truth. But this begs the question against the reasonable alternative that the primary bearers of truth are meaningful sentence and belief tokens.

The only hint of why Armstrong takes meanings qua immanent universals to be the primary truth-bearers can be found on his brief characterization of particular content ascriptions. But we are forced to “read between the lines” in lack of an explicit argument. Armstrong says:

Beliefs are essentially beliefs that something is the case. Whatever is believed to be the case may then be said to be ‘the intentional object of that belief’, using this as a technical term only. And that is a proposition. (Armstrong 2004: 13)

What follows the ‘that’ in a particular content ascription, Armstrong points out, has the form “something is the case”. Surely, then, what follows the ‘that’ can be true or false: it is a bearer of truth. But Armstrong additionally claims: “note that what follows the ‘that’... is a proposition” (Armstrong 2004: 14). If this claim were true, it would follow that a proposition is a truth-bearer in its own right. While Armstrong uses the expression ‘intentional object’ to characterize the proposition, he insists this is merely a technical term, so he warns: “I do not want to read too much metaphysics into the phrase ‘intentional objects’.” (Armstrong 2004: 13) This warning is important, since Armstrong wants to avoid reference to Platonic objects and treat a proposition as a universal but worldly semantic property that is “what makes the belief the particular belief that
it is” (Armstrong 2004: 13). An argument for naturalized propositions as primary truth-bearers we may “extract” from Armstrong’s brief remarks runs roughly as follows: What follows the ‘that’ in a content ascription has the form “something is the case” and is a proposition, so a proposition is a bearer of truth; additionally, a proposition qua immanent meaning property is what makes a belief the particular belief it is, so it follows that it is the primary bearer of truth.

The line of reasoning, if sound, would support the view that being true is a second-order property of first-order meaning properties. But the line of reasoning is not sound. The claim that the content of a belief makes it the belief it is and the claim that what follows the ‘that’ in a ‘that’-clause has the form “something is the case” are surely correct and uncontroversial. But why should we regard what follows the ‘that’ as a proposition? Notice that what follows the ‘that’-clause is obviously not a proposition, but an embedded sentence. What Armstrong must mean is that what follows the ‘that’ somehow “picks out” a proposition. How can it do so? The Platonic account provides an answer I have argued against: the content ascription asserts a relation to a proposition because the ‘that’-clause refers to a universal abstract object. But Armstrong cannot appeal to this answer. Yet, he seems to assume that something similar to the Platonist account is correct. His characterization of a proposition as an intentional object is revealing, despite his warning about reading “too much metaphysics” into it. The only account consistent with Armstrong's own metaphysics is that the ‘that’-clause, instead of referring to a transcendent object, “picks out” an immanent property. However, the ‘that’-clause surely picks out a semantic property only because the embedded sentence has that property. But then the ‘that’-clause specifies the property by means of synonymy, which is the account I have defended. This provides no support for the claim that what follows the ‘that’ has a different status from
what precedes it. The report ‘Galileo believed that the earth moves’ asserts that Galileo’s belief has the same content as the embedded sentence: both tokens have the property and both tokens are true in virtue of having this property—and because the Earth does move. So there is no support for the claim that what follows the ‘that’ is in any sense primarily true. Granting, for the sake of argument, that the meaning shared by what precedes and follows the ‘that’ is a universal does not show that the meaning is a primary truth-bearer either, since this is consistent also with truth being a first-order property of both meaningful tokens. I conclude that Armstrong’s brief characterization of content ascriptions fails to provide support for his claim that naturalized meanings also play the roles of primary truth-bearers and deserve to be regarded as “propositions”.

In this section, I have argued that Armstrong provides no support for his claim that naturalized semantic properties also play the role of being the primary bearers of truth and, consequently, should be regarded as naturalized propositions. I have not argued that meanings are not the primary truth-bearers. However, I have argued that to establish that they are requires determining that being true is a second-order property of first-order meaning properties. While Armstrong does not show that this is the case, it may be the case. Alternatively, it may be the case that the primary truth-bearers are meaningful sentence and belief tokens.

In the next section, I discuss whether postulating propositions qua abstract objects may be consistently combined with a naturalized semantics. A view along these lines is proposed by Horwich. Since this view is at odds with my claim that propositions qua abstract objects are inconsistent with a naturalized semantics (and I borrow some crucial ideas from Horwich’s theory of meaning for my own proposal), it is important that I clarify and assess the relationship
between Horwich’s propositionalism and his semantic naturalism.

1.7 Naturalized Meanings plus Abstract Propositions?

I have argued against the standard account of propositions as abstract objects which constitute the meaning or contents of thoughts and utterances and the primary bearers of truth. The main consideration against such objects is that they cannot play any causal role in language and mind. I have discussed Armstrong’s alternative proposal to naturalize propositions, regarding them as natural properties which play the same roles abstract objects were alleged to play. But there is yet another alternative to the standard account: propositions may be regarded as abstract objects that are the primary truth-bearers, but which do not play the role of determining the meanings of utterances and thoughts. Horwich’s account is precisely of this sort. He proposes a naturalized theory of meaning, which he combines with a deflationary theory of truth that applies primarily to propositions qua abstract objects. In this section, I attempt to clarify his account of the relation between naturalized meanings and propositions. Horwich’s own remarks on this metaphysical issue leave room for interpretation. I argue that under one interpretation, Horwich’s account can avoid the problems of the standard propositionalist account, while under another interpretation his propositionalism is problematic.

Horwich’s “use” theory of meaning identifies meanings with the properties that account for the acceptance of sentences, including linguistic utterances and thoughts (i.e., the tokening of mental sentences in the “belief box”), and explains meanings in terms of some core sentences that, he suggests, play a basic role in the causal explanation of why other sentences are accepted (Horwich 2005: 28). This account of meaning has good naturalistic credentials and is a crucial
component of the approach I am proposing to naturalize semantics. Furthermore, Horwich fully accepts the causal closure of the physical world which, he claims, constrains what properties can constitute meanings:

[A] person’s utterance of a given sentence on some occasion is surely explained, in part, by the meaning he attaches to that sentence; and this can be so (assuming the ‘causal autonomy of the physical’) only if [the] meanings of expressions are somehow engendered by physical phenomena. (Horwich 2010: 103n)

Horwich makes this unapologetic naturalist remark in the context of rejecting accounts according to which meanings are irreducibly normative. He is not specifically invoking the causal closure of the physical world to reject accounts of meanings as engendered or constituted by relations to abstract objects. Yet, the latter accounts are also undermined by this consideration—as I have argued—and Horwich’s theory of meaning does not appeal anyway to non-natural properties.

Horwich’s deflationary theory of truth is that truth is not a natural property susceptible of a reductive explanation. He argues that the function of our concept of truth is merely to allow us to make certain generalizations, that the meaning of the word ‘true’ is engendered by the acceptance of the trivial equivalence schema “<p> is true ↔ p”, and that no reductive explanation of the nature of truth is possible or needed (Horwich 1998b: 120-146). The equivalence schema, Horwich argues, applies primarily to propositions (<p>) and only derivatively to sentences (‘p’): the truth of the sentence ‘Snow is white’, derives from the truth of the proposition it expresses, <Snow is white>, which trivially follows from snow being white (Horwich 1998b: 129-135). His main argument for regarding propositions as objects that are the
primary truth-bearers is based on ordinary language considerations:

In ordinary language what are said to be true are the things that we believe and that our utterances express—so-called propositions. Thus, on the face of it, propositions exist, some of them… are true, and the correlated truth-like attributes of utterances and of acts of believing, asserting, etc. are the complex, derivative properties, ‘u expresses a true proposition’ and ‘the object of act x is a true proposition’. (Horwich, 1998b: 129)

Horwich bases his view that what is said and what is believed are propositional objects on the face value interpretation of inferences linking particular and general content ascriptions (Horwich, 1998b: 86-90). I have already argued against the face value interpretation at length. If I am right, it follows that if deflationism about truth turns out to be correct, a version that takes sentence tokens as the primary truth-bearers should be favored. My concern here is whether Horwich’s propositionalism is compatible with his naturalism about meaning and his acceptance of the causal closure of the physical world.

The first point to notice is that Horwich distinguishes propositions—what is said or believed—from the meanings of sentence and belief tokens and types identified and explained by his naturalized semantics: “Certainly, it is vital to distinguish between (1) the meaning (or meanings) of a sentence type, (2) the meaning of one of its tokens, and (3) the proposition expressed by that token.” (Horwich 1998a: 82). The same distinction can be made between the meaning of a belief type or token and the proposition believed, since Horwich treats beliefs as

---

30 Hartry Field proposes a sententialist version of deflationism which is similar to Horwich’s but does not posit propositions (Field 2001).
mental sentences (Horwich 2005: 30). Since his reductive account applies directly only to the meanings of sentences, this distinction seems to leave room for treating the propositions expressed by those sentences as abstract objects. Indeed, Horwich claims that this is a “small and innocuous” step that follows from introducing propositions as follows: if an utterance has the same meaning as ‘p’, then it expresses the proposition that \( p \) (Horwich 2010: 33). The claim seems to be roughly that the proposition expressed by a sentence token derives from its meaning and does not determine its meaning. This claim does not overtly violate the causal closure of the physical world, as far as meanings are concerned. (It is analogous to an epiphenomenalist account of the mental, which preserves dualism without violating the laws of nature.)

As far as truth is concerned, the order of Horwich’s account is the opposite: the truth of a sentence token derives from the truth of the proposition it expresses. Is this problematic? Does it violate the causal closure of the physical world, since propositions are regarded as abstract objects? Not necessarily. Since truth, according to Horwich, is a trivial property that plays no causal-explanatory role, it may be argued that his account of abstract propositional objects as the primary truth-bearers is harmless and does not violate the causal closure principle. For example, we may explain that Mary succeeded finding a corkscrew because her belief that it was in the kitchen drawer—which caused her to go to look for it there—turned out to be true. While truth seems to play a crucial role in the explanation of why Mary’s behavior was successful, Horwich would insist that the word ‘truth’ is only being used as a “device of generalization” and that we can explain the situation without appealing to truth: Mary’s behavior was successful because it was partly caused by a token of ‘There is a corkscrew in the kitchen’s drawer’ in her “belief box” and there was indeed a corkscrew in the kitchen’s drawer. Horwich would argue that only the
meaningful token—which can be explained naturalistically—and the location of the corkscrew—a physical fact—play a substantial role in the explanation of why Mary acted as she did and succeeded. Horwich’s contention is that truth itself does not play any causal role in the explanation; while we may use the word ‘true’ in describing the situation—as I did initially—and even to formulate the generalization that true beliefs (often) contribute to successful behavior, we only need to apply ‘true’ in accordance with the trivial equivalence schema to do so (Horwich 1998b: 44-46, 139-141).

Let me summarize what I have said so far in this section. I have argued that Horwich’s propositionalism does not overtly conflict with his naturalistic theory of meaning to the extent that (1) he distinguishes meanings from propositions, (2) he does not regard the latter as determining the former, and (3) his deflationary account does not assign any causal role to truth in the explanation of why behavior based on true belief can be successful. However, I have argued previously against his motivation for regarding propositions as the primary truth-bearers: his endorsement of the face value interpretation of attitude ascriptions. I will now discuss in more detail Horwich’s distinction between meanings and propositions.

As I pointed out, Horwich claims that positing an abstract propositional object is a “small and innocuous” step: when a sentence token is synonymous with ‘p’, it expresses the proposition that \( p \) (Horwich 2010: 33). This is the way of introducing propositions I discussed above and found not to conflict with the causal closure principle. However, the devil is in the details. When we take a closer look into Horwich’s distinction between meanings and propositions, we find that he makes some remarks that may be interpreted as being far from innocuous. On occasion, Horwich seems to treat propositions as the contents determined by a token’s meaning together
with its context of utterance. Consider the context of the passage quoted above, where he distinguishes the meanings of sentence types and tokens from propositions:

[Speakers] of English know how to use “I am hungry”—they know its unique meaning. However, on different occasions it expresses different propositions: I used it yesterday to say that I was hungry; whereas John will use it tomorrow to say that he is (then) hungry...

Certainly, it is vital to distinguish between (1) the meaning (or meanings) of a sentence type, (2) the meaning of one of its tokens, and (3) the proposition expressed by that token. (Horwich 1998a: 82)

Horwich is invoking Kaplan’s distinction between the character of a sentence token—the unvarying meaning it shares with other tokens of the same sentence type and which is determined by linguistic conventions—and the content of a sentence token—what is said by the token, determined by its character and its context of utterance—. This passage suggests that Horwich may be identifying the meaning of the token with its character and the proposition expressed by the token—what is said—with its content. Does he take the meaning of a sentence token to be merely its character? The quoted passage continues as follows:

Moreover, we must acknowledge that the use theory applies primarily to the first of these notions, and derivatively to the second (in so far as the meaning of a sentence token is simply one of the meanings of the ambiguous sentence type). As for the third notion of meaning—the proposition expressed—the use theory does not purport to give a complete account of it. However... the meaning of an
utterance token is a major determinant of which proposition is expressed...

(Horwich 1998a: 82)

If the meaning of a sentence token is the meaning of the sentence type to which it belongs, as Horwich suggests here, then it certainly is merely the token’s character when the sentence is context-sensitive. Of the three notions he distinguishes between, only the third one—the proposition expressed—seems able to accommodate Kaplan’s contents. It would be helpful to know what Horwich thinks about the relation between the meaning of a belief—which he expects his use theory to explain—and the proposition believed, since there is no character/content distinction in this case. But Horwich’s remarks on the relation between meanings and propositions are sparse and, as far as I can tell, he does not address the issue. However, his endorsement of the face value interpretation of attitude reports suggests that he identifies the content of a belief with a proposition. What attitude reports ascribe are contents and Horwich characterizes them as abstract propositional objects: “the things that we believe and that our utterances express—so-called propositions” (Horwich 1998b: 129). Further support for this interpretation comes from what Horwich says about the relation between propositions and the meanings of non-context-sensitive sentences: “[If we] restrict our attention to context-insensitive sentences (such as ‘dogs bark’ and ‘snow is white’)... it is plausible to identify the proposition expressed by an utterance with the meaning of the sentence-type to which the utterance belongs...” (Horwich 1998b: 131). All of this textual evidence seems to support the interpretation that Horwich does identify propositions and contents. Nevertheless, it has to be kept in mind that this is a topic that Horwich has not discussed in much detail. So I am not fully confident that this interpretation fully captures his view on the matter. But let me explain why Horwich’s account
would be problematic if this interpretation happened to be correct.

Once we combine Horwich’s characterization of propositions as abstract objects with his seeming identification of propositions with contents, the result is deeply troubling, rather than “innocuous”. The problem is that contents, not just characters, play a crucial role in the explanation of behavior. But, given the causal closure of the physical world—which Horwich accepts (Horwich 2010: 103n)—contents qua abstract objects cannot play such roles. Consider the following situation: Mary utters ‘I am hungry’ and Nancy responds ‘There is a sandwich in the fridge’; then Mary replies ‘Great!’, goes to the kitchen, opens the fridge, grabs the sandwich and proceeds to eat it. The explanation of Mary’s doing what she did relies on ascribing meanings to the utterances on this exchange, as well as ascribing beliefs and desires to Mary and Nancy. We ascribe to Mary’s first utterance the content that she is hungry: a content that we ascribe because we know the linguist convention for the use or the word ‘I’—the character of the word—and because we know facts about the context of utterance—that Mary is the speaker. Based on that ascription, we ascribe a mental state to Mary with that same content. The point is that the content we ascribe to the sentence token and, consequently, to a mental state of Mary, does play a causal role in the explanation of why she opens the fridge and eats the sandwich. It follows that contents, as well as characters, must be natural worldly phenomena that can play such roles. So I conclude that if Horwich identifies contents with propositions, his characterization of them as abstract objects does conflict with the causal closure of the physical world and is unable account for the role of contents in the explanation of behavior.

While Horwich identifies meanings as those properties that explain what causes the acceptance of sentences, Devitt argues that Horwich should extend his identification of meanings
to include their roles in the explanation of behavior: what sentences cause (Devitt 2011: 197-198 & 2002: 112). Horwich does provide an account of how mental sentences can cause behavior (Horwich 2005: 37-39), but he continues to identify meanings merely with what causes the acceptance of sentences (Devitt 2011: 198). The broader identification of meanings is not only well motivated, but also helps not to overlook the fact the causal closure of the physical world constrains what properties can constitute both of the kinds of meaning distinguished by Kaplan, since focusing on explanations of behavior shows that the contents we attribute ought to play causal roles. The fact that folk psychology is generally successful is evidence that the contents we do ascribe to tokens—not just the characters—often are the ones that we ought to ascribe. We need a naturalized semantics that explains the nature of characters and contents identified by their causal roles.

I have argued that Horwich’s combination of an account of propositions qua abstract objects with a naturalized semantics may either succeed or fail, depending on what specific relation between propositions and meanings is intended. In so far as Horwich sharply distinguishes meanings from propositions, does not regard propositions as determining meanings and offers a deflationary account of propositional truth, his account faces no problems. However, if we interpret Horwich as identifying meanings merely with the characters of tokens, the account is inadequate. The main problem stems from Horwich’s seeming identification of propositions with Kaplan-style contents which, I have argued, must also be naturalized and should not be regarded as abstract objects. While I see no motivation to posit propositions qua abstract objects, there is one way Horwich may avoid the problem. Horwich may regard both the characters and contents of sentence and belief tokens as natural properties which play causal
roles in the physical world and posit abstract propositions in a truly “innocuous” way, introducing them as follows: when a sentence or belief token has the same content (in Kaplan’s sense) as ‘p’ (a non-context sensitive sentence), it expresses the proposition that p. This may actually be Horwich’s more recent take on the matter, although he does not specify which aspect of meaning he intends the “innocuous” introduction of propositions to rely on (see Horwich 2010: 33). Yet, given the availability of sententialist versions of deflationism, I see no benefit in adopting such a refinement. Finally, Horwich’s main motivation for positing propositions qua abstract objects stems from ordinary language considerations. But if the face value interpretation he endorses were right (I have argued it is not), the contents we do ascribe would be abstract objects and could not play the causal roles they ought to play. In such a case, we should move on from mere ordinary language considerations to considerations about what properties we ought to ascribe to explain behavior in a scientifically viable way.

1.8 Conclusion

In this chapter, I have discussed the view that the meanings or contents of utterances and thoughts are the abstract objects called “propositions” which constitute the primary truth-bearers: what Church calls a “proposition in the abstract sense”. The main reason for addressing such a view is that it is inconsistent with a naturalized semantics of the sort I am defending and it is a view held by many philosophers. I have argued that the argument for propositions qua abstract objects based on the “One over Many” argument fails, since the Platonic “solution” faces an insurmountable difficulty: it faces a vicious relation regress. I have argued that the arguments for propositions qua abstract objects based on the alleged logical form of particular and general
content attributions also fail, since there are more plausible accounts of such attributions, which treat ascriptions as specifying semantic properties rather than referring to—or quantifying over—propositional objects. These non-propositionalist accounts, I showed, can also account for the validity of inferences containing particular and general content ascriptions. But I also pointed out that, even if the propositionalist accounts were correct about the ontological commitments of ordinary ascriptions, this would not prove that propositions exist, since folkloric opinion is *not* the ultimate guide to reality. The semantic properties we *ought to ascribe*—and which very likely are the ones we often ascribe, given the general success of folk-psychology—are properties that play causal roles in the explanation of behavior. This excludes Platonic objects as prospective meanings or contents.

We must either reject propositions altogether or look for a naturalistic account of propositions. I have, however, casted some doubt on the need for naturalized propositions. But this issue depends, I argued, on whether being true is a second-order property of first-order meaning properties. Finally, I have objected to Horwich’s combination of a naturalized semantics with an account of propositions as abstract objects—or at least one interpretation of it. The “propositional” contents he claims we ascribe in ordinary attributions must play causal roles in the physical world, so they cannot be abstract objects.

I conclude that, despite the popularity of propositions *qua* abstract objects (even among some philosophers who want to naturalize meanings!), there are no compelling reasons to believe they exist, but there are compelling metaphysical reasons to reject any abstract entities from a scientifically viable explanation of the meanings or contents of utterances and thoughts. We need a reductive explanation of what natural properties constitute the meanings of utterances and
thoughts: a naturalized semantics.
Chapter 2:

Truth, Reference and the Task of Semantics

2.1 Introduction

In the previous chapter I discussed a metaphysical issue related to the nature of meanings. In this chapter I will discuss a “methodological” issue. What is the task of semantics? The simple answer to this question is that the task is “to say what meanings are, to explain their natures” (Devitt 2012: 62). Presumably, the various semantic theories on offer are rival accounts of such natures. But it is often not clear whether they are genuine rivals, since they seem to have different purposes and disagree on what counts as a “meaning”. It is common to assume that the goal is to explain the truth-conditions of sentences and thoughts and the references of their parts —where truth is regarded as some form of correspondence with reality. But deflationists argue that truth and reference cannot play a substantial role in semantics. I will argue that no account of the nature of truth and reference should be assumed in the definition of the task of semantics, since it is likely that the dispute between correspondence and deflationary views will be settled by whether truth and reference are required to play a substantial role in the explanation of meaning. Following Devitt (1996), I argue that the task of semantics should be to explain the nature of the meanings pre-theoretically identified by the causal roles of representations. This definition of the task has the virtue of not assuming any account of the nature of truth or reference and, consequently, of not begging the question against deflationism.
2.2 The Nature of Truth and Reference

The status quo in semantics is truth-referentialism: the view that truth and reference play a central role in the explanation of meanings. The meanings of sentences are regarded as being exhausted by—or consisting mainly of—their truth-conditions, which in turn are regarded as being determined by the referential properties of words and the syntactic structure of sentences. Similarly, the truth-conditions of thoughts are regarded as being determined by the referential properties of the concepts they contain and their structure. Truth-referentialism requires truth to be a substantial property. The main substantial account of the nature of truth adopted by truth-referentialists is the correspondence theory. And the main alternative to the correspondence theory is deflationism, according to which truth is not a substantial property and cannot play an explanatory role in the theory of meaning. The deflationary conception of truth motivates some kind of “use” theory of meaning in which truth plays no explanatory role. I will briefly discuss correspondence and deflationary conceptions and argue that the dispute is likely to be settled by whether truth and reference must play an explanatory role in semantics.

The most venerable and popular view about the nature of truth is that it consists in some form of correspondence with reality. Various formulations of the view have been proposed. What they have in common is a commitment to the claim that truth is a specific kind of relational property: a property that some truth-bearers have because of a relation they bear to reality. The various formulations of the view agree on the broad characterization of this relation as one of correspondence, agreement or conformity. But they disagree on whether it involves a resemblance or merely a correlation between truth-bearers and reality (Pitcher 1964: 9-11; Kirkham 1992: Ch. 4.1). The various formulations also agree on the objective—mind-
language-independent—character of the reality that true truth-bearers are related to. But they disagree on whether true truth-bearers correspond to things or to facts (Künne 2003: Ch. 3).

2.2.1 The Canonical Correspondence Theory

What today is regarded as the canonical version of the correspondence view is the claim that a truth-bearer is true if and only if it corresponds to a fact. This idea was advanced in the early twentieth century by G. E. Moore, Bertrand Russell and Ludwig Wittgenstein. The canonical view requires an ontology of facts. Moore, Russell and Wittgenstein argued that facts are entities in their own right, existing over and above the things that compose them. The basic idea is that facts are structured combinations of objects with properties or relations. For example, the fact that a certain tree is an oak allegedly consists of an object, the tree, combined with the property of being an oak (Moore 1953: 311), while the fact that Desdemona loves Cassio allegedly is a “complex unity” of two objects, Desdemona and Cassio, connected by the loving relation which “binds [them] together” in a particular “order” (Russell 1912: 128). The canonical view also needs an account of the correspondence relation. Wittgenstein (1922) and Russell (1918) proposed that a true sentence “pictures” or “mirrors” a fact in virtue of (1) the referential relations between its words and the constituents of the fact, and (2) a relation of structural resemblance or isomorphism between the sentence and the fact (Wittgenstein 1922: 2.131 & 3.21; Russell 1918: 24-25). For example, ‘Socrates is wise’ is true if there is a fact consisting of the object and property referred to by the name and predicate, respectively, which has the same monadic structure, $\Phi x$, as the sentence. While ‘Napoleon hates Wellington’ is true if there is a fact that has the same dyadic form as the sentence, $\Phi xy$, and which has Napoleon and Wellington
and the hating relation as constituents.

While very influential in its time, the canonical version of correspondence faces serious difficulties. As J. L. Austin warned, the idea that portions of reality are isomorphic to the true sentences that represent them seems to commit “the error of reading back into the world the features of language” (Austin 1950: 119). Austin proposed that only correlations (not isomorphisms) between sentences and portions of reality are required for correspondence (Austin 1950: 115-120). Moreover, it turns out that sentences are not isomorphic to what they represent, as Russell himself noticed:

In the phrase “A is to the left of B,” even if we treat “is-to-the-left-of” as one word, we have a fact consisting of three terms with a triadic relation, not two terms with a dyadic relation. The linguistic symbol for a relation is not itself a relation, but a term as solid as the other words of the sentence. (Russell 1919: 38)

The fact represented by a true sentence ‘Rab’ allegedly is a unity of two objects, a and b, bound by a dyadic relation $R$. But the sentence is instead a unity of three words, ‘R’, ‘a’ and ‘b’, bound by a triadic logico-syntactic relation. If there is such a thing as the fact that $Rab$, it is not isomorphic to the sentence ‘Rab’.

Another problem for the canonical view is with the ontology of facts. To begin with, there seem to be no facts corresponding to negative, disjunctive or conditional statements (Strawson 1950: 154-155). Wittgenstein (1922) and Russell (1918) had a solution for this problem, which Russell called “logical atomism”: instead of explaining the truth of molecular sentences in terms of isomorphisms with logically complex facts, they explained it in terms of (1) their logical
structure and (2) the truth-values of the atomic sentences they contain: \( p \) and \( q \) is true if and only if \( p \) is true and \( q \) is true, \( p \) or \( q \) is true if and only if \( p \) is true or \( q \) is true, ‘If \( p \), then \( q \)’ is true if and only if ‘\( p \)’ is false or ‘\( q \)’ is true, etc. (Wittgenstein 1922: 5-5.101; Russell 1918: 37-41). However, this still relied on the mistaken view that atomic sentences are true in virtue of isomorphisms. Furthermore, even “atomic” facts are controversial. The idea that combinations of objects with properties or relations are sui generis entities assumes realism about universals—a controversial doctrine that I questioned in Chapter 1.\(^{31}\) We can accept that Socrates really is wise without regarding this as a combination of Socrates with a universal wisdom that can be instantiated elsewhere. Certainly there is plenty of talk of facts in ordinary language. But it looks like such talk is equivalent to talk of truths. P. F. Strawson, for example, suggests that the expressions ‘fact’ and ‘true statement’ have overlapping roles and points out:

There is no nuance, except of style, between ‘That’s true’ and ‘That’s a fact’; nor between ‘Is it true that...?’ and ‘Is it a fact that...?’ (Strawson 1950: 136)

Similarly, Quine remarks:

In ordinary usage ‘fact’ often occurs where we could without loss say ‘true sentence’ or (if it is our way) ‘true proposition’... Now so far as these uses go there is no call to posit facts... nor any difficulty in... paraphrasing away the word.

(Quine 1960: 246-247)

Strawson and Quine do not object to the ordinary uses of the word ‘fact’. Their point is rather that such uses are not seriously committed to the existence of anything other than truths.

\(^{31}\) Armstrong (1997) provides a contemporary defense of the existence of atomic facts, regarded as entities that are more than the sum of their parts.
2.2.2 The Contemporary Correspondence Theory

Given the serious problems faced by the canonical version of the correspondence theory, contemporary theorists have articulated a version of the view that relies neither on facts nor on resemblance relations. Instead, it relies on (1) an account of truth in terms of the logical structures of sentences and the referential properties of their parts—based on Alfred Tarski’s (1956 & 1944) influential work—coupled with (2) a reductive account of reference in terms of one or another kind of causal relation between words and objects—a combination which, Hartry Field (1972) argued, is necessary to provide a robust account of truth. While the canonical view regards correspondence as a resemblance-to-facts relation, the contemporary view treats correspondence as a correlation-to-objects relation.

Tarski (1956 & 1944) wanted to offer an account of truth that does justice to the classical conception of truth as consisting in some form of correspondence to reality (Tarski 1956: 153 & 1944: 342-343). Despite his intentions, however, his account turns out to be a form of deflationism (Horwich 2010: 22-24; Devitt 2010:170-173). Tarski showed how the truth of a sentence can be explained in terms of its structure and the referential properties of its parts. This is the aspect of Tarski’s work that is adopted by the contemporary correspondence theory. I will ignore for now Tarski’s account of reference, which is the deflationary part of his account discarded by the contemporary correspondence theory.²²

Tarski offered his account for fully formalized languages. I will use predicate symbolic logic as an illustration. Consider a simple sentence ‘Kr’. Following the procedure proposed by Tarski, we can specify the truth-conditions of ‘Kr’ as follows:

²² I will also ignore that Tarski (1956) did not offer an account of truth, but of truth-in-a-(specific)-language.
‘Kr’ is true $\leftrightarrow \exists x (\text{‘}r\text{’ designates } x \text{ & ‘}K\text{’ applies to } x)$. 

In plain English: ‘Kr’ is true if and only if there is an object that the name ‘r’ designates and the predicate ‘K’ applies to. Notice that the truth-conditions of the sentence are characterized in terms of its structure and the referential relations—application and designation—between its component words and an object in the world. Sentential structure and referential relations were also components of the canonical view. But here “facts” and “structural resemblances” *play no role*. These are correlation-to-objects relations instead of resemblance-to-facts relations. Given the difficulties faced by the canonical view discussed above, related to “facts” and “structural resemblance”, this is a clear advantage.

Tarski showed that the truth-conditions of an *infinite* number of sentences can be derived from a *finite* set of rules or axioms that can be *recursively* applied. For example, the truth-conditions of ‘Kr’ stated above can be derived from applying the general rule:

\[ \forall \varphi \neg \varphi \neg \varphi \text{ is true } \leftrightarrow \exists x (\text{‘}a\text{’ designates } x \text{ & ‘}F\text{’ applies to } x). \]

Where ‘a’ can be replaced by any name and ‘F’ can be replaced by any one-place predicate. So the truth-conditions of *any* sentence ‘Fx’ can be obtained. Similarly, the truth-conditions of a compound or molecular sentence ‘Kr & Jz’ can be specified by the biconditional:

‘Kr & Jz’ is true $\leftrightarrow \exists x (\text{‘}r\text{’ designates } x \text{ & ‘}K\text{’ applies to } x) \& \exists x (\text{‘}z\text{’ designates } x \text{ & ‘}J\text{’ applies to } x). \]

Which is derived from applying the truth-rule for ‘Fx’ to each conjunct *after* applying the general truth-rule for conjunction:
\[ \overline{A} \land \overline{B} \text{ is true } \leftrightarrow (\overline{A \land B} \text{ is true}). \]

Where \( \overline{A} \) and \( \overline{B} \) can be replaced by any simple or compound sentence (so the rule can be applied recursively to complex sentences that contain indefinitely many simple sentences). Notice that this characterization coincides with the one offered by Wittgenstein (1922) and Russell (1918). What sets apart Tarski is his account of “atomic” sentences. His account of “molecular” sentences is the same.

One of Tarski’s crucial insights is that the “semantic” rules should track and exploit the recursive syntactic rules. And the “syntactic” formation rules of predicate logic are finite. So, once we have formulated the corresponding finite set of semantic truth-rules that cover all the formation rules, we can characterize the truth-conditions of an infinite number of sentences.

What we need is truth-rules for: (1) sentences containing \( n \)-place predicates—\( \overline{F{a_1}{a_2}} \), \( \overline{F{a_b}} \), \( \overline{F{a_c}} \), etc.—, (2) sentences with quantifiers—\( \overline{\forall x(Fx)} \) and \( \overline{\exists x(Fx)} \)—and (3) sentences with logical operators or connectives—\( \overline{\neg{A}} \), \( \overline{A \land B} \), \( \overline{A \lor B} \), \( \overline{A \rightarrow B} \) and \( \overline{A \leftrightarrow B} \). Tarski showed how this can be done—roughly along the lines of the examples I gave above.

What is required to turn a recursive account along these lines into a correspondence theory of truth is to (1) show how it can be extended to natural languages and (2) complement it with an account of the nature of the referential relations between words and objects—a reductive account that does not rely on semantic terms, since semantic properties surely are not among the fundamental properties of the world. Tarski himself offered an account of reference in non-semantic terms. But Hartry Field (1972) showed that it did not provide a proper reduction (I discuss this below) and that it had to be replaced with a naturalistic account of reference in terms of one or another kind of casual relation between words and the world. There has been plenty of
work on how to naturalize reference in the last decades. The main contenders are theories that rely on the historical causes, the reliable causes and the etiological functions of representations, respectively. None of these proposals are without problems. Whether some reductive explanation of reference in non-semantic terms succeeds or not remains to be seen. Deflationists are skeptical because they think that a reductive explanation of reference is neither possible nor needed (Horwich 1998b; Field 1994). But the matter is not settled, and we should keep searching.

The other challenge for the contemporary theory is to extend the recursive account to natural languages. Tarski himself was skeptical, because formulating the truth-rules for a language requires that its structure has been already specified exactly. But times changed and eventually philosophers and logicians—and later linguists—started to explore how a Tarskian account could be applied to natural languages. A crucial inspiration for this optimism was the increased understanding of the syntax of natural languages provided by Chomskyan generative grammar. Richard Montague declared: “I reject the contention that an important theoretical difference exists between formal and natural languages” and proposed a Tarski-inspired theory of truth for “a formal language that I believe may be reasonably regarded as a fragment of ordinary English” (Montague 1974: 188). Similarly, Donald Davidson proclaimed: “a semantic theory of a natural language cannot be considered adequate unless it provides an account of the concept of truth for that language along the general lines proposed by Tarski for formalized languages” (Davidson 1984: 55). Consider the philosophical practice of translating natural language sentences into quantified predicate logic to reveal their logical form. Plenty of sentences can be

33 As he pointed out: “At the present time the only languages with a specified structure are the formalized languages of various systems of deductive logic, possibly enriched by the introduction of certain non-logical terms.” (Tarski 1944: 347). Tarski was also concerned about the syntactic rules of natural languages allowing for the formation of paradoxical sentences like the self-referring ‘This sentence is false’, which would make the characterizations of truth-conditions inconsistent (Tarski 1956: 157-165 & 1944: 347-349).
successfully translated. As Davidson points out, this reveals that formal languages already cover at least some the structure of natural languages—for example, Frege’s work on quantifiers reveals the logical roles of some uses of words like ‘all’, ‘every’, ‘some’ and ‘none’ in English (Davidson 1984: 29). The remaining challenge is to show that a Tarski-based approach can be extended to those sentences of natural languages that seem to resist formalization. Davidson’s analysis of action sentences and adverbial modifications is a good example of an attempt to reveal the logical form of some problematic sentences, so that a Tarskian account can cover them (Davidson 1967 & 1985). Just like with reductive theories of reference, it remains to be seen whether this can be accomplished.

2.2.3 Deflationism

According to deflationism, a reductive explanation of truth and reference is neither possible nor necessary, because truth and reference are not substantial properties with an underlying nature that could be reductively explained. I will first present the unintentional deflationism of Tarski and then the views of some contemporary deflationists.

Tarski suggested that a proper account or “definition” of truth should capture our pre-theoretical intuitions about what is for a sentence to be true by entailing the proper instances of the equivalence schema:

\[(T) \text{ ‘} S \text{’ is true } \iff S.\]

Where the sentence ‘S’ is mentioned on the left hand of the biconditional and used on the right hand. The well-known example Tarski (1944) gave of an instance of this schema in English is:
‘Snow is white’ is true if and only if snow is white.

Tarski regarded each instance as a “partial definition” of truth, because he thought that it accounted for the truth of a particular sentence. Indeed, Tarski suggested that a mere list of all the instances of the schema for a language with a finite number of sentences would provide a proper account of truth in that language (Tarski 1956: 188). He moved on to offer his recursive account because interesting languages have an infinite number of sentences. But notice that a list-like account would not constitute a “correspondence” theory. As Douglas Patterson points out, many philosophers—notably Tarski (1944: 342-343)—have held the view that instances of the schema “state correspondences between the sentences they mention and something else” (Patterson 2003: 421). But instances of the schema do not state relations of correspondence, since they do not state any relations whatsoever: their basic form is \( A \leftrightarrow B \), rather than \( Rab \). Indeed, deflationists argue that there is no need to regard the word ‘truth’ as standing for a relational property because instances of the equivalence schema are all that is required to account for its meaning and utility (Horwich 1998b; Field 1994).

Tarski complemented his recursive account of truth with a non-semantic characterization of what is for a name to designate an object and what is for a predicate to apply to an object. Notice that reductive explanations of reference in terms of one or another kind of causal link between symbols and reality (whether reliable, historical or teleological) also are non-semantic. But the “non-semantic” characterizations of reference offered by Tarski are not of this sort. In the case of ‘\( K \)’ and ‘\( r \)’, for example, they would be the following:

\[
\forall x (\text{‘}K\text{’ applies to } x \leftrightarrow Kx).
\]
Tarski proposed that a complete list of biconditionals like these, covering all the non-logical symbols of a language, would provide a proper non-semantic characterization of what is for a symbol to refer in that language. Sure enough, the right-hand sides of these biconditionals contain no semantic terms. But, as Field pointed out, they do not provide reductive explanations of the nature of reference (Field 1972). Field made this point vivid using an analogy with the reductive explanation of chemical valence. In the twentieth century, the valence of elements was explained in terms of the underlying physical properties of their atoms. But suppose that a chemist in the late nineteenth century offered the following as a “non-chemical” characterization of the valence of elements:

An element \( e \) has valence \( n \) if and only if either \( e \) is arsenic and \( n \) is 5, or \( e \) is beryllium and \( n \) is 2, or …, or \( e \) is zinc and \( n \) is 2. (See Field 1972: 362-363)

Such a list-like characterization can match all the elements with their correct numerical values. The right-hand side of the biconditional contains no chemical terms. Yet it is clear that this list-like characterization does not provide a reductive explanation of valence (ibid). Indeed, it says nothing substantial about the nature of valence. Similarly, Tarski’s list-like definitions say nothing substantial about the nature of reference. While Field saw this as a serious drawback—

34 Tarski (1956) initially focused on the relation between predicates and the objects they apply to, which he called “satisfaction”. He offered the following example of how the application of a one-place predicate can be defined: “for every \( a \), \( a \) satisfies the sentential function ‘\( x \) is white’ if and only if \( a \) is white” (Tarski 1956: 190). The formal languages he originally analyzed did not contain names (Horwich 1998b: 110n). But he did provide an account of designation (Tarski 1956: 194n; see Field 2001: 17-18), which he later explicitly included among the semantic features that determine sentential truth (Tarski 1944: 345). Tarski defined designation—just like predicate-application—in terms of satisfaction: “a given term designates a given object if this object satisfies the sentential function ‘\( x \) is identical with \( T \)’ where ‘\( T \)’ stands for the given term” (Tarski 1944: 373n). So he regarded satisfaction as the most fundamental semantic relation that determines sentential truth: “a sentence is true if it is satisfied by all objects, and false otherwise” (Tarski 1944: 353).
and proposed that we replace Tarski’s definitions with a causal theory of reference—he eventually changed his mind and came to think that truth and reference indeed have no underlying nature, but the words ‘true’ and ‘refers’ nonetheless play useful roles. I will briefly characterize the views of Horwich and Field, that differ mainly on their takes on truth-bearers.

Deflationists are not always very clear about the metaphysics of truth and reference, although they have plenty to say about the roles of the words ‘true’ and ‘refers’ (‘applies to’, ‘designates’, etc.). Consider the following schemata:

‘p’ is true if and only if \( p \).

‘a’ designates and object \( o \) if and only if \( a = o \).

‘F’ applies to an object \( o \) if and only if \( o \) is \( F \).

Deflationists like Field and Horwich argue that the roles of ‘true’, ‘designates’ and ‘applies’ in the instances of these schemata are all that is required to account for their meaning and utility, so there is no need to regard ‘truth’ and ‘reference’ as standing for relational properties. While Field (1994) is a sententialist, Horwich (1998b) is a propositionalist, so his version of the first schema is instead: \(<p>\) is true if and only if \( p \).

But what is the utility of these words? Deflationists say that they work as devices of generalization that allow us to say things that we would not be able (or it would be hard) to say without them. You can say, for example, ‘What he said is true’, expressing agreement with a statement that you may not be able to put in words. Given the equivalence schema, this sentence entails: ‘If he said that grass is green, then grass is green, if he said that Mars is a planet, then Mars is a planet, if he said that \( 2 + 2 = 4 \), then \( 2 + 2 = 4 \), etc.’. Conversely, consider this list of
sentences: ‘If grass is green, then we should believe that grass is green; if Mars is a planet, then we should believe that Mars is a planet; if \(2 + 2 = 4\), then we should believe that \(2 + 2 = 4\); etc.’. By deploying the equivalence schema, we can turn the list into the following one: ‘If ‘grass is green’ is true, then we should believe that grass is green; if ‘Mars is a planet’ is true, then we should believe that Mars is a planet; etcetera’. In turn, this can be generalized into: ‘We should believe the truth’ (Horwich 1998b: 105-106). This generalization would not be possible without the truth predicate. In virtue of the acceptance of the equivalence schema, we can formulate generalizations like this one. Moreover, deflationists argue that the equivalence schema is all that is required to account for other uses of the word ‘true’. In Horwich’s words:

The deflationist’s contention… is that whenever we deploy the concept of truth… it is playing this role: a device of generalization. Moreover, it doing so requires… no more and no less than the equivalence schema. Thus the basis for our use of the truth predicate is indeed our acceptance of the instances of that schema…

(Horwich 1998b: 106)

Horwich indeed claims that the role and meaning of ‘true’ are explained by our acceptance of the schema. And similar considerations are offered for ‘designates’ and ‘applies to’ (or ‘is true of’).

But what about the metaphysics of truth? What do deflationists say about the nature of truth itself, rather than about the role of ‘true’? This is a more difficult question than it may seem, since discussions of deflationism “tend to blur the distinction between the linguistic and the metaphysical” (Devitt 2010: 157). What is clear is that, based on their account of the role of

---

35 Horwich uses the propositional version of the equivalence schema in his own examples. Given my objections to propositions presented in Chapter 1, I prefer Field’s sentential version.
the truth predicate, they argue for the negative thesis that it does not stand for a substantial relational property. Horwich is happy to describe truth as a “logical” property—although he insists that is has no underlying nature. Presumably he should say that there is no truth property—or no such thing as truth. So there is no such thing as the nature of truth. Anyway, if the truth predicate does not stand for a relational property, then it does seem that there is no place for a correspondence theory of truth.

How can we figure out whether we need a correspondence theory or deflationists are right and we do not need such a theory? On one hand, the correspondence view of truth is very appealing and seems to provide the proper framework for a truth-referential semantics. On the other hand, the deflationary view of the role of ‘true’ seems very plausible. But if the equivalence schema explains everything about the role of ‘true’, then the term does not stand for a relational property of the sort required to play an explanatory role in semantics—although the truth term may serve in semantics as a device of generalization. The main argument in favor of the correspondence theory and against deflationism is that truth and reference play substantive roles in the explanation of meaning (Devitt 2010: 178-181). Deflationists have attempted to address this concern by offering so-called “use” theories of meaning where truth and reference do not play any explanatory roles (e.g., Brandom 1994 & Horwich 1998a). But it remains to be seen whether these theories can succeed explaining meaning and, moreover, if in case of succeeding they do not appeal to substantive relations between language and the world that may turn out to constitute truth and reference. Given this situation, it looks like the debate on the nature of truth and reference is very likely to be eventually settled after we have figured out what is the best explanation of meaning and whether truth and reference play substantive roles in it or not.
Consequently, we cannot settle the debate on the nature of truth and reference before engaging in the search for a proper theory of meaning—and we cannot assume that the correspondence theory or deflationism is correct before learning more about the nature of meaning.

2.3 Identifying and Explaining Meanings

The task of explaining the nature of any property should rely on a relatively uncontroversial identification of the property to be explained. But semantic theories often proceed without such an identification or rely on controversial identifications which are simply taken for granted but assume a particular account of the nature of meanings, such as the assumption that a theory of meaning must explain the truth-conditions of truth-bearers and the referential properties of their parts. I shall argue that the task of semantics must be specified on the basis of a non-question-begging identification of meanings.

As Devitt points out, the task of investigating the nature of meanings is analogous to the tasks of investigating the natures of genes (Devitt 2012: 62). Likewise, Horwich suggests that the scientific investigation of what constitutes water provides a model for what a semantic theory should accomplish: it should account for what properties constitute meanings (Horwich 2005: 76-77). The suggestion of Devitt and Horwich is that the task of semantics is to find, for meanings, an account of the sort exemplified by the discoveries that genes are DNA sequences and water is $H_2O$. The natures of genes and water—like the nature of meanings—are constituted by more fundamental properties. But notice that the tasks of explaining the nature of genes and

---

36 And the properties that constitute meanings, genes and water in turn are constituted by even more fundamental properties. The *chemical* properties of DNA and $H_2O$ are explained by their underlying *physical* properties. Similarly, whatever properties constitute meanings surely are not directly fundamental physical properties, but some properties whose nature needs further explanation.
water—unlike the task of explaining the nature of meanings—had a clearly defined “subject matter”, since genes and water were “identified relatively uncontroversially” before undertaking the investigation of their nature (Devitt 2012: 62). Furthermore, the identification of genes and water served “clear and familiar theoretical or practical purposes” (Devitt 2012: 62). The characteristic features by which genes and water were identified—and which made their identification useful for our theoretical or practical purposes—needed explanation: why do genes and water have these features? Consequently, investigating their underlying nature was clearly a worthwhile task: it was clear (1) what to explain and (2) why the explanation was needed. In sum, there was a clear identification of what needed explanation: “[The] first step is semantics is to identify meanings. We can then move on to explain the meanings so identified.” (Devitt 2002: 108).

What are these “meanings” we are trying to explain? How should we identify them? And why is it worth seeking such explanations? What are the explanations needed for? The methodological question of how meanings should be identified before attempting to explain their nature is often overlooked (Devitt 2012: 61-63). But this situation is deeply troubling. On one hand, “our ordinary talk of “meanings”... is very vague” (Devitt 2011:197). So the claim that the task of semantics is to explain the nature of “meanings” requires further specification. On the other hand, theorists often simply take for granted certain views of the tasks of semantics, without properly motivating them: “It is troubling that much semantic theorizing proceeds with inexplicit reliance on apparently ad hoc views of the semantic tasks.” (Devitt 2012: 62). The problem is that competing theories do not only disagree on what is the best explanation of the
nature of meanings, but also on what needs explanation: “It is far from clear what counts as a meaning that needs explaining. Indeed, the intractable nature of semantic disputes largely stems from differing opinions about what counts.” (Devitt 2012: 62).

Many theorists assume that the task of semantics is to explain what underlying properties determine the truth-referential properties of symbols. In other words, they take for granted an identification of meanings with truth-referential properties at the semantic level and seek a reductive explanation of such properties. But this simply begs the question against deflationary views according to which truth and reference have no underlying natures. While properties that determine truth and reference may indeed play a role in the explanation of meaning, deflationism should not be “ruled out by a definition” of the semantic task (Devitt 1996: 63-64). There is a panoply of theoretical options and none of them should be either ruled out or favored by a definition of the semantic task. On one extreme, there are theories according to which a reductive explanation of truth and reference is all that is needed to account of the underlying nature of meanings (e.g., Fodor 1987). On the other extreme, there are theories that attempt to reductively explain meanings merely in terms of functional-roles which do not substantially determine truth and reference (e.g., Harman 1987). Between these extremes, there are theories according to which a reductive explanation of truth and reference is crucial but does not exhaust an account of the nature of meanings. One alternative is to regard the very same non-semantic properties that determine truth and reference also as constituting Fregean “modes of presentation” (e.g., Peacocke 1992, Devitt 1996, 2001b & 2012).\footnote{Horwich points out that the distinction that Frege (1948) draws between sense and reference remains at the semantic level (Horwich 1998a: 52). But just like there are attempts to explain the underlying nature of reference, there are attempts to explain the underlying nature of Fregean senses or modes of presentation. Peacocke (1992) explains the meanings of concepts in terms of “primitively compelling” functional-roles that constitute modes of presentation substantially determining their reference. Similarly, Devitt (1996) naturalizes} Another alternative is to regard the underlying
nature of meanings as composed by two independent factors: a factor that determines truth and reference plus a functional-role factor that does not determine truth and reference (e.g., Block 1986). All of these proposals are presented as offering accounts of the nature of meanings: they are allegedly engaged in the same task, asking the same question and offering different competing answers. For this to be genuinely the case, the specification of the task of semantics should not beg the question against any of these proposals (Devitt 1996: 64).

How can we decide among all the competing accounts of the nature of meanings? And how can we determine that they are genuinely competing explanations of the same phenomena, rather than explanations of different aspects of linguistic and mental representations? Consider, for example, my thought that snow is white. It is uncontroversial that my thought has the meaning or content that snow is white, that it is true and that one of the concepts it contains refers to snow. But let us consider what underlying properties my thought and the concepts it contains have. As it turns out, a wide range of theorists agree on what these various properties are, although they disagree on which among them constitute meanings and on whether they substantially engender truth-referential properties. Most theorists agree that my thought and the concepts it contains have complex functional-roles in my mind: that they are causally connected to various other thoughts and concepts, to various inputs from my perception and outputs to my behavior. Yet, they disagree about whether all, some or none of these roles constitute the meanings of my thought and my concepts. Additionally, most theorists agree that one of the concepts in my thought has links to snow in the distal/external world and, furthermore, that such

the meanings of non-primitive words covered by a description theory in terms of some of their functional-roles: those that constitute the modes of presentation that determine their reference. Interestingly, Devitt (1996, 2001b & 2012) suggests that the causal links to reality—whether historical-causal, informational and/or teleological—that determine the reference of primitive words should also be regarded as constituting modes of presentation: these are causal modes rather than descriptive modes.
links are—at least part of—what constitutes the meaning of the concept. Yet, they disagree about whether such links substantially determine the reference of the concept and on whether they should be regarded as part of its (non-reference-determining) functional-role. Regarding truth and reference: even deflationists agree that representations do have truth-referential features, although they regard them as trivial (e.g., Horwich 1998b, Field 1994). Regarding modes of presentation: “direct reference” theorists agree that there are properties that substantially determine reference, although they do not include them among the components of meaning (e.g., Salmon 1986: 70-71, Soames 2002: 19-20; see Devitt 2001b & 2012 for a criticism of this view); while deflationists agree that the meanings of words and concepts determine their reference, although they do not regard such “determination” as substantial (Horwich 1998a: 68-71, 2005: 79-81). Finally, regarding functional-roles: every author who is a realist about mental representation—and this includes most authors, except a few behaviorists and eliminativists—agrees that thoughts and concepts have causal roles in cognition, although some authors deny that these roles constitute the meanings of any concepts (Fodor 1987: 71-95). While different theorists agree that representations and their parts have all these features—functional-roles, connections to reality, etc.—they disagree about which among them constitute their meanings.

Why should we regard meanings as truth-referential properties, including or not modes of presentation, as functional-roles or as some combination of these features? Lacking an explicit specification of the purposes of semantics, the dispute may seem to be merely about what should be honored with the label of “meaning”. William Lycan humorously expresses this concern with his “Double Indexical Theory of Meaning”:

$\text{MEANING} =_{\text{def}} \text{Whatever aspect of linguistic activity happens to interest me now.}$
While Lycan’s concern is mainly with theories of linguistic meaning, we can rephrase it to also include theories of mental content:

\[
\text{MEANING} =_{\text{def}} \text{Whatever aspect of linguistic or conceptual activity happens to interest me now.}
\]

To be fair, the competing approaches in semantics often do have theoretically motivated differences. The disputes (or at least some of them) seem to be substantive rather than merely verbal. For example, purely functional-role theorists are not simply “more interested” in functional-roles than on truth and reference. At least some of them are motivated by deflationary views according to which truth and reference are trivial features that cannot explain meaning or content (e.g., Field 1994). In contrast, truth-referential theorists are motivated by the view that truth and reference are substantial properties of linguistic and mental representations (e.g., Fodor 1987).

But granting that the debate is not merely verbal does not remove the underlying problem. As Devitt points out, “to avoid Lycan’s mockery, we must specify a subject matter that is worthy of investigation, we need an explication that is not ad hoc” (Devitt 1996: 55). We cannot determine what theory offers the best account of the nature of meaning without a clear specification of worthwhile explanatory tasks that are not ad hoc. What we need is an identification of the meanings that need explanation which does not presuppose any specific

---

38 Others are instead motivated by the idea that content must be “narrow” or “within the head” in order to explain behavior. This is the case of Ned Block (1986). Block does not adopt a deflationary view that would be incompatible with a truth-referential semantics. Instead, he combines a truth-referential semantics with a “narrow” version of functional-role semantics as parts of his two-factor theory of meaning.
account of their nature. The dispute among various semantic theories will remain hopeless as long as we do not have a non-question-begging identification of what needs to be explained and, consequently, a measure of what may count as a satisfactory explanation.

In what follows, I shall argue that the investigation of the nature of meaning-properties should follow the steps involved in the reductive explanation of any property, starting with an uncontroversial identification of the property and of the characteristic features by which we identify it before studying its underlying nature. In the next section, I argue that we should identify meanings as the properties that are responsible for the causes and the behavioral effects of linguistic and mental representations. Since we need an explanation of these causal roles, investigating the nature of the properties responsible for them is a worthy task.

2.3.1 How to Identify Meanings and Investigate their Nature?

In this section, I discuss the methodological issues of how to identify meanings and how such identification should inform the investigation of their nature. I shall argue, following Devitt, that we should first identify examples of properties that ordinary people recognize as semantic and establish what these properties are supposed to do and what are the purposes for which they are ascribed. This provides a preliminary target for the investigation of the nature of meanings: the investigation should examine the examples of meanings to find out which underlying properties constitute meanings and, consequently, are responsible for or explain what meanings do and how they serve the purposes for which they are ascribed. But it should be kept in mind that the preliminary identifications are in principle defeasible, and we may need to revise them in light of the investigation of the nature of meanings. I will proceed as follows: I shall first discuss how the
investigation of the nature of any property should proceed and then apply the findings to the methodology of semantics.

Consider the general methodological question of how do we determine the nature of any property: “How do we tell what is the nature of some property, being an \( F \), that we ascribe?” (Devitt 2012: 66). The answer to this question, as Devitt points out, “breaks into two stages”:

First, we must identify some apparently uncontroversial examples of \( F \)’s and non-\( F \)’s. Second, we must examine the examples to determine the nature of being an \( F \). The second stage is a straightforwardly scientific one. The preliminary first stage may not be. It involves using “identification experts” who may be scientists but may be just plain folk. (Devitt 2012: 66-67)

Examining the nature of a property is the second stage of the investigation. The first stage is to identify clear and apparently uncontroversial instances of the property. The successful scientific investigations of the nature of various properties clearly illustrate these two different stages and how they are related. They provide us with a good general model for how to proceed in semantics. I will focus on the two paradigmatic examples that I mentioned above: the investigation of the nature of water and genes.

Let us consider the task of explaining the nature of water. The chemical discovery that being a sample of water is constituted by or reduces to being made of \( \text{H}_2\text{O} \) molecules accomplished this task. In order to accomplish this task, chemists examined samples of water that were identified pre-theoretically and uncontroversially—without having an account of their underlying chemical composition. Samples of water were identified by characteristic features,
like being samples of the colorless liquid that quenches thirst, flows in rivers, freezes at certain cold temperatures (0 degrees Celsius) and boils at certain high temperatures (100 degrees Celsius at sea level). The task of explaining the nature of water, consequently, had a clear and worthy target: what is the nature of the identified samples of water, in virtue of which they have their characteristic features? Examining the samples, chemists discovered that they are made of \( \text{H}_2\text{O} \) molecules. Based on this discovery, chemists were able to explain why water has its characteristic features. For example, the physico-chemical behavior of \( \text{H}_2\text{O} \) molecules at different temperatures explains why water freezes and boils at the temperatures it does. It is important to distinguish water’s nature from its causal role. The nature of water is being made of \( \text{H}_2\text{O} \) molecules. The causal role of water includes, among other things, its behavior as a solid, as a liquid or as a gas at different temperatures. While water was identified by its causal role, chemists did not identify water with its causal role. Rather, they hypothesized that there must be some underlying chemical composition that constitutes the nature of water and that explains why water has its causal role or characteristic behavior.

Let us consider the task of explaining the nature of genes. In this case, the identification of genes relied entirely on a scientific theory. The Mendelian postulation of genes as entities playing causal roles in heredity—the transmission of innate traits from organisms to their offspring—was well-supported by empirical evidence. But there was an explanatory need to account for the nature of these theoretically identified entities. Again, we should not confuse what the entities are or their nature with what they do or their causal roles. Mendelian genetics discovered that there are some things that play certain causal roles in heredity, but it did not discover the nature of these things. What are genes made of and how can they play their causal
roles in heredity? Molecular genetics provided the needed explanation. This happened many
decades after the work of Gregor Mendel, as a result of the discovery by James Watson and
Francis Crick of the helix-shaped structure of DNA: Molecular genetics discovered that genes
are made of segments of DNA molecules which are responsible for the transmission of innate
traits from organisms to their offspring. Molecular genetics relied on Mendelian genetics to
identify what needed explanation (Devitt 1996: 73).

The way scientists discovered the natures of genes and water illustrate some basic
features of the methodology involved in the investigation of the nature of a non-fundamental
property—a property that has an underlying nature. Notice that the account of the underlying
nature of a property involves two different levels: the “higher” or “superficial” level where the
property to be explained is identified and a “lower” or “underlying” level where the nature of the
property is explained. Basically, the account of a property’s underlying nature is a reductive
explanation. There are philosophical controversies about the character of such explanations. Do
they ontologically reduce the higher-level phenomenon to the lower-level phenomenon? And
how is the gap between the levels bridged in the inference from the explanans to the
explanandum? These are important questions and they do concern the metaphysics and
methodology of theories of the nature of meaning. However, my current goal is to make some
general methodological suggestions leaving room for different answers to these deeper questions.
The task of investigating the nature of meaning can benefit from an elucidation of the
uncontroversial aspects of reductive explanations, since the debate among different reductive
accounts of meaning often proceeds without explicitly attending to even the most basic
methodological matters.

39 See Kim (2005: 93-120) for a detailed discussion of these matters.
My *first methodological suggestion* is that the investigation should involve the two stages distinguished by Devitt: the examination of the nature of the property should rely on a *prior* identification of *apparently uncontroversial* examples of the property. In the case of the property of being water, both scientists and ordinary people were able to identify samples of the property. No controversial scientific theory was assumed in such identifications. In the case of the property of being a gene, in contrast, Mendelian genetics was assumed. However, Mendelian genetics identified genes *without having any theory of their underlying nature*: this is the specific sense in which the prior identification of what needs to be explained should be uncontroversial.\textsuperscript{40} Since the explanation of the nature of the property will take place at a lower-level, the methodological suggestion is that we need an apparently uncontroversial identification of the property at the higher level *before* examining its nature at a lower level.

My *second methodological suggestion* is that an analysis of how examples of a property are identified should establish what the property is supposed to *do* and what are the *purposes* for which it is identified. Samples of water are pre-theoretically identified by their superficial characteristics, including their observable behavior: water *does* things such as quenching thirst when drunk, freezing when cooled and boiling when heated. Identifying water serves clear theoretical purposes, like predicting its behavior at different temperatures, and clear practical purposes, like quenching thirst by drinking it, making ice cubes by cooling it or brewing tea by heating it. Similarly, Mendelian genetics identifies examples of genes by what they *do*, by their characteristic causal roles in heredity: “the subject matter of investigation is already identified

\textsuperscript{40} As James Watson reminds us in his personal account of events that led him and Francis Crick to discover the “double helix” structure of DNA, the leading hypothesis until then was that genes were “special types of protein molecules” rather than strings of DNA (Watson 1968: 12). Both the DNA theory and the Protein theory were compatible with the Mendelian identification of genes.
relatively uncontroversially... because we already have a fairly good idea of what genes... do, which is a basis for identifying them” (Devitt 2002: 107). Identifying genes serves clear scientific purposes in explaining and predicting heredity. Diagnosing what a property does or is supposed to do and the purposes for which we identify it are intimately linked, since identifying a property presumably serves our purposes because of what the property does. Often, elucidating the purposes for which we ascribe a property can lead to an elucidation of what the property does, of its causal role. An elucidation of what a property does is not an explanation of its nature. But it should inform the investigation of such nature.

My third methodological suggestion is that the examination of the underlying nature of the identified higher-level property—what Devitt calls the “second stage” of the investigation—should have as a goal to reveal what lower-level property explains what the identified property does—its causal role—and how it serves the purposes for which it is identified. The account of the nature of water as composed of H$_2$O molecules explains the characteristic features by which water was identified. Similarly, the account of the nature of genes as composed of strings of DNA explains the causal roles by which they were identified. These accounts also explain how the higher-level identifications serve our purposes: how it comes about that identifying certain genes we can predict the transmission of certain innate traits or that identifying samples of water we can predict at what temperature they will freeze. That the reductive explanation of a higher-level property must explain what lower-level property is responsible for its causal role is a basic and uncontroversial requirement. Suppose that we discovered a lower-level property of samples of water that did not explain why those samples freeze at 0 degrees Celsius, or a lower-level property of samples of genes that did not contribute to an explanation of their causal role in
heredity. Surely we would have no reason to regard such lower-level properties as constituting the *natures* of water and genes. That a lower-level property explains the causal role of a higher-level property is crucial evidence in support of the claim that the former constitutes the latter.

It is time to apply these methodological suggestions to the investigation of the nature of meanings. The first step is to identify uncontroversial samples of properties identified as meanings. The second step is to find out what these identified meanings do—or are supposed to do—and what are the purposes for which they are identified. The last step is to look for what underlying non-semantic properties constitute meanings. A basic requirement to determine whether a non-semantic property indeed constitutes a meaning is that it must explain the causal role *by which* and *for which* the meaning is identified.

My methodological suggestions are far from original. They correspond to similar suggestions made by Devitt and Horwich. Devitt argues that we need to first identify uncontroversial samples of meanings and then investigate their nature (Devitt 1996: 72-81). But he also pays attention to the intermediate step I suggested, arguing that we should find out what are the purposes for which we ascribe meanings and that those purposes should inform our investigation of the nature of meanings (Devitt 1996: 57-62). Horwich shares this view. His following passage serves well as a summary of what I have said about how semantics should proceed:

An underlying property U constitutes a relatively superficial property S if and only if the co-extensiveness of U and S explains why S is manifested in the characteristic ways that it is. For example, we judge that ‘being made of H₂O
molecules’ constitutes ‘being a sample of water’ because, on the basis of the assumption that water is made of H2O, we can explain why water is a colorless, tasteless liquid that boils at 100° C. In the same way, in order to identify how meaning-properties are constituted, we should look for underlying non-semantic properties that can explain the symptoms of those meaning-properties. (Horwich 2005: 76-77)

It is crucial to identify correctly meaning-properties and their characteristic features or “symptoms”, since the task of investigating the nature of meanings will largely consist of finding out what non-semantic properties can explain such “symptoms”.

How do we identify meanings in the first step of the investigation? Since the competing philosophical theories of meaning—including semantic-level analyses—are not neutral in the required sense, we cannot simply rely on them in the way that molecular genetics relied on Mendelian genetics. What we need instead is some apparently uncontroversial examples of properties that “identification experts” recognize _pre-theoretically_ as semantic, and we need to establish what features of these properties the experts rely on in their identification. In this case, like in the case of the identification of water, the relevant identification experts are ordinary people. As Devitt points out, ordinary people have the competence “to produce token thoughts and utterances to which the [meaning] properties are ascribed” (Devitt 1996: 78). More importantly, ordinary people are competent to _ascribe_ meanings to the utterances and thoughts of others. These ascriptions provide the uncontroversial examples we need: “To identify examples in the preliminary first stage, we look to ordinary attitude ascriptions” (Devitt 1996: 78). This does not mean that philosophers of language must engage in an anthropological or sociological
study of folk ascriptions, as interesting as that may be. As Devitt points out, “theorists can count themselves among the expert folk” (Devitt 1996: 78). Of course, the theorist must set aside philosophical assumptions about the nature of meaning. The goal is to (1) find samples of meanings and (2) analyze what are they supposed to do—what do we ascribe them for _qua_ folk semanticists—before (3) proceeding to study their underlying nature.

2.3.2 Identifying Meanings by their Causal Roles

Devitt and Horwich propose a way of identifying meanings that is neutral, in the sense that it does not assume any particular account of the nature of meaning. Basically, they suggest that we should identify meanings by the _causal roles_ of representations. While Horwich (1998a, 2005, 2010) focuses on the _causes_ of linguistic and mental representations, Devitt (1996) originally focused on their _effects_. But Devitt (2002 & 2011) later suggests that we should focus on _both_ aspects of their causal roles. In the previous sections I argued that we need a non-question-begging identification of meanings. I shall argue in this section that we should identify meanings as the properties that we ordinarily ascribe based on the causes and behavioral effects of linguistic and mental representations. These properties can be recognized pre-theoretically as semantic—without assuming any theory of meaning—and we need an explanation of their nature, so investigating them is a worthy task. Notice that identifying meanings _by_ their causal roles is a very different proposal than that of identifying meanings _with_ their causal roles.

The topic of how folk semanticists identify meanings is one I already covered in Chapter 1, where I analyzed ordinary attitude ascriptions. I argued, following Devitt, that we ascribe

---

Devitt (1997) expanded his identification of meanings to also include the _causes_ of representations.
contents to the utterances and thoughts by means of ‘that’-clauses that function as indefinite singular terms that specify the meanings of such utterances and thoughts. These ascriptions provide us with identified samples of what needs to be explained—and such identifications are pre-theoretical in the required sense: they do not assume anything about the underlying nature of the identified properties. I also talked in Chapter 1 about one of the main purposes for which we ascribe meanings: namely, to explain other people’s behavior. Indeed, Devitt suggests that we should focus on the purposes for which the folk make attitude ascriptions using ‘that’-clauses: one of them is to explain behavior, but another one is to use other’s thoughts and utterances as guides to reality (Devitt 1996: Ch. 2 & 2012: 1.1).

Consider the ascription ‘Nancy fears that it will rain’. Such ascription may play a causal role, for example, in the explanation why Nancy is constantly looking at the sky. Or consider the ascription ‘Mary believes that it is raining’. It may help us to explain why Mary picks up an umbrella before leaving her house. Alternatively, consider the ascription ‘The weather guy said that it will rain’. This ascription serves as a guide to reality. When a person’s thoughts or utterances are reliable, we can learn about the world by ascribing the right meanings or contents to them. Notice that we typically ascribe meanings to thoughts and utterances based on what we take to be the causes of such tokens, which in turn helps to explain their effects. For example, if we witness that John runs away after seeing a dog, we may ascribe to him the thought that the dog is dangerous. The content we ascribe is identified both by its presumed causes and effects.

It is important to recall two points I made in Chapter 1. One is that our ordinary
ascriptions of content to explain other people’s behavior are often successful. Now we can add that our ascriptions of content to other’s utterances and thoughts to learn about reality are also often successful. This success provides very good evidence that the meanings ascribed in folk-psychology are real phenomena that play causal roles in the physical world. This means that we are overall correct in our pre-theoretical identifications of meanings and that investigating the underlying nature of such meanings is a worthwhile task. Also, it is worth insisting here that the meanings thereby identified must be natural phenomena, since otherwise they could not play the causal roles they do play. As I argued before, what is required is a naturalized semantics.

It is worth considering for a moment the case of attributions of content to non-human languages. Cognitive ethologists attribute “referential” languages to bees, prairie dogs, monkeys, chickens, etc., because of their causal role of conveying information about the world from one organism to another. The attributions also play a causal role in the ethologists’ explanation of animal behavior. Let us consider a couple of examples. Cognitive ethologists have discovered that vervet monkeys use three distinct vocalizations to warn their peers of the presence of leopards, eagles and snakes, respectively (Cheney and Seyfarth 1990). They have learned this because each of these alarm calls is paradigmatically caused by the presence of the relevant predator and the calls trigger three different kinds of avoidance behavior that are appropriate to avoid being hunted by that particular kind of predator. Similarly, chickens use two distinct vocalizations to warn of the presence of aerial predators and land predators (Evans and Marler 1995). Again, each vocalization is paradigmatically caused by the presence of one or the other kind of predator and triggers an avoidance behavior in the audience that is appropriate to the particular kind of predator being announced. In these cases of non-human languages we clearly
see what sorts of causes and effects are relevant to the identification of meanings. Both the “typical” causes and effects of utterances are very directly connected in these simple languages to the purposes for which meanings are ascribed to them. The utterances serve the purpose of guiding conspecifics away from predators and the meanings ascribed by scientists serve the purpose of explaining the animals’ behavior.

I end with a short note on the meanings of words and concepts. By focusing on ‘that’-clauses in the case of ascriptions to humans, I have played attention to the meanings of complete utterances and thoughts. But as folk semanticists we also ascribe meanings to the words and concepts that compose them. The meanings of words and concepts are usually identified within ‘that’-clauses. If I ascribe to Mary the belief that it will rain, I am ascribing to her a concept with the same meaning as the word ‘rain’ in the embedded sentence in the ‘that’-clause. Similarly, if I report ‘John said that it will rain’, I am likely ascribing to one of the words uttered by John the same meaning as the word ‘rain’—even if John spoke in German and my report serves as a translation. The meanings of words have to be abstracted away from the meanings of sentences where they appear. But in the end the meanings of words are also pretheoretically identified by their causal roles. As Horwich points out: “[We] recognize pretheoretically… that what people say is due, in part, to what they mean. I assent to ‘That’s red’, when I do, partly because of what I mean by the word ‘red’.” (Horwich 1998a: 47). Also: “[The] symptom of a word’s meaning is its overall use, principally, the collection of sentences containing it that are accepted and the circumstances under which this is done.” (Horwich 2005: 77). Horwich’s very reasonable point here is that we identify the meanings of words by the way they are used and, more specifically, as what is causally responsible for such use. So a crucial explanatory target for a theory of
meaning is to explain in non-semantic terms what properties are indeed causally responsible for the use of words.\footnote{42}

Notice that the identifications here proposed regarding what meanings it is the task of semantics to explain are theoretically neutral. They do not beg the question for or against truth-referential approaches, whether directly referential or “Fregean”, functional-role approaches, etc. Any existing or new approach is free to attempt to explain the underlying nature of the identified meanings. The virtue of these identifications is that they provide a non-question-begging definition of what is it the task of semantics to explain.

2.4 Conclusion

I have argued that neither a correspondence nor a deflationary view of the nature of truth should be assumed in the definition of the task of semantics, largely because it is an open question which of these views will end up succeeding. Moreover, the dispute is likely to be resolved based on whether substantial truth and referential relations turn out to be required to explain the nature of meaning. In general, I argued that we need a pre-theoretical identification of meanings. I suggested that we should identify meanings by the causal roles of utterances and thoughts, and that such identification specifies the task of semantics. Finally, I argued that the ‘that’-clauses deployed in ordinary attitude ascriptions in order to explain the behavior of others and in order to learn about reality provide the needed identifications.

\footnote{42}{As I argued before, following Devitt (2011: 197-198 & 2002: 112), Horwich should expand his identification of meanings to include also behavioral effects (see Chapter 1, Section 1.7).}
Chapter 3: Basic-Acceptance Semantics

3.1 Introduction

In the first part of this dissertation, Chapters 1 and 2, I focused on fundamental metaphysical and methodological issues related to the nature of meanings and the task of semantics. I argued that meanings are phenomena that play causal roles in the physical world and that we need a naturalistic explanation of their underlying natures. I also argued that the task of semantics is to explain specifically the meanings pre-theoretically identified by the causal roles of linguistic and mental representations. A proper theory of the meanings of words should account for what non-semantic properties are responsible for their overall uses.

The main aim of the second part of this dissertation, from Chapter 3 to Chapter 6, is to explore what sort of naturalistic theory promises to explain what non-semantic properties constitute meanings. My main focus is on the meanings of human words and their mental counterparts. I will propose a combination of (1) a “use” theory along the lines of Horwich’s theory of meaning with (2) a teleological account of the functions of symbols. In this chapter I briefly review the range of available theories of meaning, briefly sketch the combination that I will propose and focus on motivating a “use” theory along the lines of Horwich’s theory. But I also point out the problems faced by the theory. In a nutshell, the main advantages are that it promises to explain properly the overall uses of human words/concepts, including their complex roles in inferential processes, and also that it can account for the variety of meanings that human words like ‘red’, ‘bachelor’ and ‘or’ exhibit. One problem with the theory is that it risks
collapsing into a form of truth-referentialism in spite of being presented as a match for a deflationary view of reference and truth. But this problem does not concern me in particular, since I am happy to embrace such a collapse. I argue that in such a case we would anyway need a truth-referential version of the theory. A more serious problem is that Horwich’s theory is unable to account for misuses of words due to ignorance or error. It is this problem that motivates the search for a teleological component that I undertake in the following chapters.

3.2 The Place of Meaning in the World

What is the place of meaning in the world? In virtue of what do our linguistic utterances and thoughts have the meanings they have? The nature of meaning is one of the most fundamental issues at the core of both the philosophy of language and the philosophy of mind. As Jerry Fodor puts it: “It appears increasingly that the main joint business of the philosophy of language and the philosophy of mind is the problem of representation itself: the metaphysical question of the place of meaning in the world order.” (1987, p. xi)

The debate on the nature of meaning involves multiple dimensions. One of the many questions that a theory of meaning has to answer is what is the relation between the meanings of linguistic utterances and thoughts. Some theories of meaning have been originally proposed in the philosophy of language and their primary targets are the meanings of words and sentences in natural languages (e.g., Frege 1948 and Kripke 1980). Others have been proposed in the philosophy of mind and their primary targets are the meanings or contents of thoughts and concepts (e.g., Fodor 1987 and Block 1986). Others simply treat language and thought as a seamless whole, assuming that we use the same words to think and to communicate and that their
meanings do not depend primarily on one of these uses (e.g., Sellars 1954 and Brandom 1994).

It is clear that the meanings of linguistic utterances and the meanings of at least some thoughts are intimately related. After all, one of the primary uses of language—likely the primary use—is to express thought. As Fodor, Bever and Garrett point out: “there is much to be said for the old-fashioned view that speech expresses thought, and very little to be said against it” (Fodor, Bever, and Garrett 1974: 375). Someone can express the belief that the weather in New York is horrible by uttering ‘the weather in New York is horrible’ or a number of sentences in other languages with the same meaning. All these sentences and the belief they express have the same meaning or representational content. Additionally, the meanings or contents of many of our concepts—probably most of them—come from the meanings of words. Most of us can think about genes because we have heard of them: the content of our concept comes from the meaning of the word. However, this does not entail that linguistic meaning has the ultimate explanatory priority over mental content. After all, some people must have originally thought about genes to coin the word ‘gene’, which was then passed along from speaker to speaker until we acquired our concept from it (Devitt & Sterelny 1999: 152; Devitt 1996: 157-158).

How are linguistic and mental meanings related? Many authors think that the meanings of linguistic utterances depend on conventions (Lewis 2002) that ultimately link them to the meanings of thoughts (Grice 1989). Given the obvious conventional character of the words of natural language, together with the consideration in favor of the ultimate priority of thought just mentioned, I think that this account must be heading in the right direction—although there is room for filling in more details about the nature of the conventions involved in language. I also support the promising though controversial hypothesis that thoughts take place in an internal
language of thought (Harman 1973; Fodor 1975). A very plausible explanation of the *productivity* and *systematicity* of mental states like beliefs and desires is that they have sentence-like structure (Fodor 1987: 147-154). More importantly, the language of thought hypothesis offers the best available account of the *mental processes* involved in thinking, which it explains as computational operations performed over syntactically structured inner representations (Fodor 1987: 143-147). Devitt reasonably argues that the later consideration is more decisive, since other kinds of structures—e.g., map-like structures—may also be productive and systematic (Devitt 1996: 154-156).

If these views are correct, as many theorists of meaning expect today, part of what we need to do in order to find the place of meaning in the world is to explain the nature of the relevant conventions linking the words used in natural languages to their mental counterparts, but the most fundamental task is to explain the sources and nature of the meanings of words in the language of thought. I will later elaborate an account of how the meanings of linguistic words depend on the meanings of mental words. While I endorse the Gricean view about the priority of thought, I will not adopt the particular mechanism he proposes to account for the literal meanings of words. Following Millikan (2005), my account of linguist conventions will depart from Lewis’ in being *historical* and teleological—involving the historical *purposes* of linguistic conventions. For the moment, however, I will simply talk about “words”, including their uses in thought and communication. We need an account of the meanings of words independently of what account of the relation between the semantic properties of mental and linguistic representations turns out to be true. Treating language and thought initially as a “seamless whole” will simplify the discussion of other dimensions of the debate.43

43 Millikan (2005: 84) complains that some authors treat language and thought as a “seamless whole”. I will
In virtue of what do words mean what they do? While a number of theories have been proposed to answer this fundamental question, there is no agreement in sight. The purpose of the second part of this dissertation is to propose a new answer to the question. My proposal borrows components from already existing theories. Mainly, I combine the core component of Horwich’s (1998a) use-theory of meaning—namely, that the meanings of words depend on their *roles in some sentences*—with the core component of teleosemantics (Millikan 1984; Papineau 1984)—namely, that the meanings of words depend on their *historical functions or purposes*. The novelty of my approach comes from this particular combination of components, which has not been proposed before but, I shall argue, promises to overcome some of the most serious shortcomings of available theories of meaning, including those I borrow components from. In a nutshell, my proposal is that words mean what they do in virtue of their roles in some sentences that have the historical function of governing their use.

Before articulating and defending my own proposal, I shall briefly mention the main competing theories in the debate on the nature of meaning. My focus in the reminder of this chapter will be on the virtues and problems faced by Horwich’s “use” theory.

### 3.3 The Range of Available Theories

There is a panoply of theories attempting to explain why words mean what they do. According to the *description* theory (Frege 1948; Russell 1905), the meanings of words depend on *links to other words* in descriptions that determine their reference. This theory has plausibility for words like ‘bachelor’, but is essentially incomplete: the reference of ‘bachelor’ may depend on

---

44 Searle (1958) modified the theory to rely instead on *clusters* of descriptions.
that of ‘unmarried men’, but what determines the reference on ‘unmarried’ and ‘men’? According to another family of theories, the meanings of words depend instead on *links to the external world* that determine their reference: *indicator* theories (Dretske 1981; Fodor 1987) claim that the relevant links involve the reliable causes of words, other theories (Kripke 1980; Devitt 1981) rely on their *causal-history*, while *teleosemantic* theories (Millikan 1984; Papineau 1984) rely on their historical functions—which they explain in terms of natural selection or an analogous evolutionary process. The three kinds of theories in this family are typically *atomistic*, including *only* links to reality and excluding any links to other words among the links that determine meaning.\(^45\) The description theory can be characterized instead as *molecularist*, since it includes *some* links to other words.

According to yet another family of theories (Block 1986; Harman 1982; Brandom 1994; Horwich 1998a), the meanings of words depend on connections to inputs from perception, outputs to behavior and roles in inference that determine their *uses or functional-roles*.\(^46\) (What I am calling “functional” roles are often called “conceptual” or “inferential” roles. As Devitt (1996) points out, talk of “functional” roles is preferable because the latter two encourage the overlooking of links to inputs and outputs.) Functional-role or “use” theories are typically *holistic*, including *all* the inferential, perceptual and behavioral connections of words among the links that determine their meanings (Block 1986; Harman 1982; Brandom 1994). In contrast, Horwich (1998a) claims that the meanings of words depend *only* on their functional-roles in *some sentences*, which may link a word *either* to the world, to some other words or to rules of

---

\(^{45}\) An exception is Papineau (1984; 1987), who includes *all* the functional links of words in their meanings.

\(^{46}\) The view has its origins in the work of Wittgenstein (1953) and Sellars (1954).
inference. This functional-role theory is molecularist rather than holistic.\footnote{Peacocke (1992) proposes a similar theory based on the primitively compelling functional-roles of concepts.}

The mainstream approach in semantics is truth-referentialist: the view is that meanings depend on properties that determine the \textit{references} of words which, together with syntactic structure, determine the \textit{truth-conditions} of sentences containing them. Frege (1948) proposed that the meanings of words include their \textit{references} and their \textit{senses} or “modes of presentation”: the associated descriptions that determine their references. Based on this distinction, Frege offered a neat account of how co-referring words can differ in meaning: different senses can determine the same reference. Most theories relying on causal links to the external world regard meanings instead as \textit{exhausted} by their referential properties, so co-referential words cannot differ in meaning. Devitt (2001b), however, expands the Fregean notion of \textit{senses} to include \textit{causal} modes or reference constituted by links to the world. This allows him to offer a neat Fregean account of how co-referring words can differ in meaning, even when their meanings depend on links to the world rather than on associated descriptions.

In contrast with this truth-referential tradition, most functional-role or “use” theories claim that the functional-roles involved in meaning \textit{do not determine reference}. Some theorists (e.g., Block 1986) propose that meanings are composed of \textit{two factors}: a purely internal and non-reference-determining \textit{functional-role factor} and a separate \textit{referential factor} involving links to the external world. Other theorists reject altogether the referential factor, since they endorse \textit{deflationary} views (Field 1994; Horwich 1998b) according to which truth and reference are not substantive properties and, consequently, cannot play an explanatory role in the theory of meaning. Use-theories that endorse deflationism (Harman 1987; Horwich 1998a; Brandom 1994)
include links to the world in their accounts: their functional-roles are not purely internal. However, they claim that these external links do not determine reference. A less common—but promising—proposal is that some of the functional-roles of words do determine their reference: Devitt (1996) argues that the meanings of some words—those covered by a description theory—include some functional-roles that determine their reference, while Peacocke (1992) argues that the referential properties of all concepts are determined by their primitively compelling functional-roles.

Whether the meanings of words depend on functional-roles and whether these roles determine reference are two different questions. The expressions “functional-role semantics” and “use-theory” are often used to characterize non-referential theories or at least theories that include a non-reference-determining factor. But this common usage obscures the fact that functional-roles may substantially determine reference. In what follows, I will depart from this common usage and I will not assume that functional-roles do not determine reference.

I end this brief survey presenting another fundamental dimension on which theories of meaning diverge. Most of the theories mentioned above are in the business of naturalizing semantics: explaining meaning in terms of properties that words have in the natural world, such as the reliable causes (e.g., Fodor 1987), the historical causes (e.g., Devitt 1981), the evolutionary functions (e.g., Millikan 1984) or the causal/functional-roles (e.g., Harman 1982) of their physical tokens. However, other theories abandon naturalism, claiming that the meanings of words depend on norms prescribing how they ought to be used that cannot be reduced to any

---

48 The common usage is also inconsistent. Most authors that adopt it consider, for example, Peacocke (1992) as a functional-role theory of the content of concepts, despite its substantial account of reference determination.
aggregate of natural facts (Kripke 1982). Brandom’s (1994) inferentialism—a use theory based on irreducible normativity—is the most developed version of this non-naturalist approach.

The panoply of theories offered to explain meaning shows that this is a very difficult enterprise. I shall argue that many of the current proposals contain at least a kernel of truth, providing some important insight into the nature of meaning, but that every single one of them is undermined by one or another serious problem. Each proposal combines a set of claims. While some claims of a proposal may contain important insights, other claims of the same proposal may be problematic. To make progress, we need a new approach that combines the insightful claims of different theories while discarding the problematic ones. I will soon discuss some of the main insights and problems.

I have presented in this section a brief map of the territory: the vast range of available accounts of why words mean what they do. In the next section, I will clarify what are the two main components that my own proposal combines to answer this question. Afterwards, I will argue for each of the components and articulate how I propose to combine them.

3.4 A New Approach to Explain Meaning

The new approach I propose to explain why words mean what they do combines elements from functional-role semantics and teleosemantics. This combination, I shall argue, provides a framework to incorporate the insights of further theories. The basic idea is that words mean what they do in virtue of the core functional-roles they are “supposed” to have in order to perform their historical functions: the functions that account for why the words exist and have been kept

---

49 Kripke’s (1982) influential work does not endorse this view. He presents it as a reading of Wittgenstein’s (1953) discussion about rule-following.
in circulation. These core functional-roles have the historical function of governing word usage.

The functional-role component of my approach comes specifically from Horwich’s (1998a; 2005) use-theory of meaning, according to which the meanings of words depend on their core functional-roles: their roles in some basic sentences that users (speakers / thinkers) accept or regard as true. Horwich argues that these basic sentences govern the uses of words (2005: 37-42) and vary in kind for different words, connecting them either to the external world, to other words or to rules of inference (2005: 48-49). I borrow this set of central claims from Horwich’s use-theory. I shall call this set of claims “basic-acceptance semantics”.

Horwich’s use-theory, however, includes other commitments—beyond basic-acceptance semantics—that I do not borrow or endorse. Horwich (1998b) is one of the leading advocates of deflationism, so his use-theory is committed to the claim that meanings do not substantially determine reference. But Devitt (2002) points out that his use-theory risks collapsing into truth-referentialism, since it includes causal connections between words and the world that are among the candidates to naturalize reference. I shall remain agnostic about deflationism, since basic-acceptance—given its links to the world—may indeed determine reference. Devitt (2002; 2011) also shows that Horwich’s use-theory cannot tell apart misuses—due to ignorance and error—from the correct uses of words. This problem, I shall argue, stems from Horwich’s commitment to semantic dispositionalism: the view that meanings depend on causal dispositions or regularities underlying word usage. I shall reject any form of dispositionalism, since it cannot account for the difference between the correct and incorrect uses of words (Kripke 1982). My approach basically replaces Horwich’s dispositionalism with the historical account of the functions of symbols proposed by teleosemantics.
The teleosemantic component of my approach is the core claim shared by all available\textsuperscript{50} and possible teleosemantic theories; namely, that the meanings of symbols depend on their historical functions, which are determined by the selective history underlying their use and/or the mechanisms responsible for it. These are functions that explain why symbols exist. The selective processes may take place at the genetic level or at the level of learning and cultural inheritance. A combination of these levels is needed to account for the meanings of all words and concepts. Teleosemanticists argue that symbols need not perform their functions reliably or regularly. They need only perform them often enough to be kept in circulation (Millikan 1984). Teleosemanticists also argue that a theory of meaning based on historical functions can explain why some uses are correct, while others incorrect: the correct ones are those where symbols perform their historical functions (Millikan 1984; Papineau 1984; Neander 1995). Any theory of meaning committed to these central claims is a version of teleo-semantics, since it adopts a teleological or teleonomic account of functions—based on selective history—to explain the semantic properties (i.e., the meanings) of symbols. My approach is committed to these central claims, so it is a version of teleosemantics—even though it combines them with basic acceptance semantics.

Available teleosemantic theories, however, typically share further commitments that I will not endorse. The currently available versions of teleosemantics claim that the historical functions of all symbols involve links to the external world. But this is plausible neither for words whose meanings are likely to involve only links to other words (Devitt & Sterelny 1999: 156-157) nor for the terms of logic and mathematics (Horwich 1998a: 53). I shall argue that the

historical functions of words involve instead their roles in basic sentences that link them either to the external world, to other words or to rules of inference. Also, available teleosemantic theories are committed to the historical functions of words substantially determining their reference. I accept this is very likely to be true. But I shall remain agnostic about deflationism and, consequently, open to the possibility that the historical functions of words may not substantially determine their reference.\footnote{Millikan (1993: 231-235) argues that a teleosemantic theory must involve substantive (i.e., correspondence) reference and truth, since the historical functions of symbols relate them to real distal events in the external world. While her argument rules out a verificationist notion of truth as warranted assertability, which threatens realism about the external world, it does not rule out a deflationary notion. As Devitt (2010: 34-35) shows, deflationism is fully compatible with realism. If what teleosemantics needs to work is realism, nothing more than deflationary truth is required.} The goal is to explain the underlying nature of the meanings pre-theoretically identified by the causal roles of words. If the non-semantic properties that end up explaining meanings turn out to provide a proper reductive explanation of truth and reference, as I think it is likely, truth-referentialism will be vindicated. But I think plenty of progress can be made before fully settling that issue.

A combination of basic-acceptance semantics and teleosemantics, I shall argue, can simultaneously overcome some of the most serious problems of available theories while incorporating their valuable insights. In particular, the combination can overcome the shortcomings that Horwich’s theory and available teleosemantic theories have on their own. Horwich’s theory, I shall argue, cannot tell apart the misuses from the correct uses of words; but the combination I propose can do so, since its teleosemantic component is tailored to explain why some uses are correct. Standard teleosemantic theories, I shall argue, cannot account for words whose meanings depend on links to other words or to rules of inference and cannot explain how the meanings of words govern their complex inferential uses in cognition; but the
combination I propose can do so, since its basic-acceptance component is tailored to explain how
the meanings of all kinds of words govern their various uses. I will not argue that my approach
answers all the questions a theory of meaning must answer. Some questions will remain open—
notably, whether the features that explain meaning substantially determine reference. But I shall
argue that they can be answered within the framework provided by my approach.

In this section, I have presented the two components I propose to combine into a new
approach to explain the nature of meaning: the core claims of Horwich’s use-theory—i.e., basic
acceptance semantics—and the core claims of teleosemantics. I have also mentioned which
aspects of Horwich’s theory and standard teleosemantics I either propose to reject or remain
neutral about. In what follows, I will argue that a theory of meaning does need to include each of
the two components I am combining and reject the problematic aspects of the theories these
components come from. I will also argue that the fate of deflationism is still undecided, so we
should remain initially neutral about it. Finally, I will elaborate on how to combine basic-
acceptance semantics and teleosemantics. I shall start by motivating basic-acceptance semantics.

3.5 Why Basic-Acceptance Semantics?

The claim that the meanings of words depend on their roles in some sentences that govern their
use is the central claim of Horwich’s use-theory. I will first elaborate on his theory and the main
arguments for it. Then I will discuss the main problems faced by the theory. Based on this
assessment, I shall argue that an account of meaning should adopt basic-acceptance semantics,
reject semantic dispositionalism and remain initially neutral about deflationism.
3.5.1 Horwich’s Use-Theory of Meaning

As I discussed in Chapter 2, the prevailing view is that truth and reference play a substantial role in the explanation of meaning. A challenge to this truth-referentialism comes from deflationary views according to which truth and reference are not properties with an underlying nature and, accordingly, cannot play a metaphysically substantial or significant role in explanations of any kind, including explanations of why words mean what they do. I pointed out that deflationists (Field 1994; Horwich 1998b) base their view on a plausible account of the role of the truth term. Whether their account of the nature of truth is equally plausible is controversial (Devitt 2001a). Nevertheless, if deflationism is true, we need an alternative approach to explain the nature of meaning. And the best—perhaps only—alternative seems to be the “use theory” or functional-role semantics, since the uses or roles of words in language and thought—which depend on their connections to other words, inputs from perception and outputs to behavior—presumably do not substantially determine their reference.\footnote{As I pointed out in Chapter 2, Section 2.2.3, and in this chapter, Section 3.3, this may turn out to be a wrong assumption. But notice that two-factor versions of functional-role semantics endorse this assumption but reject deflationism. Their case relies on the need of a functional-role factor in psychological explanation (Block 1986 & 1987). Deflationary versions of functional-role semantics (Harman 1987; Horwich 1998a) typically make a similar case about psychological explanation: their theories are not exclusively motivated by deflationism.} Horwich (1998b) provides one of the most elaborate and influential defenses of the deflationary theory of truth and reference. His use-theory of meaning (Horwich 1998a & 2005) is intended to be compatible with—and largely motivated by—deflationism.\footnote{Other theorists that combine a “use theory” with deflationism include Brandom (1994), Field (1994) and Harman (1974, 1982 & 1987).}

Horwich’s use-theory of meaning (UTM) sides with naturalist versions of functional-role semantics (FRS), which understand functional-roles in terms of causal connections of words (e.g., Block 1986 & 1987; Harman 1968 & 1982), and against non-naturalist versions that offer...
irreducibly normative accounts of meaning (e.g., Brandom 1994 & 2000). The aim of UTM is to offer a reductive account of meaning constitution in terms of non-semantic and non-normative facts. Unlike other versions of FRS, however, UTM does not identify the meaning of a symbol with its overall use or functional-role. Rather, UTM identifies the meaning of a word with what explains its overall use. So UTM regards merely as the explanandum what most versions of FRS conceive as the explanans. What makes UTM a version of FRS is that its explanans is a subset of “explanatory fundamental” uses or functional-roles: UTM claims that the meaning of a word is constituted by the basic sentence acceptance property that explains its overall deployment. Horwich (2005: 28) offers the following “short crude statement” of UTM:

The meaning of a word, w, is engendered by the non-semantic feature of w that explains w’s overall deployment. And this will be an acceptance-property of the following form:—‘that such-and-such w-sentences are regularly accepted in such-and-such circumstances’ is the idealized law governing w’s use (by the relevant ‘experts’, given certain meanings attached to various other words).

UTM characterizes the use of a word in terms of its presence in sentences that are regarded as true or accepted, including not only uttered sentences, but also sentences deployed in thought—i.e., sentences in the speaker’s “belief-box”. According to UTM, the totality of accepted w-sentences (sentences containing the word w) can be explained by a small subset of accepted w-sentences. This hypothesis appeals to “a familiar inferential model” to explain the overall use of

---

54 As Brandom (1994, p. 625) characterizes his own view: “the story is one in which it is norms all the way down”.
55 A minor typographical mistake has been corrected in this quotation.
56 Indeed, UTM treats uniformly the use of expressions in speech and in the language of thought (Horwich 2005: 31). Basically UTM treats language and thought as a “seamless whole”. Devitt (2002) suggests that UTM should give instead priority to the mental. I will follow Devitt’s suggestion when I combine basic-acceptance semantics with teleosemantics.
words (Horwich 2005: 43). The idea is that the acceptance of most w-sentences deployed in speech and in theoretical and practical reasoning can be explained as the result of inferences from other w-sentences; but there must be some w-sentences that (1) serve as the inferential basis for other w-sentences but (2) cannot themselves be inferentially derived from previous w-sentences. According to UTM, the acceptance of these core w-sentences is *explanatorily basic* and constitutes the meaning of w because it accounts for the acceptance of all other w-sentences.57

One of the main advantages of UTM over other versions of FRS comes precisely from not identifying meanings with overall uses. A serious problem with FRS is that it seems to lead to semantic holism, since the functional roles of a symbol include the totality of its causal/inferential connections. Although proponents of FRS usually endorse this result (Block 1995; Harman 1993), semantic holism is extremely implausible (Fodor and Lepore 1992; Devitt 1993 & 1996). As Fodor puts it, “Meaning Holism really is a *crazy* doctrine” (Fodor 1987: 60). Semantic holism entails that the meaning of a word in a user’s idiolect changes whenever the user accepts new sentences or rejects old sentences containing the word. This leads to the radically absurd consequence that a word in a user’s idiolect cannot mean the same over time and that no two users can ever mean the same, given the obvious discrepancies between people’s webs of beliefs. UTM, however, is a *non-holistic* version of FRS, since it only regards some core uses as meaning constitutive. If the meaning of ‘bachelor’ stems from the user’s acceptance of ‘Bachelors are unmarried men’, it will not change when user comes to accept a new sentence, such as ‘Bachelors are unhappy’.

57 The various w-sentences contain other words, not just w, so their acceptance is not explained merely in terms of the core w-sentences, but rather by the core sentences governing each of these words.
Another virtue of UTM is that it is a non-solipsistic version of FRS. A persistent worry about FRS is that it seems to be a purely internalist doctrine, since it relies on the roles of words within the cognitive system. But compelling arguments for physical externalism show that meanings must involve relations between words the external world (Putnam 1975; Kripke 1980).

Two-factor theories attempt to avoid this problem by adding a referential factor that connects words to the world to the solipsistic—purely internal—functional-role factor that plays a role in the user’s psychology (Block 1986). However, they face serious problems that stem from their postulation of two separate factors. One problem is that they do not offer an account of what coordinates the separate factors (Fodor 1987: 82-83). To put it bluntly, nothing in the theory prevents a word from having, for example, the functional role of ‘cat’ and referential connections dogs. Another problem is that the external factor plays no role in the psychological explanation of behavior. This is deeply troubling, since the explanation of behavior must account for how we deal with the world (Harman 1988 & 1999: 235-243). In contrast with two-factor theories, UTM makes room for externalism by including connections to the external world among the core functional-roles of words. These are wide functional roles that reach out into the world instead of narrow ones encapsulated in the head. The basic acceptance-property governing the use of ‘cat’, for example, may involve accepting ‘This is a cat’ in the presence of cats. Accordingly, this wide functional role explains the various uses of ‘cat’, in theoretical and practical reasoning leading to intentional behavior, in terms of the links to cats—in the world—of sentences that govern these inferential processes. The connections to the world and the psychological roles of the word are fully coordinated into a single factor, so the external link plays a substantial role in the explanation of behavior. Harman (1987) also proposes a wide or non-solipsistic version of FRS.
But his theory is hopelessly holistic.

By being non-solipsistic and non-holistic, UTM has clear advantages over other versions of FRS. If we need a functional-role or use-theory, UTM is the best candidate. But why should we favor UTM over non-use-theoretical alternatives? One reason may be deflationism about reference and truth. If deflationism is right, we need a use-theory and the arguments above suggest that we should favor a basic-acceptance account—e.g., UTM—over other use-theories. However, Horwich does not base his case for a use-theory exclusively on deflationism—which is good because the argument based on deflationism is not too strong, given the risk of UTM collapsing into a form of truth-referentialism (see Section 3.6.1). He also argues that UTM has two crucial advantages over non-use-theoretical alternatives: its generality and its explanatory power (1998a: 53). In what follows, I will present these two arguments and argue that together they show that we need an acceptance-based account of meaning independently of the fate of deflationism. In other words, that we need an acceptance-based account even if the meanings of words turn out to substantially determine their references. I start with Horwich’s generality consideration.

3.5.2 The Generality of a Basic-Acceptance Account

The meanings of different words vary considerably, so theories that are plausible for some terms usually are not plausible for others. UTM, in contrast, can account for all classes of terms. Horwich (2005: 26-27) gives the following examples of the meaning constitutive properties we may expect to find according to his use-theory:

---

58 In fairness, Peacocke’s (1992) theory of concepts is also non-solipsistic and non-holistic. Indeed, it is very similar to UTM in crucial respects.
“bachelor”’s meaning is engendered by the fact that its basic regularity of use is our acceptance of the sentence, “The bachelors are the unmarried men.”

“red”’s meaning stems from the fact that its law of use is a propensity to accept “That is red” in response to the sort of visual experience normally provoked by observing a clearly red surface.

The meaning of “water” is constituted by the fact that the law explaining its overall use is that we accept, “x is water ↔ x has the underlying nature of the stuff in our seas, rivers, lakes and rain.”

“neutrino” means what it does in virtue of our unsupported acceptance of the conditional, “∃øT(ø) ⇒ T(neutrino)”, where “T(neutrino)” is a formulation of neutrino theory.

“and” means what it does because the fundamental regularity in its use is our acceptance of the two-way argument schema, “p, q // p and q”.

Notice the variety of meaning constituting properties allowed for by UTM. The meaning of ‘bachelor’ is accounted for by its association with other words. The meaning of ‘red’ is causally related to red surfaces. The meaning of ‘neutrino’ is determined by its role within a theory. And the meaning of ‘and’ is determined by an argument schema or rule of inference. The shape taken by these different basic acceptance properties accommodates insights from description theories (‘bachelor’), indicator or causal theories (‘red’), traditional FRS (‘neutrino’) and even solipsistic FRS (‘and’). The generality of UTM cannot be matched by any of these other theories. The description theory is plausible for terms like ‘bachelor’, but cannot account for terms like ‘red’,
‘horse’ and ‘water’. Causal theories are plausible for terms such as ‘red’, ‘horse’ and ‘water’, but implausible for terms like ‘bachelor’ and ‘neutrino’. And purely internalist versions of FRS give plausible accounts only of logical terms like ‘and’. None of these theories provide a unified reductive account of the meaning constituting properties of all words (Horwich 1998a: 53).

Due to its generality, a basic-acceptance semantics—such as UTM—has a crucial advantage over almost every other theory offered to explain why words mean what they do. There is, however, an alternative to UTM that can accommodate different kinds of meaning. As Michael Devitt (2002: 113-114) points out:

The alternative is a “moderate” truth referentialism along the following lines.

“Primitives” get their meanings from referential relations explained by some sort of direct causal link to reality, an informational, teleological, or historical-causal link, or some combination of these; proper names and natural kind words are likely primitives. Other words get their meanings from referential relations explained by their “definitional” links; the words are inferentially associated with others that determine their reference; they are covered by “description” theories of reference; ‘bachelor’ is a likely example.

Devitt’s moderate truth-referentialism incorporates the insights of different theories that regard meanings as constituted by reference-determining properties.59 So the approach is diametrically opposed to Horwich’s UTM, which is committed to deflationism and does not regard meaning constituting properties as reference-determining. Notice that by making room for description theories, informational theories, teleological theories and historical-causal theories, moderate

59 Devitt (1996) and Devitt & Sterelny (1999) argue for this moderate view.
truth-referentialism offers a plausible account of the meanings of different kinds word, unlike any of these particular theories taken by itself.

Moderate truth-referentialism explains the meanings of primitive words in terms of direct links to reality and the meanings of non-primitive words in terms of indirect links to reality provided by connections to other words. However, even this moderate approach does not seem particularly well-suited to explain the meanings of the terms of logic and mathematics (e.g., ‘and’ or ‘plus’), which appear to be related to their roles in inference rather than to any direct or indirect link to the external world (Field 1977; Harman 1987). An acceptance-based theory, like UTM, is even more general than moderate truth-referentialism, allowing the meanings of different words to depend on basic sentences that link them either to the world, to other words or to rules of inference. The links to rules of inference are crucial to account for the meanings of the terms of logic and mathematics. But moderate truth-referentialism only includes links to the world and to other words. So UTM has an advantage even over moderate truth-referentialism. As Devitt (2002: 117) acknowledges: “I do think that the use theory has an advantage over its rivals in the handling of the terms of logic and mathematics”. Fodor (1994: 110) makes a similar point:

I’m inclined to think that maybe there is no objection to the idea that “+”, “and”, “all” and the like have the meanings they do because they play a certain causal role in the mental lives of their users. This would, of course, be to accept a distinction in kind between the logical and the nonlogical vocabularies. (The semantics for the former would be a kind of ‘use’ theory, whereas the semantics

---

60 Field (1977) gives an inferential account of the meanings of logical terms. Similarly, Harman (1987) argues that the meanings of logical and mathematical words depends mostly on their roles in inference: given his semantic holism, he claims that links to perception also play some role.
for the latter would depend on nomic, specifically mind-world, relations.)

It may seem surprising that Fodor accepts a use-theory for logical vocabulary, since he is one of the fiercest opponents of use-theories in general—at least for non-logical words (Fodor 1987: 71-84). However, Fodor’s account of meaning—for non-logical words—is far from moderate. On the contrary, he proposes a radical atomism, according to which the meanings of all—non-logical—words depend exclusively on their links to the world (Fodor 1987 & 1994). Both a basic-acceptance semantics (like UTM) and Devitt’s moderate truth-referentialism have a clear advantage on this regard.

A general theory of meaning has to account for all words, presumably explaining what is the common property in virtue of which they all have meanings. While a basic-acceptance theory has a clear advantage over all of its rivals on this regard, I suggest that a modification of moderate truth-referentialism can also achieve the required generality: it can treat logical words as auxiliaries whose meanings also depend on properties that contribute to the truth-conditions of sentences, explained in terms of the links of logical words to rules of inference. Moderate truth-referentialism can hold the following:

The meanings of all words depend on properties that determine the truth-conditions of sentences: primitive words get their meanings from links to the world that determine their references, non-primitive words get their meanings from links to other words that determine their references, and logical words get their meanings from links to rules of inference that determine their contributions to the truth-conditions of sentences.
I conclude that the generality consideration favors a basic-acceptance theory like UTM over the vast majority of rival theories of meaning, but a moderate truth-referentialism—expanded to include links to rules of inference—matches the basic-acceptance theory in generality. While generality alone does not favor a basic-acceptance theory over a moderate truth-referentialism, generality is a crucial desideratum. A plausible unified reductive account of meaning will have to be along the lines of theories like basic-acceptance semantics (e.g., UTM) or an expanded moderate truth-referentialism, which can accommodate all kinds of words.

In the next section, I present the second consideration offered by Horwich in favor of UTM. I shall argue that this consideration does favor a basic-acceptance approach over all of its rivals. Still, I shall argue that a moderate truth-referentialism can again match a deflationary basic-acceptance approach if expanded further to include basic-acceptance. The resulting version of moderate truth-referentialism is, of course, not a rival to basic-acceptance semantics, because it includes basic acceptance, but it is a rival to UTM. After arguing for the core claims of UTM—i.e., basic-acceptance semantics—, I shall discuss the main problems faced by UTM. These problems do not stem form basic-acceptance semantics per se, but from UTM’s commitments to deflationism and dispositionalism.

3.5.3 The Explanatory Power of a Basic-Acceptance Account

The feature of a basic-acceptance semantics—like UTM—that truly favors it over non-acceptance-based rivals is its explanatory power. Basic-acceptance semantics offers a plausible explanation of how meanings can govern use, since it can appeal to the inferential model to explain how the total use of a word is derived from a fundamental acceptance property. In
contrast, Horwich (1998a: 47) argues, theories according to which the meanings of words do not depend on the acceptance of some sentences containing them cannot explain how these meanings govern the use of words in theoretical and practical inference:

One of the properties of meaning that we recognize pretheoretically is that what people say is due, in part, to what they mean. [...] And this explanatory feature of meaning is immediately accounted for by the use theory. For the central component of that theory is that the property which constitutes a word’s having the meaning it does is that its use is governed by a certain explanatorily fundamental acceptance property. And it is indeed quite clear [...] how the total use of a word might be derived, in light of circumstantial factors, from a basic ‘law’ of use—whereas it is relatively unclear how any other sort of property of a word [...] would constrain its overall use.

Let me unfold the points Horwich makes in this passage step by step.

_Pretheoretically_, the meaning of a word seems to be what governs its use, since people use words according to what meanings they attach to them. (Notice that this crucial point is in tune with the identification of meanings and the definition of the task of semantics I defended in Chapter 2.) If someone utters ‘Bachelors are unhappy’ or has a token of this sentence in his “belief box”, this is surely due _in part_ to what the word ‘bachelor’ means (1998a: 47). Additionally, a central _theoretical_ motivation for attributing meanings and, consequently, for articulating semantic theories is precisely our need to explain people’s behavior, which includes their linguistic behavior (Devitt 1996 & 2002). So a _fundamental desideratum_ for a theory of
meaning is that *it must explain how people use words*. (This again follows for the methodological considerations discussed in Chapter 2.)

Before I elaborate on how a basic-acceptance account (e.g., UTM) can explain word deployment, let me point out that standard use theories face a serious circularity problem if they attempt to explain word usage in terms of meaning, since they claim that meaning is use. Use theories that identify meanings with the *overall use* of words seem to have things upside down: they attempt to explain meaning in terms of use, when what we need instead is to explain use in terms of meaning. Basic-acceptance semantics (e.g., UTM), however, only claims that some *core uses* determine meaning: it does not identify the meaning of a word with its overall use. Furthermore, UTM claims that the core use of a word—i.e., its basic acceptance property—explains its overall use. So UTM is different from other use theories in not having things upside down. The explanatory power of this use theory is precisely given by the fact that it offers a plausible explanation of how meanings can govern use.

Horwich claims that the total use of a word derives from a fundamental acceptance property (Horwich 1998a: 47). How does it derive? The explanation appeals to the inferential model, which is the only working model of thinking we have: “no other model has been suggested for how we might explain what needs to be explained.” (Horwich 2005: 44). Given the deployment of rules of inference, the acceptance of some w-sentences—sentences containing a word w—can bring about the acceptance of other w-sentences, so some uses of a word can be derived from—and consequently explained by—others. While most uses of a word are surely the result of inferences from other w-sentences, these causal/inferential chain must start somewhere. There must be a core set of accepted w-sentences that serve as the inferential basis for other w-
sentences\textsuperscript{61} but are not themselves inferentially derived from other w-sentences. These core sentences govern the overall use of the word because they function as fundamental premises or “axioms” in the system of inferences containing the word. But if this is the case, these core sentences must be—at least part of—what the meaning of the words depends on, since they play a substantial causal role in the explanation the word’s overall use. As long as our theoretical motivations to attribute meanings are related to the need to explain the overall deployment of words, basic-acceptance semantics offers indeed a plausible account of the nature of meaning.

Notice that basic-acceptance semantics can also explain how mental content governs non-linguistic behavior, since the fundamental acceptance properties of a word can as well govern behavior through practical inference. Basic-acceptance semantics—e.g., UTM—exploits the inferential model without identifying meaning with overall inferential roles, so it preserves the pre-theoretical intuition that meaning determines use. Also, it relies on the inferential model without restricting meanings to inferential roles, since fundamental acceptance properties can link the meanings of words to the external world. The various uses of ‘cat’ in inferences leading to intentional behavior, for example, may be explained by the acceptance of ‘This is a cat’ in the presence of cats. Links to the world play a crucial role in the explanation of the use of many words and the intentional behavior resulting from practical inferences containing them.

Horwich claims that “it is relatively unclear how any other sort of property of a word [...] would constrain its overall use” (Horwich 1998a: 47). In other words, he claims that a basic-acceptance approach offers the best explanation. I take it that non-use-theoretical approaches must anyway incorporate a use theory to explain the use of logical and mathematical words (see

\textsuperscript{61} Together with other factors, such as the meanings of other words in those derived w-sentences, inputs from perception, etc.
Section 3.5.2). So I will focus on words whose meaning plausibly depends on links to some other words or to the world. The description theory plausibly claims that the meaning of ‘bachelor’ depends on its association with ‘unmarried man’. Indicator, causal-historical and (standard) teleosemantic theories plausibly claim that the meaning of ‘cat’ depends on its links to cats. Basic-acceptance semantics (e.g., UTM) includes these very same links to explain the meanings of ‘bachelor’ and ‘cat’, respectively. The crucial difference is that the above mentioned theories—or a moderate truth-referentialism that combines them—do not articulate these links as part of basic-acceptance properties. Can the links explain by themselves the overall use of these words, independently of the words having the links while embedded in some basic sentences? Arguably not. Most uses of these words in theoretical and practical inferences—in thinking—must result from previous sentences containing them, since only sentences can function as premises in inferences. I take this to be rather uncontroversial.62 Those who endorse non-acceptance-based theories can go along with this and claim that the word simply preserves its meaning in these inferential processes. But the widely accepted inferential model of thinking does impose a constraint in what properties can conceivably govern the use of words—and, consequently, explain their meanings—, since when we get to the original uses of the words in the causal/inferential chain, they must also be uses in sentences for the very same reason: isolated words cannot function as premises. The acceptance of ‘Bachelors are unmarried men’ associates ‘bachelor’ with ‘unmarried men’, as the description theory plausibly suggests, but it also explains how the association can govern the use of the word in other sentences containing it. The acceptance of ‘This is a cat’ in the presence of cats links ‘cat’ specifically to cats in the external world, as indicator, causal-historical and teleosemantic theories plausibly suggest, but it also can

---

62 At least for anyone who accepts the language of thought hypothesis and a “computational” theory of the mind.
serve as a premise in inferences containing ‘cat’, thereby governing other stimulus-independent uses of the word. So links to other words and to the world can indeed govern a word’s overall use, provided they are links the word has when embedded in some sentences containing it.

While UTM incorporates causal connections to the world in its account of the meaning of some words, Horwich denies that they thereby determine the reference of such words. Interestingly, however, Horwich (2005: 44) points out that a non-deflationary version of the basic-acceptance approach can be articulated, by postulating basic-acceptance properties that substantially determine the referential properties of words. Given the generality consideration, a plausible non-deflationary version of basic-acceptance semantics should be combined with a moderate truth-referentialism. Such a version of a basic-acceptance semantics would have the following shape:

i. *Primitive words* get their meanings from the acceptance of basic sentences with links to the world that determine their references and govern their use (e.g., the acceptance of ‘This is a cat’ in the presence of cats determines the reference of ‘cat’ and governs its use).

ii. *Non-primitive words* get their meanings from the acceptance of basic sentences that link them to other words that determine their references and govern their use (e.g., the acceptance of ‘Bachelors are unmarried men’ determines the reference of ‘bachelor’ and governs its use).

iii. *Logical words* get their meanings from the acceptance of basic sentences that link them to truth-preserving inference rules and govern their use (i.e., the acceptance of the schema
‘p, q \& p and q’ determines the truth-conditional meaning of ‘and’ and governs its use). This version of moderate truth-referentialism clearly matches Horwich’s deflationary UTM in explanatory power. But arguably it does so thanks to incorporating basic-acceptance semantics.

The moral we should draw is that the “explanatory power” consideration strongly favors a basic-acceptance account of meaning over all of its non-acceptance-based rivals. If deflationism turns out to be right, we need a deflationary basis-acceptance account. If deflationism turns out to be wrong, we need a moderately truth-referentialist basic-acceptance semantics. Basic-acceptance semantics is strongly motivated independently of the fate of deflationism.

In Section 3.5.1, I presented Horwich’s UTM and argued that it has crucial advantages over other versions of FRS. In Section 3.5.2, I argued that basic-acceptance semantics—the core set of claims of UTM—has a crucial advantage over almost every other theory of meaning in terms of generality, except an expanded moderate truth-referentialism that includes links to rules of inference for logical words. Basic-acceptance semantics and moderate truth-referentialism are the only plausible candidates for a general theory that accounts for the meanings of all words. In this section, 3.5.3, I argued that basic-acceptance semantics has a crucial advantage over every other theory of meaning in terms of explanatory power: the meanings of words must depend on basic-sentences containing them in order to govern their use. I also argued that this argument is independent of deflationism, so that it does not favor a deflationary basic-acceptance semantics—like Horwich’s UTM—over a moderately truth-referentialist basic-acceptance semantics.

---

63 This sketch is inspired on one offered by Horwich (2005: 44-45).
64 At least if we do not want to “abandon meaning altogether” (Devitt 2002: 106).
I have so far abstracted away from the problems faced by Horwich’s UTM and focused on the virtues and insights of a basic-acceptance semantics. However, Horwich’s UTM does face some serious challenges. I shall discuss these challenges in the following sections. Based on my assessment, I shall argue that basic-acceptance semantics must be combined with teleosemantics and that the result may turn out to be deflationary or moderately truth-referentialist.

3.6 Problems with Horwich’s Use-Theory

While basic-acceptance semantics is strongly motivated by its generality and explanatory power, Horwich’s UTM faces two main problems pointed out by Devitt (2002 & 2011): (1) UTM risks collapsing into truth-referentialism, and (2) UTM is undermined by problems of ignorance and error. I shall argue that neither of these problems stems from basic-acceptance semantics. The first problem arises from UTM’s commitment to deflationism. Given the risk of a collapse, I shall argue that basic-acceptance semantics should remain agnostic or neutral about deflationism until we determine whether it effectively collapses or not into a form of truth-referentialism. The second problem arises from UTM’s commitment to a dispositionalist account of the functions of words. I shall argue that UTM is indeed undermined by this problem and that basic-acceptance semantics must be combined with a different account of the functions of words in order to offer a plausible explanation of why words mean what they do.

Before discussing the main problems faced by UTM, let me make a brief point about Horwich’s take on the relation between the meanings of linguistic and mental representations. Horwich treats language and thought as a “seamless whole”. As far as he is concerned, his theory is about the words used by people both to think and to communicate with each other. As he puts
it: “I favour a uniform account, which will deal in the same way with both overt and mental terms” (Horwich 2005: 31). In a critical assessment of UTM, however, Devitt reasonably argues that UTM should give instead priority to the mental. Devitt’s main point is that UTM’s “linguistic theory which seeks to identify and then explain the meanings of linguistic words depends on the prior identification and explanation of the meanings of mental words”, and that this is a consequence of UTM’s “reliance on accepting a sentence” (Devitt 2002: 109). Recall that, for Horwich, accepting a sentence involves regarding it as true and relying on it as a premise in inferences, which includes not only asserted sentences, but also sentences deployed in thought—i.e., sentences in the speaker’s “belief-box”. The point is that “accepting a sentence” is fundamentally a psychological notion. Inferences take place in the mind. In order to rely on a sentence as a premise in theoretical and practical inferences, the sentence must be tokened in though—i.e., stored in the “belief-box”. This alone should suffice to show that UTM should give priority to the mental.

Suppose that Mary believes that bachelors are unhappy and utters ‘Bachelors are unhappy’. What explains Mary’s linguistic behavior? The best explanation we have—I am not sure there even is another plausible one—is the standard one based on Mary’s thoughts: she believes that bachelors are unhappy, she wants to communicate her belief, therefore she utters a sentence that has the same meaning as her belief according to the linguistic conventions of her community (Devitt 2002: 110). This sort of explanation is based on our old and very successful folk-psychology. The idea behind it is reflected in the saying that language expresses thought (Devitt 2002: 111). It should be obvious that there is a lot to say in favor of this view and almost nothing to say against it. But the view in effect assigns explanatory priority to the mental. Let us
consider now Horwich’s explanation of the overall use of a word in terms of basic-acceptance. To follow the same example, UTM would explain Mary’s uses of ‘bachelor’ in her utterance and belief by her acceptance of ‘The bachelors are the unmarried men’. Should a proper version of UTM deal “in the same way with both overt and mental terms”? The first point to notice is that the basic sentence will have to be tokened in the “belief-box” in order to govern inferences with the word by serving as a premise. The second point to notice is that, given the explanation of Mary’s utterance, the basic sentence will only cause the utterance by first causing the belief. UTM can indeed explain both overt and mental uses, but the explanations are not symmetrical: it can explain the overt use by explaining the mental one. So I conclude that UTM indeed should give priority to the mental. Of course, for many explanatory purposes it is convenient to ignore the complex relations between language and thought treating them simply as a “seamless whole”. But the point remains that in an ultimate analysis thought takes priority.

3.6.1 Risk of Collapse

Devitt points out that both UTM and a moderate truth-referentialism account for the meanings of non-primitives such as ‘bachelor’ in terms of their associations with other words. For these terms, Devitt argues, the only difference between the two competing theories is in their accounts of the nature of reference: “truth referentialism is committed to a substantial reference relation, the use theory, to deflationary reference” (Devitt 2002: 114). Therefore, relevant evidence in favor of one theory of meaning over the other has to be found elsewhere: in their accounts of the meanings of primitives.

Moderate truth referentialism explains the meanings of primitives in terms of their
reference determining causal relations to reality. The use theory has to choose between also explaining the meanings of these words in terms of their causal relations to reality or explaining them instead in purely intra-linguistic terms. As Devitt points out, well-known arguments for externalism count against the second alternative (Devitt 2002: 115). A use theory that follows this path is an **extremely implausible** theory of meaning. To be plausible, the use theory must explain the meanings of primitives in terms of their causal relations to reality. But then, Devitt argues, the theory faces the following challenge:

[It] needs to distinguish itself from truth referentialism. Truth referentialism holds that these meaning-constituting causal links explain reference. The use theory needs to say why they don’t (because the use theory holds that reference is deflationary and not open to this sort of explanation). (Devitt 2002: 114-115)

It is important to notice that this challenge is faced specifically by Horwich’s deflationary UTM, rather than by basic-acceptance semantics in general. As I argued above (Section 3.5.3), basic-acceptance semantics is motivated independently of deflationism—and may even take a truth-referentialist form—so its fate does not go hand in hand with the fate of deflationism.

How can UTM explain why all of these causal links are not reference determining ones? Horwich may claim that the causal links only determine meaning and that ascriptions of referents to these words are derived trivially from the application of the deflationary reference schema (*see* Chapter 2, Section 2.2.3). But the problem is that deflationism entails that reference cannot play any explanatory role in meaning. Why should we expect the meanings of primitives to be constituted exactly by the same causal links that truth referentialism regards as reference
determining if reference does not play any explanatory role in the account? Truth referentialism has good reasons to hold that we need these causal links: we need them to explain reference.

In his reply to Devitt, Horwich makes two points. I already presented the first point above (Section 3.5.3): Horwich suggests that a non-deflationary version of the basic-acceptance approach can be articulated, by postulating basic-acceptance properties that substantially determine the referential properties of words (Horwich 2005: 44). This of course does not prevent UTM from collapsing into truth-referentialism. Horwich’s point is merely that if there were a collapse, it would not undermine basic-acceptance semantics per se. Horwich’s second point is intended to defend specifically his deflationary UTM. He argues that we should not expect basic-acceptance semantics to take the shape required by moderate truth referentialism, since the plausibility of deflationism shows that there is no reason to assume that meaning properties must constitute the “is true of” relation (Horwich 2005: 45). While Horwich agrees that some words may turn out to have relational acceptance properties, he insists that the use theory does not require it:

[The use theory] leaves room for ‘physical’ externalism, insofar as it does not preclude the possibility that the conditions of acceptance involved in certain meaning-constituting properties will include aspects of the environment. But it is not committed to externalism, insofar as it allows that the best explanations of overall use may turn out never to call for that kind of basic acceptance property. (Horwich 2005: 49n.)

Devitt warned that if UTM treats the meanings of primitives as determined by causal relations to
reality, it *risks* collapsing into truth referentialism. Horwich replies that UTM is *not committed* to externalism and UTM’s explanations may turn out to *never* involve causal links to reality. If this happened to be the case, UTM would certainly not collapse. However, a purely internalist UTM would not only avoid a collapse: it would also be undermined by the powerful arguments for externalism, as Devitt originally warned (Devitt 2002: 115). So the sensible conclusion to draw is that UTM indeed *risks* collapsing into a form of truth-referentialism.

Given similar worries, Field suggests that we should merely adopt a *methodological deflationism* in the theory of meaning: “a methodological policy, which if pursued could lead to the discovery that deflationism in the original sense (‘metaphysical deflationism’) is workable or could lead to the discovery that inflationism is inevitable” (Field 2001: 119). We should leave open the possibility that a fully satisfactory version of basic-acceptance semantics may turn out to be one that complies with the requirements of a moderate truth-referentialism—along the lines I sketched above (Section 3.5.3). I have to confess that, unlike Horwich, I am inclined to believe that this is a *likely* outcome.

### 3.6.2 Problems of Ignorance and Error

Another issue pointed out by Devitt (2002 & 2011) is that UTM faces serious problems of ignorance and error. These problems, I will argue, stem from UTM’s commitment to dispositionalism, not from basic-acceptance semantics itself.

Horwich relies on a dispositionalist model of inference to explain the overall use of words, on one hand, and the nature of basic acceptance properties, on the other. Accordingly, Horwich regards the overall use of words as a certain kind of regularity, and the core uses that
constitute their meanings as an “explanatory basic regularity”. The meaning constituting properties of words are, according to Horwich, the laws that govern their use. These laws are articulated as basic acceptance properties which, in turn, are dispositions or tendencies to accept certain sentences in certain conditions. Certain sentences have to be regularly accepted in certain conditions in order to constitute an explanatory basic use. Even Horwich’s account of the division of linguistic labor is purely dispositionalist. A speaker qualifies as sharing the public meaning of a word, even when he is too ignorant or makes too many mistakes, as long as he is disposed to defer to the experts in the use of that word in his linguistic community.

Horwich’s dispositionalist accounts of fundamental acceptance properties and of the division of linguistic labor face serious problems of ignorance and error. As Kripke (1982) has shown, any dispositionalist theory of meaning will fail to properly characterize incorrect uses as incorrect, because speakers/thinkers often have dispositions to make mistakes and often lack the right dispositions. A purely dispositionalist theory does not have the resources to exclude the dispositions to make mistakes from being meaning constitutive. So it is no surprise that Horwich’s use theory, as it stands, fails to offer a plausible account of meaning. I will argue that Horwich’s dispositionalist account of fundamental acceptance properties indeed leads to a serious disjunction problem when applied to relational meanings. On one hand, Horwich claims that the problem of error is a pseudo-problem. But I will reply that this position leads to implausible attributions of meaning. On the other hand, Horwich offers an account of the division of linguistic labor that anyway may help to overcome the problem. But I will argue that the problem persists for mental words that are not part of a public language and that the proposed solution even fails for words that are indeed words of a public language since, as Devitt (2002)
shows, Horwich’s theory of linguistic deference is itself undermined by serious problems of ignorance and error.

Let us suppose that UTM’s hypothesis for the meaning of ‘horse’ is that it constituted by our basic tendency to accept ‘This is a horse’ in the presence of horses. However, surely we have a tendency to accept ‘This is a horse’ in the presence of not only horses, but also muddy zebras, the odd cow, etc. But then this is the actual law of use of ‘horse’, with the result that applying ‘horse’ to muddy zebras, the odd cow, etc., does not constitute a misuse of the word. Horwich, however, claims that the meaning constituting laws of use are idealized laws. So his theory does make some room for error. If a speaker has a tendency to only occasionally accept ‘This is a horse’ in the presence of muddy zebras and odd cows, but a regular tendency to accept it in the presence of horses, then the use theory can characterize his misuses as misuses: simplicity considerations will determine that the relevant idealized law is only the one that involves horses.

However, idealized laws of use will successfully abstract away from errors only when most actual uses of the word are correct. The dispositionalist use theory still requires meaning constituting sentences to be regularly accepted. But this makes too little room for error. The problem is that errors may often be more frequent than the use theory allows. In these cases, UTM will attribute the wrong meaning to the word. In the ‘horse’ example, it will attribute a disjunctive meaning-constituting relational property to speakers that fail more often than they succeed in identifying real horses. Or consider Putnam’s (1975) ‘beech’ and ‘elm’ example. In this case, plain ignorance of how to apply the terms (rather than error committed during their application) may lead to consistent misapplication of the words by many speakers who share Putnam’s inability to distinguish beeches from elms.
Horwich argues that the problem of error is a pseudo-problem. I will get back to this below. But he also offers an account of the division of linguistic labor that is intended to make room for people misusing words without changing their meanings:

In order for an individual member of the community to mean a certain thing by a given word, it is not necessary that he himself uses it precisely in accordance with the regularity that fixes the meaning of the word type. What is needed is, first, that there are acknowledged experts in the deployment of the term—experts whose usage is determined by some such regularity; second, that the individual is disposed to defer to the experts—i.e. to accept correction by them; and consequently, third, that his use of the term conforms to that regularity at least to some extent. In these circumstances, even when the speaker’s use of a word is fundamentally abnormal, we none the less attribute the normal meaning to him; and that normal meaning is constituted by the regularity that explains the overall use of the word by those ‘specialists’ to whom the rest of us are prepared to defer. (Horwich 1998a: 86)

There are two things to be said against this. In the first place, the proposed solution only is available for abnormal uses of words with a public usage. So the problem of error persists for words that are not part of a public language—e.g., mental words coined by an individual thinker or written words coined for someone’s private use. In the second place, even for words with a public usage, the solution relies on the disposition of users to defer to experts. The original problem of error was related to the theory’s dispositionalist account of basic acceptance properties. But, as Devitt (2002) demonstrates, a dispositionalist account of the division of
linguistic labor is doomed to face analogous problems.

I will elaborate the first problem with a simple thought experiment. Suppose that Robinson Crusoe, alone on his island, coins a new word to name a small long-eared species of rodent he has encountered. When speaking to himself and writing in his diary he calls these creatures ‘gavagais’. One day he manages to construct a precarious boat to escape from his solitary confinement. In sailing out to sea, however, he finds himself in a storm that unfortunately brings him to another uninhabited island. On this new island, Robinson Crusoe finds a small mammal that looks exactly like a gavagai, so he thinks these creatures are gavagais and the new entries on his diary also allude to these creatures as gavagais. However, unknown to him, the creatures on the new island are not really gavagais. They are not even rodents, but marsupials that have evolved into a very similar shape to that of gavagais because of their similar ecological niches.

In a case like this, how can the use theory explain error and misrepresentation? It cannot appeal to semantic deference, since this is a word coined by a solitary speaker. Robinson Crusoe is the only “expert” in the use of the term. The use theory cannot avoid the conclusion that the meaning of Robinson Crusoe’s ‘gavagai’ is constituted by his tendency to accept ‘This is a gavagai’ in the presence of either gavagais or their marsupial lookalikes. But surely Robinson Crusoe is misrepresenting the small marsupials as being gavagais. Since the misrepresentation is systematic, the use theory cannot abstract away from it appealing to the ideal law of use for the word. Since the word is not a word of a public language, the use theory cannot appeal to semantic deference to account for the proper meaning of the word. The conclusion is that a dispositional characterization of basic acceptance properties fails to provide a plausible account
of meaning.

The second problem with Horwich’s proposed solution is that his dispositionalist account of semantic deference faces itself serious problems of error. Recall that Horwich’s theory of deference imposes three requirements for a speaker who uses a word abnormally to qualify as using the word with its public meaning: first, there must be experts on the use of the word; second, the speaker must be disposed to defer to the experts; third, the speaker use must conform to the public use to some extent. As Devitt (2002) points out, these requirements can hardly be met:

Consider proper names. To meet Horwich’s second requirement, the user of a name has to acknowledge her ignorance and hence be prepared to defer. But surely many ignorant users of a name do not acknowledge their ignorance. On Horwich’s theory these users will not be using the name with the same meaning as the deferrers and the experts. My guess is that most users of most names will be in that category. Next, the requirement demands that each deferrer identify experts to defer to. How? They cannot be identified simply as experts on the meaning of the name, on pain of circularity. It is surely unlikely that most deferrers will be able to manage the required identification [...]. Finally, Horwich’s first requirement is that there be acknowledged experts. But surely the ignorant will often defer to someone that they think is an expert who in fact is not. In sum, people will often not defer where they should; they will often try to defer but fail; they will often refer to a nonexpert. (Devitt 2002: 118-119)
The account of deference fails for exactly the same reasons as the account of basic acceptance properties. People often have dispositions to make mistakes (error) and often lack the required dispositions (ignorance). In the case of deference, as Devitt points out, people may have dispositions to make mistakes in whom to defer to, in when to defer, and people may lack the disposition to defer to anyone whatsoever. Again, these failures may often be regular, so appealing to regularities or laws of usage does not solve the problem.

Horwich has developed an ingenious reply to the objection that UTM cannot overcome the problem of error. In a nutshell, he argues that the “problem of error” is a pseudo-problem that results from an unwarranted inflationism about truth and reference. Given inflationary assumptions, he claims, we wrongly expect to be able to “read off” the referent of a word from its meaning constitutive property. But deflationism frees us from the requirement that a meaning constitutive property should substantially determine the reference of a word. According to deflationism, the reference of a word is instead trivially determined. So, Horwich concludes, there is no need to impose on the theory of meaning any substantial requirements related to how the correct reference of a word is determined by its meaning. (See Horwich 1998a: 65-71 & 2005: 63-84.)

I have two points to make against Horwich’s attempt to disregard the problem as a pseudo-problem. First, it seems to me that Horwich’s “dissolution” fails because it would undermine his own reasonable hypotheses about words’ meanings. Consider the hypothesis that the meaning of horse is engendered by the acceptance of ‘This is a horse’ in the presence of horses. This is similar to many of the reasonable hypotheses offered by Horwich himself. But in order to really embrace the view that there is no problem of error, rather than just flirt with the
idea, Horwich must be willing to bite the bullet and say that the meaning of ‘horse’ is engendered by the acceptance of ‘This is a horse’ in the presence of horse-looking things; so that the word would be correctly used when applied to muddy zebras or odd cows.

The second problem I find with Horwich’s alleged dissolution of the problem of error is that I do not think that the problem is merely a truth-referentialist one. While the problem has often been articulated in truth-referentialist terms, the underlying problem with dispositionalism is deeper and occurs even outside of the theory of meaning. As teleological theorists have pointed out, a dispositionalist theory of biological function is in quite an analogous situation to a dispositionalist theory of meaning (I elaborate more on this on Chapter 4). The function of sperm, for instance, seems to be to fertilize eggs. However, most sperm never performs this function. And similar problems arise for the functions of biological organs, innate behaviors, acquired behaviors, etc. Biological devices often have dispositions to malfunction and even malfunction regularly. So a purely dispositionalist/regularist account of functions is unable to characterize malfunctioning as malfunctioning.

The problems of malfunctioning with dispositionalist theories of biological function are clearly not based on any inflationary assumptions. But they will also arise for a functionalism about content that adopts a similar dispositionalist account of the functions of symbols. Consequently, I think that deflationism does not truly remove the underlying problem. Even a deflationary UTM should give a proper account of malfunctioning in the deployment of words and concepts. But as long as UTM is committed to pure dispositionalism, it is not clear how it can offer such an account.
There is a use-theorist that does take seriously the problem of dispositionalism being unable to underscore a proper distinction between correct and incorrect uses. Robert Brandom (1994 & 2000) argues for a form of use theory that he calls “inferentialism” and that, notably, is based on norms rather than dispositions or regularities of use. Brandom’s elaborate theory is non-naturalist, since it relies on irreducible normativity. This alone makes it unattractive, given the arguments for a naturalized semantics I have given in Chapter 1. But notice that someone with different views on that issue may suggest that Horwich should similarly adopt a norm-based account of use. Meanings would be engendered according to such a theory by basic norms of acceptance, instead of basic acceptance regularities. The meaning of ‘horse’, it could be said, is engendered by a norm like: You ought to accept ‘This is a horse’ in the presence of horses.

The alternative solution, which I suggested earlier, is to combine basic-acceptance with a teleological account of the functions of symbols. This combination allows the basic-acceptance approach to make plenty of room for ignorance and error while remaining within the sane confines of naturalism. The meaning of ‘horse’, for example, may be said to be engendered by the fact that the acceptance of ‘This is a horse’ in response to the perception of horses is supposed to govern its use, while the meaning of ‘cardiologist’ may be said to be constituted by the fact that its use is supposed to be governed by the acceptance of ‘Cardiologists are doctors specialized heart diseases’—where the relevant “supposed” is not a prescriptive one, but the kind of “supposed” involved in the natural fact that kidneys are supposed to filter blood. This requires the notion of “function” that applies to biological organs to be “normative” in this non-prescriptive sense and representations (including acquired ones) to have these same sorts of functions. I will discuss these issues in the following chapters.
3.7 Conclusion

In this chapter I have briefly surveyed the range of available theories of meaning, suggested an approach that combines Horwich’s basic-acceptance semantics with teleosemantics, and elaborated on the virtues and problems of UTM: Horwich’s version of basic-acceptance semantics. I argued that the generality and mainly the explanatory power of basic-acceptance semantics favors it over other alternatives, although I pointed out that a basic-acceptance semantics with these advantageous features can take a truth-referentialist form. I also argued, following Devitt, that UTM risks collapsing into truth-referentialism and fails to give a proper account of relational meanings, because it is undermined by serious problems of ignorance and error. I have suggested that the risk of collapse into truth-referentialism does not undermine basic-acceptance semantics. If basic-acceptance semantics fails to give a plausible account of meaning given the problems of ignorance and error of UTM, then we either have to reevaluate the inference to the best explanation based on the explanatory power consideration or somehow fix the theory to avoid the problems of ignorance and error. Since I think that the explanatory power consideration favors the use theory over any theory not based on basic sentence acceptance properties, I will explore the second alternative: to fix the use theory to avoid the problems of ignorance and error. Overcoming the problems of ignorance and error is the main virtue of the combination of basic-acceptance and teleosemantics. In the next chapter, I will introduce the idea of teleosemantics and analyze in detail the teleological notion of function.
Chapter 4:

Teleonomic Functions

4.1 Introduction

In Chapter 3, I argued in favor of a basic-acceptance account of meaning, but objected that Horwich’s articulation of such an account is undermined by serious problems of ignorance and error. My aim is to propose a teleological or teleonomic version of basic-acceptance semantics which promises to overcome such problems. There are various already existing “teleosemantic” theories that rely on teleology to offer accounts of meaning that avoid falling prey of problems of ignorance and error. The approach I will be proposing departs in crucial ways from existing theories, but also borrows elements from them. Some available teleosemantic theories are quite complex and the model of functions they rely on requires some inspection before delving into how it can illuminate the nature of meaning. In this chapter, I will briefly discuss the nature of teleosemantic theories and discuss in some detail the etiological account of functions that such theories rely on. I will argue that the main advantage of the etiological account is that it accounts for the “normative” character of functions within a fully naturalistic framework. This is the feature that makes etiological functions particularly attractive to semantics.

4.2 What is Teleosemantics?

The core feature of teleosemantic theories is the following: they rely on a teleological account of functions to explain the nature of meanings. But most of the available versions of teleosemantics share another feature: they rely on direct links to reality to explain the references and/or truth-
conditions of representations. In this section, I briefly present these points, which I discuss in more detail afterwards. I suggest that the only defining feature of teleosemantics is the first one. A theory of meaning can be teleosemantic even if it does not have the second feature.

The first common feature of available versions of teleosemantics is that they all attempt to explain the nature of meanings in terms of what symbols or representations are supposed to do or what is their purpose given the history of their use and/or the history of the underlying mechanisms responsible for their use. What they propose, first and foremost, is that the meaning or content of a symbol or representation depends on its etiological functions: the functions that explain why the symbol or representation exists. Since the historical functions of an item explain why the item exists in terms of its purposes, of what the item is for, the etiological account is often called “teleological”. This account of functions has its roots in biology and the philosophy of biology. A kidney, for example, has the function of filtering blood. A teleological account proposes that filtering blood is a purpose of kidneys (among others, such a producing certain hormones) because they have been naturally selected for doing this. One of the main virtues of this account is that it makes room for malfunctioning. A failing kidney can be said to have the function of filtering blood even if it is not currently performing it properly or at all, because the explanation of why it exists relies on how having kidneys benefited ancestor creatures—by filtering waste and excess fluids out of their blood—aiding their survival and reproduction, which resulted in the inheritance of the trait by their descendants. Notice that while “malfunctioning” or “failing” to perform a function are, in a qualified sense, “normative” notions, the etiological explanation is fully naturalistic. I discuss etiological functions in more detail in the following sections of this chapter.
A teleosemantic theory basically makes use of the etiological explanation of functions to account for the meanings of symbols or representations. If symbols or representations are supposed to be used in certain ways because they have etiological functions, then their misuses can be accounted for in fully naturalistic terms. If we strictly follow the biological account which relies on Darwinian natural selection operating over hereditary traits, a teleosemantic account may seem to be available only for innate symbols or representations. Some teleosemanticists do limit the scope of teleosemantics to a subset of symbols or representations that are innate or the result of purely innate mechanisms (Sterelny 1990 & Neander 1999). But other teleosemanticists have argued that the account can be extended to cover also symbols or representations that are the result of learning (Millikan 1984, Papineau 1984 & Dretske 1988) and even to the words of natural languages which are transmitted culturally rather than by biological reproduction (Millikan 1984). Whether the scope of a teleosemantic theory is limited or not to innately determined symbols or representations, what makes it a version of “teleosemantics” is that it relies on an etiological account of the functions of symbols to explain the nature of their meanings. This is the defining feature of teleosemantics.

The second common feature of most available versions of teleosemantics is that they resort to teleology specifically in order to explain the contents of representations in terms of direct links to reality that determine their references and/or truth-conditions. Like indicator or informational theories, causal-historical theories and description theories of reference—but unlike many “use” or functional-role theories of meaning—all available versions of

---

65 In the case of words and concepts, the direct links are alleged to determine their references—which in turn contribute to the truth-conditions (or satisfaction-conditions) of sentences and thoughts containing them. In the case of simple and non-compositional representations—such as those generated by frog’s bug-detectors (discussed in Chapter 5)—a teleosemantic theory is best characterized as providing direct links that determine truth-conditions.
teleosemantics are truth-referential. Indeed, teleosemantic theorists tend to simply identify the contents of representations with their references and/or truth-conditions. Like indicator and causal-historical theories, most available versions of teleosemantics rely on direct links to reality to explain representational content. The links are direct in contrast to the indirect links of description theories of reference—according to which the link between a symbol and reality is mediated by its links to other symbols (e.g., ‘bachelor’ is indirectly linked to bachelors through its direct links to ‘unmarried’ and ‘man’). In sum, available teleosemantic theories belong to the large set of truth-referential theories of meaning and, within this set, to the sub-set of theories that rely on direct links to the world. The crucial motivation for resorting to teleology in order to explain the nature of these direct links can be appreciated by comparing teleosemantic theories with indicator or informational theories—since the former are tailored to avoid one of the most serious problems faced by the latter.

According to indicator or informational theories (Stampe 1977, Dretske 1981 & Fodor 1987), the references and/or truth-conditions of symbols or representations are determined by their reliable causal links to reality. A mental token of ‘horse’, for example, is said to indicate or carry information about the presence of a horse because tokens of that type are reliably caused by the presence of horses. Indicator theories have the virtue of being fully naturalistic. But they are undermined by serious problems of misrepresentation. Due to ignorance or error, tokens of ‘horse’ are sometimes caused by zebras or distant cows, for example. The problem is that indicator theories do not have the resources to characterize these instances as genuine misrepresentations. Since ‘horse’ is reliably caused by horses, zebras and other horse-looking things, the pure indicator theory unfortunately entails that ‘horse’ refers to horses, zebras and
Available teleosemantic theories attempt to overcome the problem of misrepresentation by relying on a teleological characterization of the direct links between representations and reality. A simple option combines indication with teleology, claiming that a representation $R$ stands for $X$ if and only if it has the etiological function of indicating $X$ (Dretske 1988). A related informational alternative holds that $R$ stands for $X$ if and only if the mechanism that produces it has the etiological function of responding to $X$ (Neander 2013). These versions of teleosemantics may be regarded as teleological variants of indicator or informational theories. They rely on etiological functions to isolate the proper causes of representations—what conditions in the world they are supposed to be caused by—and are able to characterize tokens that are not properly caused as misrepresentations. A representation may have the function of indicating or being tokened in response to exclusively $X$ even if it fails to be reliably caused only by $X$. Other versions of teleosemantics depart further from indicator or informational theories, focusing on the effects rather than the causes of representations. They claim that a representation $R$ stands for $X$ if and only if it is supposed to be tokened in the presence of $X$—or when $X$ obtains—in order to bring about the beneficial effects it has the function of producing (Millikan 1984 & Papineau 1984). Notice that what matters here is that $R$ is tokened in the right circumstances, regardless of what causes such tokens. The circumstances may be the right ones even when $R$ is not caused by what it represents. These versions of teleosemantics rely on etiological functions to isolate under what conditions in the world representations are supposed to be tokened—and are able to characterize tokens that are not tokened in the proper conditions as misrepresentations. I will propose a hybrid between these two approaches. The hybrid account claims that a representation...
$R$ stands for $X$ if and only if it is supposed to be caused by $X$ in order to bring about the beneficial effects it has the etiological function of producing. I will argue that this combination overcomes some problems faced by teleosemantic theories exclusively based on either the causes or the effects of representations. I discuss these various teleosemantic approaches in Chapter 5.

All of these teleosemantic theories attempt to explain truth-referential content in terms of direct links between representations and reality. However, this common feature is not a defining characteristic of the teleosemantic approach itself. In principle, the etiological account of functions can be exploited to articulate teleosemantic variants of other kinds of theories of meaning. Consider a moderate truth-referentialism, according to which the meanings of primitive words/concepts depend on their reference-determining links to reality and the meanings of non-primitive words/concepts depend on their reference-determining links to other words/concepts (Devitt 1996). The place for teleosemantics within such an approach, it may be assumed, is among the candidate theories for the meanings only of primitives, while a description theory is expected to account for the meanings of non-primitives. However, a description theory can also be articulated in teleological terms—basically, characterizing the reference-determining links to other words/concepts of a non-primitive as those it is supposed to have in order to be used properly.66 Furthermore, I will suggest that teleology can be used to articulate a teleosemantic version of a “use”, functional-role or conceptual-role theory. More specifically, I will argue that a basic-acceptance semantics that relies on an etiological account of functions promises to overcome the serious problems of misrepresentation faced by Horwich’s basic-acceptance semantics.

66 Papineau (1987: 78-80 & 93) indeed offers a teleosemantic description theory for non-observational concepts, which supplements his (direct-links based) teleosemantic theory for observational concepts. The version of the description theory he adopts, however, is not the most plausible one (see Devitt 1991: 431-432). But the problems do not specifically stem from its teleosemantic character. I will get back to this in Chapter 6.
theory. I discuss these matters and propose a combination of basic-acceptance semantics with teleosemantics in Chapter 6.

Before delving into the functions of symbols or representations, discussing different teleosemantic approaches and articulating a teleosemantic version of basic-acceptance semantics, I will discuss in more detail the etiological account of biological functions that serves as the basis for teleosemantics. While the etiological account of functions discussed below can and has been applied beyond the realm of the strictly biological—for example, to the realm of “cultural evolution”—its application to strictly biological phenomena is largely responsible for its naturalistic credentials which, in turn, are responsible for much of the philosophical interest in applying it beyond the strictly biological. The focus of the remainder of this chapter is on biological functions.

4.3 The Etiological Account of Biological Functions

The etiological account of biological functions is usually called “teleological” because it ascribes “purposes” to biological items. But this must be qualified, since it departs in crucial ways from traditional forms of teleological thinking. In the Aristotelian tradition, purposes or goals were understood as “final causes”—roughly, the counterpart of $X$ being the efficient cause of $Y$ would be $Y$ being the final cause of $X$—which were postulated not only to explain why seeds grow into plants, but also why stones fall to the ground. The etiological account of functions does not rely on this rightly discredited notion of “final causes”. Another traditional form of teleological thinking—which is still alive in folkloric opinion—relates the purposes of biological items to their alleged “design” by an intelligent, supernatural, creator. This view is at odds with a
naturalistic view of the world. In contrast, the contemporary form of teleological explanation offers a naturalistic account of the “purposes” of biological items that relies on the brute, blind, and ultimately physical forces governing biological evolution.

While the etiological account is the mainstream view about the nature of functions in the philosophy of biology, it is not the only naturalist account available. Its main alternative is the account of functions in terms of the causal roles played by items within the larger systems to which they belong. The etiological and the causal-role accounts of functions are often regarded as rivals. But they both may be appropriate as accounts of different phenomena. However, in the case of specifically biological functions the etiological account has a clear advantage over the causal-role account: it explains how things can have a function even when they fail to perform it. I contrast the etiological and the causal-role accounts of functions in Section 4.7 of this chapter.

Eyes are for seeing, hands are for grabbing, wings are for flying, ears are for hearing, and so forth. The purpose of hearts is to pump blood, the purpose of kidneys is to filter blood, the purpose of calluses is to protect skin from friction, and so forth. In some sense of the words ‘for’ and ‘purpose’ these are all uncontroversial claims. But the characterization of biological functions in these teleological terms has to be taken with a grain of salt. Ordinary people often describe functions teleologically, but they also often think of such functions as intended or deliberate purposes: the kinds of purposes conferred to artifacts by their intelligent designers and users, which biological items would also have if they were the products of intelligent design. The question is how can a naturalistically respectable “teleological” account of functions be offered for biological items, since scientific biology has shown that they are the result of non-intelligent forces. As Peter Godfrey-Smith points out:
Teleology has been a huge topic in the philosophy of biology; here I mean the status of our tendency to treat biological activities in terms of purposes, goals, and proper functions. It is usually assumed that the intentions of an intelligent designer or user of an object can be the basis for teleological description in a straightforward way. The problem is whether and how these terms can be used in the absence of this overt role for intelligence. (Godfrey-Smith 2009: 12)

Let us consider for a moment the functions of artifacts. Pens, for example, are for writing. That is their purpose because they have been intentionally designed for it. Now consider a stone used as a paperweight. If it was polished to be comfortable to handle and painted over to be aesthetically appealing, it has been intentionally designed to be a paperweight and a piece of decoration. Alternatively, the stone may have been chosen or “selected” by someone because it was already smooth and simply used as a paperweight without modifying it. It anyway has the function of a paperweight—i.e., preventing paper from being blown away by wind—because of the use it has been deliberately selected for and is being put to. However, stones qua stones have no purposes. It would be a serious mistake to ascribe purposes to them, “projecting” into the world the kinds of purposes that humans deliberately confer to artifacts. Unlike stones, however, wings, eyes, hearts and kidneys do have “purposes”—in some sense of the word—_independently_ of the “intentions of an intelligent designer or user” (to use Godfrey-Smith’s phrase). How can this be?

The contemporary account of teleology in the philosophy of biology regards evolution by _natural selection_ as the ultimate source of “purposiveness”. Charles Darwin discovered in the 19th century that the complex phenotypical traits of organisms are the result of a process involving inheritance, modification and differential survival and reproduction. Organisms inherit
characteristics from their ancestors, but they are not perfect copies or reproductions, since there are occasional random mutations. These modifications sometimes increase or decrease a creature’s chances of surviving before it reproduces. Consequently, favorable modifications tend to be maintained in—and unfavorable modifications tend to be filtered out from—populations of reproducing organisms over time. As Darwin pointed out, the process is analogous to that of the artificial selection of traits made by human breeders of domesticated animals and plants. The difference is that the natural history of organisms involves no deliberate selection whatsoever. Nonetheless, some variations are favored and others are filtered out as a matter of brute fact. This is what Darwin called “natural selection”. Intelligence plays no role in the process, but the results of the process often look as if they had been deliberately designed. We may say that biological traits are “designed” by natural selection, as long as we keep in mind that this manner of speaking, while useful in some contexts, is either merely figurative or has a technical meaning that departs from the ordinary usage of ‘design’.

The process of evolution by natural selection explains how organisms and their traits descend from common single-celled ancestors which did not have most of those traits. It explains why creatures have eyes, hearts, kidneys, wings, hands, and so forth. The standard evolutionary explanation is gradualist. Eyes are complex structures which are the result of cumulative small changes—each of which was favorable—starting with simple mutations that made some skin cells light-sensitive. There may also be non-gradual changes or “saltations”. But notice that the selective process does not stop working once an organ has “fully” evolved. Even if the eyes of many species have not changed in millions of years, their maintenance in the reproducing population still requires explanation. Random mutations keep occurring and traits that are not
continuously favored by natural selection tend to turn into vestigial traits and disappear over time. This means that the maintenance over time and the current existence of a complex functional trait in a population requires an explanation in terms of natural selection even if that particular trait originally was the result of a non-gradual change.

Darwin lacked an adequate theory of inheritance. But his contemporary Gregor Mendel discovered a crucial piece of the puzzle: there are units—now called “genes”—that are discretely transmitted from parent organisms to their offspring and which play a crucial role in the inheritance of phenotypical characteristics. Almost a century later, Francis Crick and James Watson discovered the structure of the DNA molecules stored in cell’s chromosomes. This helped to establish that genes are strings of DNA—instead of some kind of protein molecule, as many expected at the time (Watson 1968: 12). The synthesis of evolutionary theory with Mendelian genetics and eventually with molecular genetics has proven to be extremely fruitful. Genetics not only contributes to the explanation of the underlying mechanisms involved in the inheritance of phenotypical traits. It also contributes to the explanation of the phenotypical variations required for evolution to take place. At least some of such variations result from random mutations in strings of DNA. Molecular genetics also provides crucial evidence of the common descent of existing organisms and their links in the tree of life. It must be mentioned, however, that relatively recent research shows that genes are not the only factors involved in heredity, so treating them as the single “unit of inheritance” is an oversimplification. Organisms also inherit plenty of traits in non-genetic ways, ranging from the epigenetic inheritance of cellular structures and molecular machinery (other than DNA) to the “cultural” inheritance of behaviors that occurs even in simple organisms that are able to learn from their parents.67

67 See Sterelny and Griffiths (1999: Ch. 5) for a discussion of non-genetic forms of inheritance.
A controversy in the field of evolutionary biology that is particularly relevant to the philosophy of biology involves the “units of selection”. Classical Darwinism regards natural selection as operating mainly at the level of individual organisms, regarded as the entities that inherit traits and may succeed or fail to survive and reproduce. But some theorists—such as George Williams (1966) and Richard Dawkins (1976)—propose that genes are the real units of selection in biology. According to this view, the evolutionary process selects those genes that have favorable phenotypical “expressions”—instead of those organisms that have favorable traits. Genes are regarded as the main beneficiaries of the process. Other theorists propose that natural selection operates at multiple levels. Richard Lewontin (1970), for example, suggests that natural selection acts on any entities that display heredity, variation and differential fitness—whether they happen to be genes, organisms, populations or even species. The etiological account of functions discussed below works regardless of which of these views on the units of selection turns out to be correct. While the way some versions are formulated does seem to make assumptions about the units of selection, such assumptions can be easily avoided. As long as a biological trait is the result of a selection process, the trait can have an etiological function. The idea that selection processes also operate at the level of learning processes (discussed in Chapter 6) does assume a sort of multi-level view. However, it is worth noticing that even Dawkins—the most vehement advocate of the gene-centered view—is open about natural selection operating on more than one level. Indeed, he is known for proposing that there are units of cultural inheritance, that he baptized as “memes”, which are also units of selection.

The idea behind the contemporary teleological account is that the process of natural selection not only explains why biological items exist, but also what are their functions. An
influential philosophical defense of the account was offered by Larry Wright (1973). I will briefly discuss his proposal before presenting the main current formulations of the view.

4.4 Wright’s Defense of Teleology

Wright suggested that the questions “What is the function of $X$?” and “Why do $C$’s have $X$?” are both requests for “functional explanations” and that the answers to them are the same (Wright 1973: 155). For example, the basic common answer to “What is the function of the heart?” and “Why do humans have a heart?” is “To pump blood” (Wright 1973: 155). Of course, a more complete answer to the second question requires an evolutionary explanation: hearts have been naturally selected in order to pump blood. This is precisely what Wright proposed (Wright 1973: 162-164). According to this account, a function depends on causal history or etiology: “functional explanations... concern how the thing with the function got there... they are etiological, which is to say «causal»...” (Wright 1973: 156). More specifically, what matters are the effects of having $X$’s for ancestors—which Wright called “consequences”. We have hearts because the hearts of our ancestors pumped blood, which contributed to their survival and reproduction: “the consequences... must be invoked to explain why $X$ is there” (Wright 1973: 162). Notice that this account does not rely on an Aristotelian view of teleology in terms of “final causes”. There is nothing that may raise questions about “backward” causation in it. There is only the “forward” or “efficient” causation involved in selective history. The effects of $X$ can turn into causes for $X$ existing now precisely because they are past effects. This non-vicious circle is made possible by the reproduction of traits. Nonetheless, the etiological account may be said to make legitimate the teleological characterization of functions. According to Wright, his analysis
“shows what it is about functions that is teleological... [it] provides an etiological rationale for the functional «in order to»...” (Wright 1973: 162). So talk of hearts existing in order to pump blood turns out to be correct if ‘in order to’ is understood etiologically.

Contemporary advocates of the etiological account often follow Wright in making use of explicitly teleological language to characterize functions: the function of an item is said to be its “purpose” or what it is “supposed” to do, which is a result of evolutionary “design”. The account of functions itself is called “teleological”. However, the intended meanings of these words is strictly etiological:

A thin form of teleological description can be grounded in a Darwinian view. For example, the Darwinian can say that the function of a body part is the thing it does that has led to its being favored by natural selection. In that thin sense, the function is what that structure is “supposed” to do. This is a very deflationary sense of “supposed to.” (Godfrey-Smith 2009: 13)

In fact, since the etiological account of functions is at odds with traditional teleological views, some biologists prefer to use the word ‘teleonomy’ instead of ‘teleology’ in order to talk about “those parts of traditional teleological thinking that can be given a foundation in the operation of natural selection” (Godfrey-Smith 1996: 16). While the etiological account of functions is sometimes called “teleonomic” in philosophical contexts, it is more often called “teleological”. But the difference is merely terminological. I will here use ‘teleology’ and ‘teleonomy’ as equivalent terms and I will use words like ‘purpose’, ‘supposed’, ‘in order to’ and ‘design’ in their “thin”, strictly teleonomic or etiological, sense.
Let me make another brief terminological remark. Current teleologists often call functions “proper functions”. This is a technical term. The qualifier ‘proper’ is often intended to distinguish the teleological notion of function from other ones, such as the causal-role or dispositional one. While this usage is widespread, it also invites misinterpretation, since ‘proper’ contrasts with ‘improper’, yet it is unclear what an “improper function” would be. I will simply use the word ‘function’ or the longer ‘etiological function’ or ‘historical function’ when context may require disambiguation. The qualifiers ‘etiological’ and ‘historical’ contrast with ‘non-etiological’ and ‘ahistorical’, which properly apply to alternative notions of function. The word ‘proper’ is best used to qualify performances of a function. An item may be said to perform properly or improperly its function. Notice how odd it would sound to say that an item performs properly its proper function.

While some aspects of Wright’s analysis have been rejected by later teleologists, the core etiological point of the proposal as I described it above remains the same. But Wright actually did more than define functions etiologically. He suggested that the two conditions for \( Z \) to be the function of \( X \) are: “(a) \( X \) is there because it does \( Z \)” and “(b) \( Z \) is a consequence (or result) of \( X \)’s being there” (Wright 1973: 161). While the first condition is clearly intended to be historical—although it should be rephrased as ‘\( X \) is there because past \( X \)’s did \( Z \)’—the second condition requires \( X \) to presently do \( Z \). However, suppose that an organism has a kidney that does not work. The kidney is there because the kidneys of ancestors of the organism filtered blood. It is there in order to filter blood. Yet this particular kidney does not filter blood. Wright argued, correctly, that in such a case there is still a functional explanation of \( X \) being there, so \( X \) does have the function of doing \( Z \); but he also claimed that this “departs from the paradigms in a systematic but

---

68 The term was coined originally by Millikan (1984). But has since been used by many other authors.
“intelligible way” and that the word ‘function’ should be italicized “to make its use plausible and appropriate” in cases where an X fails to do Z (Wright 1973: 167). This is because in such cases the second condition he proposed for having a function is not met.

Current teleological accounts of functions draw a very different conclusion than Wright did: biological items that fail to perform their functions have these functions in the exact same way as those that succeed in performing them. Current versions rely merely on an etiological condition. The two main formulations of the current teleological account are those offered by Karen Neander (1991 & 1995) and Ruth Millikan (1984 & 1993). I will present their accounts and discuss along the way some qualifications proposed by Paul Griffiths (1993) and Peter Godfrey-Smith (1994).

4.5 Neander’s Proposal

According to Neander, biological functions are “effects for which traits were selected by natural selection” (Neander 1991: 174). She offers the following more detailed definition:

It is the/a proper function of an item (X) of an organism (0) to do that which items of X’s type did to contribute to the inclusive fitness of 0’s ancestors, and which caused the genotype, of which X is the phenotypic expression, to be selected by natural selection. (Neander 1991: 174)

Neander is interested specifically in the biological notion of function. But she suggests that the broader definition of function that applies also to non-biological items such as artifacts also relies on selection (Neander 1991: 175). This matches Wright’s claim that the functions of artifacts
result from “conscious consequence-selection”, while those of biology result from natural selection (Wright 1973: 163-164). Neander follows Wright in arguing that what is peculiar to biological functions is that they are determined by natural selection. But notice that her definition—unlike Wright’s—does not even mention what items presently do. In fact, Neander explicitly denies that the functions of biological traits are defined by anything other than their evolutionary history:

[For] a trait to have a proper function is not for it presently to have any actual causal role... or disposition. Instead, a trait has a proper function if there is something that it is supposed to do. According to my etiological theory, a trait is supposed to do whatever it was selected for by natural selection... Functional norms seem to be determined by a history of selection... (Neander 1991: 183)

The basic idea is that biological functions are not what items do, but what they are supposed to do given their selective history. The main advantage of this view—which accounts for its popularity in the philosophy of biology—is that it is able to account for malfunctioning: “It is the function of kidneys... to filter wastes from blood because that is what kidneys did in ancestral organisms that caused them to be favored by natural selection... and this fact remains true even if renal failure becomes universal” (Neander 1991: 183).

A failing kidney is malfunctioning or failing to perform its function. But this requires it to have the function of filtering blood. Otherwise, it would not be failing to perform it. If $X$’s having the function to do $Z$ requires in any way that $X$ actually does $Z$, then there is no room for characterizing $X$ as malfunctioning when it does not do $Z$. What this shows, Neander points out,
is that the notion of a biological function is a “normative notion” and that “a normative notion is not ahistorical” (Neander 1991: 168). The sense in which functions are “normative” obviously is not the strong prescriptive sense: kidneys do not have an “obligation” or “duty” to filter blood. Rather, it is a weaker teleonomic sense: kidneys are “supposed” to filter blood because kidneys were naturally selected for filtering blood—so if they fail to filter blood, they fail to do what explains what they were selected for. But having a biological function must be “normative” in this “thin” sense, since otherwise malfunctioning would not be even possible. Neander’s main argument for the etiological account of biological functions is that only a historical account can be “normative” in the required sense and properly characterize malfunctioning.

Neander sometimes characterizes natural selection as proportionally increasing the distribution of genes and their phenotypic expressions in a population (Neander 1991: 174). But Griffiths warns that proportional increase should not be a requirement. Populations often have various competing traits, some of which have a reduced number in the population, which anyway are there due to natural selection; so Griffiths argues that the function of a trait is simply what it has been selected for, independently of any proportional increase (Griffiths 1993: 414). Of course, many traits do proportionally increase due to natural selection. Griffiths qualification is simply that this should not be built into the etiological account of functions. Griffiths proposes the following definition of biological function:

Where $i$ is a trait of systems of type $S$, a proper function of $i$ in $S$’s is $F$ iff a selective explanation of the current non-zero proportion of $S$’s with $i$ must cite $F$ as a component in the fitness conferred by $i$. (Griffiths 1993: 415)
Notice that this definition explicitly requires selection to account only for the current existence of a trait, regardless of how many members of a population have the trait. Another difference with Neander’s definition is that Griffiths’ “leaves implicit various points about the gene/phenotype relation” (Griffiths 1993: 415). This is an advantage given that there is debate in biology about the precise details of such relation (Griffiths 1993: 416). Neander’s formulation seems to assume the controversial view that genes are the single units of selection. But this unnecessary commitment is easily avoided. In any case, Griffiths intends his definition to apply exclusively to biological functions. The only kind of selection he has in mind is natural selection.

4.6 Millikan’s Proposal

Millikan offers a more ambitious etiological account of functions that is even more general than Griffiths’ variant and is intended to apply not only to innately determined functions, but also to the functions of any items that result from the interaction between the innate mechanisms of an organism and its environment, as well as functions of items such as cultural artifacts and words of human natural languages which, Millikan argues, have selective histories even though they are not histories of natural selection operating over innate traits. In this section, I will focus on how her account applies to biological functions.

Millikan distinguishes between what she calls “direct proper functions” and “derived proper functions” (Millikan 1984: Ch. 1 & 2)—which I will simply call “direct” and “derived” functions. Her account of direct functions basically matches Neander’s and Griffiths’ accounts when applied to biological items. Although she introduces some interesting qualifications, the idea roughly is that an item’s function is determined by its selective history. Her account of
derived functions, in contrast, goes beyond: it allows for the ascription of functions to the 
products of selected mechanisms—products that have not been selected themselves, although 
they are the result of selected items which have the direct function of producing them.

I will first present Millikan’s account of direct functions and then elaborate on her 
account of derived functions. An item $A$ has a direct function to perform $F$, Millikan suggests, if 
and only if:

$A$ originated as a “reproduction” (to give one example, as a copy, or a copy of a 
copy) of some prior item or items that, $due$ in part to possession of the properties 
reproduced, have actually performed $F$ in the past, and $A$ exists because (causally 
historically because) of this or these performances. (Millikan 1993: 13)

Notice that Millikan does not explicitly mention natural selection or even selection of any kind 
here. Nonetheless, her definition of function clearly captures the basic features of the etiological 
account and is intended to rely precisely on selective processes. First, the direct function of an 
item depends on the causal history that explains why the item exists. As Millikan points out, her 
account “looks to the history of an item to determine its function rather than to the item’s present 
properties or dispositions.” (Millikan 1993: 13). This is intended to provide an account of 
functions that allows for things being able to have them even when they fail to perform them. 
Second, the relevant causal history is one that involves the reproduction of items—where 
reproduction is understood as a physical process that generates similar copies of the items. 
Consider a series of similar items, $I_1, I_2, I_3,... I_n$, were each successive item is a reproduction or 
copy of the previous one. According to Millikan, the members of this series form a
reproductively established family” (Millikan 1984: 18-25). The links between the members of such a family are causal, so two items do not belong to the same family merely because of resembling each other. Of course, two items can be members of the same family because they have a common ancestor: the causal links of reproduction typically produce family trees.

Millikan’s proposal is that an item $A$ has performing $F$ as a direct function when (1) $A$ is a member of a reproductively established family, (2) some ancestors of $A$ in the family actually performed $F$ and (3) $A$ exists because of this. If $A$ is a biological trait, the causal-historical explanation of why it exists in virtue of its ancestors performing $F$ is provided by natural selection: performing $F$ contributed to the ancestors’ survival and reproduction, so the trait that allowed the ancestors to perform $F$ was passed on to their descendants in the family.

What kinds of items are members of a reproductively established family? Genes are a clear example. Each token gene is literally a copy or reproduction of ancestor genes of the same type. The function of a token gene is to do whatever the ancestors in its lineage did that explains why the lineage has survived and, consequently, why the token gene exists. But what about phenotypic traits such as organs and innate behaviors? Neither of them are direct copies of previous organs or innate behaviors. Millikan draws a distinction between “first-order” and “higher-order” reproductively established families to clarify this issue (Millikan 1984: 23-25). Genes belong to first-order families, while organs and innate behaviors belong to higher-order families, since they are only reproduced via the reproduction of genes. For example: “although my heart is not a copy of my parents’ hearts, it was produced... in accordance with the proper functions of certain of my genes which were directly copied from my parents’ genes” (Millikan 1984: 25). This means that hearts do belong to reproductively established families—they are
reproduced items—and that their function is to do what their (higher-order) ancestors did which explains why they exist (the full explanation, of course, has to include the selection of genes). The same considerations apply to innate behaviors such as sneezing. The functions of organs and innate behaviors are direct functions: “the sorts of purposes that we ordinarily attribute to biological devices such as hearts and inherited behaviors are direct proper functions, granted that our guesses about the evolutionary histories of these devices are correct—our guesses about their «reasons for survival»” (Millikan 1984: 28).

As formulated, the account attributes functions to reproduced items that exist because they have been historically selected to perform certain tasks. But without some further qualification, this account seems to have a problem with vestigial items. The same may be said of Neander’s account. Consider the vestigial eyes of mole rats. Mole rats live underground and are blind. But they still have vestigial eyes, although completely covered by a layer of skin. Now consider the malformed eyes of an individual blind rabbit. The explanation of why the eyes of a mole rat and the blind rabbit exist involves the evolutionary history of their ancestors, whose eyes were selected for seeing. In the case of the blind rabbit, the etiological account explains why the eyes are failing to do what they are supposed to. That is the main virtue of the account over non-historical ones. But the account seems to wrongly imply that the eyes of the mole rat are also failing to do what they are supposed to. The solution to this problem relies on the fact that natural selection is not only responsible for the formation of a trait, but also for its maintenance. Millikan does point out that “the main business of natural selection is steady maintenance of

---

69 Millikan’s account of higher-order reproductively established families does seem to be effective in including phenotypical traits. But notice that her discussion seems to assume a gene-centered view of natural selection. Why not treat hearts as reproduced or inherited items on their own right? Why should genes be regarded as the first-order reproduced items?
useful traits against new intruders in the gene pool” (Millikan 1989: 173). However, this feature is not part of her definition of direct functions. Godfrey-Smith argues that the historically recent maintenance of a trait by natural selection is “important enough to make this a constitutive part of the concept of function” (Godfrey-Smith 1994: 356). He proposes a “modern history” account according to which biological functions are “dispositions or effects a trait has which explain the recent maintenance of the trait under natural selection” (Godfrey-Smith 1994: 356). This qualification allows the etiological account to distinguish vestigial from non-vestigial traits. The eyes of the mole rat—unlike those of the blind rabbit—have not been recently maintained by natural selection for seeing, so they do no have a function that they are failing to perform.

Millikan argues that items can have functions even if they have not been selected: these are not direct functions but derived functions. An item $A$ has a derived function to perform $F$, Millikan suggests, if and only if:

$$ A \text{ originated as the product of some prior device that, given its circumstances, had performance of } F \text{ as a proper function and that, under those circumstances, normally causes } F \text{ to be performed by means of producing an item like } A. \quad \text{(Millikan 1993: 13-14)}$$

Millikan’s account of derived functions is intended to illuminate how things that are “new under the sun”, which do not have a selective history of their own, can nonetheless sometimes have biological functions (Millikan 1984: 45). The functions of these items are derived from the direct functions of mechanisms that have been selected precisely for producing them. The account of derived functions is also etiological, since it is firmly grounded on selective history.
Millikan uses the example of chameleons’ camouflage to illustrate the need for an etiological account of derived functions (Millikan 1984: Ch. 2). Chameleons have mechanisms that allow them to rearrange the pigment patterns of their skin. Some species—such as Smith’s dwarf chameleon—use their ability to change color as camouflage: they change their color so that it matches the color of their background, making them less visible to predators. Now suppose that a chameleon sits on a surface with a color pattern that none of its ancestors has ever encountered before. Presumably, the resulting color pattern of the chameleon is supposed to match the color of the surface it is sitting on, thereby making it less vulnerable to predators. But an orthodox etiological account apparently cannot account for this. This particular background pattern has not been encountered before, so there cannot be any selective history determining specifically that the pigment pattern of the chameleon’s skin is supposed to match it. Godfrey-Smith explains the problem by distinguishing between innate mechanisms or structures and the states that they produce:

It is one thing to say that a mechanism has a biological function, and another to say that a particular state of this mechanism has a function... Structural features... are products of an evolutionary history, a history of heritable variation in fitness. But states... are not the right sort of things to have such a history... A state may or may not profit an organism, but its nature is not the product of the success of previous states of the same type, and has no propensity to lead to the future survival and proliferation of the type. (Godfrey-Smith 1989: 542)

It seems that the states produced by innate mechanisms cannot have etiological functions. However, some states produced by some innate mechanisms prima facie seem to have purposes.
Presumably, the entirely new pigment pattern of the chameleon has the purpose of making the chameleon less visible to predators by making it blend with its surroundings. The phenomenon of camouflage is widespread in the animal kingdom. Animals often have color patterns that make them less visible to predators or prey. When such a color pattern is fixed—as it is in most cases—the orthodox etiological account has no problem explaining why the pattern has a function: it is a phenotypical trait that has been naturally selected for camouflage purposes. Yet, the non-fixed and variable color patterns of some species of chameleon, octopus and cuttlefish also have camouflage purposes, which must be somehow the result of the evolutionary history of these species.

Millikan’s account of derived functions expands the etiological account to explain how the products of innate mechanisms can indeed have functions. Even if the entirely new color pattern of a chameleon cannot itself have been specifically selected, the mechanisms responsible for the pigment rearrangement have been naturally selected for producing variable color patterns to match variable backgrounds. Protecting the chameleon from predators by making it blend with its current background is a direct function of the mechanisms. This function is performed by means of producing states of a certain general type: skin color patterns that match the color patterns of the background. The state of the chameleon sitting on a background not encountered before by its ancestors is precisely of this general type. The evolutionary explanation of why the mechanisms exist explains why the new pattern of skin color exists. It also explains its function: the purpose of the pattern is to make the chameleon less visible. While the state has not been selected for performing this function, it nonetheless has a function that is derived from the selected or direct function of the mechanism that produces it (Millikan 1984: 39-45).
Millikan’s account of derived functions has a crucial application in the explanation of the functions of non-innate states and behaviors that are the combined result of the innate mechanisms of an individual organism and its interaction with its environment. Consider the following case: “If a rat becomes ill within a few hours after eating a specific food, it will later shun all foods that taste the same. For example, if the rat eats soap and soon becomes ill, thereafter it will refuse to eat soap.” (Millikan 1993: 224). The mechanism responsible for this behavior is innate. It was selected because it sometimes made ancestor rats avoid poisonous foods. The mechanism has the direct function of preventing poisoning. However, the specific flavors avoided by an individual rat will depend on the rats’ experience: the rat learns to avoid them after becoming ill. This learning is the result of the interaction between a specific innate mechanism and the rat’s experience. The rat cannot learn, for example, to avoid foods that “merely look the same or that are found in the same place as the food eaten prior to illness” (Millikan 1993: 224-225). The learned behavior has not been itself naturally selected, yet it is the product of a naturally selected innate mechanism “designed” to enable the rat to learn to avoid eating certain foods. The learned or acquired behavior and the inner states in the rat’s nervous system responsible for it have a function or purpose. When a rat has learned to avoid soap-tasting substances, its behavior and the inner states that cause this behavior have the function of preventing poisoning. Yet neither the behavior nor the inner state has a history of selection of its own. Millikan’s account of derived functions provides a compelling explanation how such learned or acquired states and behaviors can have etiologically grounded functions. They have functions derived from the direct functions of the learning mechanisms that exist in order to produce them in response to variable environmental circumstances (Millikan 1993: 225-228).
The claim that functions are determined by history may be initially counterintuitive. Let us consider two cases presented by Neander as alleged counterexamples to the etiological account of functions. The first is the case of William Harvey’s discovery of the function of the heart two centuries before Darwin’s discovery of the process of natural selection. Since Harvey had no historical notion of the origins of hearts, it may be argued that the function he discovered cannot be etiological (Neander 1991: 189). The second is a thought experiment: suppose that lions came to exist instantly a few minutes ago by a purely accidental arrangement of molecules. The organs of “instant lions” cannot have etiological functions, since they do not have any selective history, but it looks like they have functions nonetheless (Neander 1991: 189). The advocate of the etiological analysis has compelling replies to these alleged counterexamples. But let us explore for a moment the alternative view. What could be the ahistorical functions of the organs of instant lions and the function discovered by Harvey if it were not a historical function?

The main ahistorical naturalistic alternative to the etiological account is the view that functions are defined by the causal roles that items play within larger systems. The chief proponent of the view is Robert Cummins, who argues that the function of an item is its causal contribution to some overall capacity of the system of which it is a part. He proposes the following definition:

\[ x \text{ functions as a } \phi \text{ in } s \text{ (or: the function of } x \text{ in } s \text{ is to } \phi) \text{ relative to an analytical account } A \text{ of } s \text{'s capacity to } \psi \text{ just in case } x \text{ is capable of } \phi\text{-ing in } s \text{ and } A \text{ appropriately and adequately accounts for } s \text{’s capacity to } \psi \text{ by, in part, appealing} \]
Cummins idea is that an overall capacity of a complex system is explained by decomposing the system into its various parts, showing what each of these parts does and how the combination of the contributions of each of the parts accounts for how the system as a whole is able to have the analyzed capacity. Cummins calls this kind of explanation “functional analysis”. A clear illustration of functional analysis, he suggests, is provided by the schematic diagrams used in electronics: “Since each symbol represents any physical object whatever having a certain capacity, a schematic diagram of a complex device constitutes an analysis of the electronic capacities of the device as a whole into the capacities of its components.” (Cummins 1975: 760).

Here is a simple example: the function of each of the diodes in a circuit that turns alternate current (AC) into direct current (DC) is to allow the flow of electricity in only one direction, while the function of the capacitor is to smooth the peeks of voltage that result from the rectification performed by the diodes (the diodes only make one of the outputs always positive and the other always negative, but they do not remove the cyclical variations in voltage); the overall capacity of the circuit is explained by the causal contributions made by each of these components and how they are arranged.

According to Cummins, the explanation of the functions of biological organs is also provided by a functional analysis: “It is appropriate to say that the heart functions as a pump against the background of an analysis of the circulatory system’s capacity to transport food, oxygen, wastes, and so on, which appeals to the fact that the heart is capable of pumping.” (Cummins 1975: 762). In this example, the overall capacity to be analyzed is the circulatory system’s capacity to circulate blood and thereby transport nutrients, oxygen, waste, hormones,
etc., to and from cells throughout the body. It is simply undeniable that this capacity of the circulatory system is the result of the interlocking capacities of its parts and that the specific contribution of the heart is that it works as a pump in the system. An analysis of how the circulatory system works (when it works properly) must surely have the basic features of a Cummins-style functional analysis. Consequently, it has to be acknowledged that the idea that functions are causal roles played within a system has some initial plausibility.

Let us consider the two alleged counterexamples to the etiological account presented by Neander: Harvey’s discovery of the function of the heart and “instant lions”. It may be argued that the function discovered by Harvey was merely a causal-role function—since he knew nothing of evolutionary history—and that this illustrates what kind of functional explanation is really employed in physiology. This is a rather weak argument. But many philosophers who are skeptical about the etiological account seem to share the intuitions that support it. Similarly, it may be argued that the organs of instant lions have causal-role functions. A functional analysis will reveal that their kidneys have the function of filtering blood, their hearts have the function of pumping blood, and so forth. The etiological account cannot attribute functions to instant lions, but the causal-role account does not have this limitation. Indeed, the functions ascribed by the causal-role account to instant lions are the same it ascribes to real lions. These points give reasons to favor the causal-role account over the etiological account to those who have the intuitions that Harvey discovered all about the function of the heart and that the organs of instant lions really would have the same biological functions as those of real lions. What can the advocate of the etiological account reply?

The first point to notice is that the advocate of the etiological account can and should
accept that Harvey did discover the contributions of the heart (when it is functioning properly) to
the circulatory system and that the causal-roles of the organs of instant lions would indeed be
those ascribed by a Cummins-style functional analysis if such creatures suddenly came to exist.
The second point to notice is that the causal-roles of organs do play a crucial explanatory role in
the etiological account as well, although with an important qualification. The causal-roles of
*properly working* hearts within the circulatory systems of *ancestors*, for example, are what
explain why their descendants have hearts. The hearts of present day creatures *are there* only
because the hearts of past creatures pumped blood and contributed to the circulation of nutrients,
oxygen and waste, which thereby contributed to their survival and reproduction, leading to the
genes for hearts being passed on to their descendants. The etiological account does not require a
heart to *presently* pump blood in order to have the function of pumping blood. This is the main
disagreement between etiological and purely causal-role accounts. But a causal-role analysis is
anyway already part of the etiological account of what has been selected by natural selection.

The crucial debate is not whether causal-roles matter, but whether the purely causal-role
account is the proper account of biological functions. Cummins explicitly denies that the
evolutionary history of an organ defines its function and argues that the function of an organ
depends on its current dispositions:

> [If] the function of something in a system $s$ is to pump, then it must be capable of
> pumping in $s$. Thus... to attribute a function to something is, in part, to attribute a
> disposition to it. If the function of $x$ in $s$ to $\phi$, then $x$ has a disposition to $\phi$ in $s$.
> (Cummins 1975: 757-758).
If the kidneys of an instant lion have the disposition to filter blood and thereby contribute to the capacities of its circulatory and urinary systems, this is their biological function according to Cummins. This would be an advantage over the etiological account if the intuition that the kidneys of instant lions have the biological function of filtering blood were correct. However, that intuition is highly suspicious and, more importantly, the causal-role account has a major drawback that outstrips any advantages it may seem to have over the etiological account.

By requiring an item to actually have a causal role or disposition in order to have a function, the causal-role account cannot explain what would it be for the item to malfunction or fail to perform its function. Consider a real lion—i.e., not an “instant lion”—with a non-working kidney that does not have a disposition to filter blood. Surely the kidney is malfunctioning. But according to the causal-role analysis this kidney does not have the function to filter blood to begin with, since it has no such disposition. I take it that most advocates of the causal-role analysis of functions accept that evolutionary history explains why real lions have kidneys. What they deny is that this is what defines the function of kidneys. Regarding Harvey’s discovery, they may want to say that what Harvey discovered is the function of the heart, while Darwin merely discovered the historical origins of the heart. However, by denying that evolutionary history defines the functions of organs and other biological traits, the causal-role account is simply unable to characterize them as malfunctioning or failing to perform their functions, as Neander (1991 & 1995), Millikan (1993 & 1999) and others have emphasized.

How can the etiological account handle the alleged counterexamples? Regarding “instant lions”, the etiological account must claim that their organs do not have functions of the sort that allow for ascriptions of malfunctioning. When a real lion’s kidneys do not filter blood properly
or at all, they are failing or malfunctioning because they have been selected for filtering blood. When an instant lion’s kidneys do not filter blood, they are not malfunctioning, since they are not *supposed* to filter blood to begin with. Surely we should say, when they do filter blood, that they have this causal-role. This is true of both real and instant lions. But the purely causal-role account is wrong to characterize the functions of the organs of real lions in the same way as those of the organs of instant lions. As Neander points out, “theories which imply that instant lions... would have proper [biological] functions do not capture the distinction between what an item does and what it is supposed to do...” (Neander 1991: 180). Consequently, “wayward intuitions about instant lions might have to be revised in the light of the fact that biology has and needs a notion of a «proper function» that is normative” (Neander 1991: 180). There is the intuition some people have that the organs of instant lions would have the same functions as those of real lions. But there is also the strong intuition that things that have functions in some sense have “purposes”: that they are “supposed” to perform their functions but may “fail” to do so. The dilemma is that an account of the nature of functions cannot be consistent with both intuitions. The only way out of the dilemma that preserves the second intuition is to revise the first one. While the traditional teleological intuition was associated with notions that are incompatible with naturalism, the etiological account is not. So a firm commitment to naturalism does not provide any reason to favor a causal-role account over an etiological one.

Regarding Harvey’s discovery, what the etiologist must say is that he indeed discovered what hearts do when they function properly—which, unknown to him, turned out to be also what hearts have been naturally selected for doing. That Harvey did not know about natural selection is irrelevant to the issue of the nature of biological functions (Neander 1991: 175-176; Millikan
1993: 15). Just like Mendel discovered genes without having a clue about their underlying nature as strings of DNA, Harvey discovered the function of the heart without having a clue about its underlying etiological nature. The intuition that what he discovered was merely a causal-role function is simply wrong. The etiological functions of items are first identified by the causal-roles of normal well-functioning samples. As Neander points out, this is how physiology works: “The primary physiological analysis necessarily abstracts away from maladaptive environments... and the infinite possible pathways of pathology... by describing what the components do when they are functioning properly.” (Neander 1995: 117-118). Harvey abstracted away from failing hearts and thereby described how hearts work when functioning properly. That hearts are supposed to pump blood, however, can only follow from some teleological or teleonomic account of biological functions. Harvey himself was a teleologist, likely of the theological kind (Neander 1991: 176). So he likely thought he had discovered what hearts are supposed to do given their intelligent design. If this is the case, then we have to say that he was wrong about the source of natural teleology. The heart is indeed supposed to pump blood, but this is because it has been naturally selected to do so.

4.8 Conclusion

The etiological account relies on the selective history that accounts for the existence of trait in order to explain its function. The crucial advantage of this account over dispositionalist alternatives is that it offers a fully naturalistic account of what an item is supposed to do and what counts as malfunctioning or failing to perform its function. It is this feature that is of interest to semantics. Teleosemantic theories promise to explain the nature of meanings in a way
that offers a naturalistic solution to problems of misrepresentation. I will discuss the main available teleosemantic approaches in the next chapter.
5.1 Introduction

The etiological account of functions discussed in Chapter 4 has the crucial advantage over dispositionalist alternatives of offering a fully naturalistic account of malfunctioning. Teleosemantic theories exploit this feature to offer a fully naturalistic account of the “normative” character of meanings. If symbols or representations have etiological functions that determine how they are *supposed* to be used, then they can be said to have meanings that are independent of their current uses. When a representation or symbol is not deployed the way that accounts for its existence, such deployment can be characterized as a misuse. In this chapter I will discuss in depth the main teleosemantic approaches, including those that focus on the effects of representations to explain their meanings, and those that focus on their causes. I will argue that each of these approaches has insights but also drawbacks and that a hybrid account that explains meanings in terms of both the inputs and outputs of representations is needed. I will start with a discussion of indicator or informational theories of meaning and the problem of misrepresentation they face—which is a major motivation for favoring a teleosemantic account.

5.2 Indicator or Informational Semantics

According to indicator or informational semantics, the meanings of symbols or representations are basically their referential properties, which are determined by their direct causal links to reality. More specifically, the relevant links are alleged to involve *reliable causation*. A symbol
or representation is claimed to refer to what reliably causes tokens of it. Consider the reliable causal correlation between smoke and fire or between the rings of a tree and its age. Smoke and tree rings can be said to “carry natural information about” or “indicate” the presence of fire and the age of trees, respectively, because of the law-like correlations involved. Similarly, a token representation is said to indicate or carry natural information about the presence of those things which its type is lawfully or nomically linked to. This is a very demanding requirement for indicating or carrying information. Dretske, for example, proposes: “A signal $r$ carries the information that $s$ is $F = \text{The conditional probability of } s's \text{ being } F, \text{ given } r... \text{ is 1}”$ (Dretske 1981: 65). Suppose that $r$ is a perceptually induced mental state that represents that something is a horse—say, the word ‘horse’ in an agent’s language of thought. Dretske’s requirement entails that an actual horse must always be present when ‘horse’ is tokened for ‘horse’ to carry information about or indicate specifically horses.

An obvious problem is that the indicator theory, even if it worked, would only apply to perceptually induced representations: what about the stimulus-independent tokens of ‘horse’? The indicator theory is at best incomplete and needs to be complemented with an account of how the meanings of stimulus-independent tokens derive from the meanings of stimulus-dependent tokens (Devitt and Sterelny 1999: 157). Another problem is that the indicator theory only applies to symbol or representation types that have direct links to reality: what about words like the ‘bachelor’—and their mental counterparts—whose meanings depend on those of other words? This is a problem of scope: the theory does not have the generality required to provide us with a theory of meaning that covers all kinds of word (Devitt and Sterelny 1999: 156). These problems may be solved by restricting the indicator theory to a theory of stimulus-dependent meaning and
combining it with other theories that provide the missing explanations. However, the indicator theory faces a serious problem that undermines its account of stimulus-dependent meanings. As Godfrey-Smith points out: “informational semantics has been hounded by a problem with error... and no other problem has hounded the theory so persistently” (Godfrey-Smith 1989: 533-534).

Stimulus-dependent tokens of ‘horse’ are not only caused by horses, but also by a number of horse-looking things, such as distant cows or muddy zebras. A token of ‘horse’ caused by a muddy zebra is being misused: this is a clear case of misrepresentation. But the word ‘horse’ is causally linked to all of these things, so the indicator theory wrongly entails that it must mean “horse or distant cow or muddy zebra or...”. The indicator theory cannot characterize misuses as misuses: it must characterize tokens of ‘horse’ caused by muddy zebras as correct uses of the word. As Godfrey-Smith points out, the problem is the following:

[Any] environmental state that can cause a representation token is included in the class of environmental states informationally linked to that representation type. Anything that can cause you to think ‘Horse!’ contributes to the content of that representation type, and as a result error is impossible. (Godfrey-Smith 1989: 537)

Fodor attempts to overcome the problem by invoking an alleged “asymmetric dependence” of incorrect uses on correct uses: “the fact that cows cause one to say ‘horse’ depends on the fact that horses do; but the fact that horses cause one to say ‘horse’ does not depend on the fact that [cows] do” (Fodor 1987: 108). The basic idea is that the meaning fixing law-like regularity underlying the use of a symbol is exclusively the one that is not asymmetrically dependent on other regularities. Fodor thinks that this simple proposal shows why ‘horse’ applies correctly to
horses and not muddy zebras. While ingenious, however, this suggestion does not truly avoid the problem. Why should we suppose that the most basic regularity is precisely the one linking ‘horse’ with horses? It seems rather that the most basic regularity is the one linking ‘horse’ with horse-looking things, of which actual horses, distant cows and muddy zebras are all instances (Godfrey-Smith 1989: 539-540; Devitt and Sterelny 1999: 159-160).

The indicator theory, while initially attractive, does not have the resources to make room for misrepresentation. The problem arises due to the dispositionalist character of the theory: $x$ being reliably caused by $y$ is an actual disposition governing occurrences of $x$. As Saul Kripke argues, dispositionalist theories of meaning constitution cannot properly account for misrepresentation, since the users of representations often have dispositions to make mistakes while using them and often lack the dispositions to correctly deploy them (Kripke 1982: 28-35). A theory that reduces meanings to the actual regularities or dispositions underlying the uses of representations cannot avoid the problem of error. As I argued in Chapter 3, this is the case for Horwich’s account of the meanings of words in terms of law-like regularities involving the acceptance of basic sentences containing them. This is also the case for the indicator account of the meanings of representations in terms of their reliable causes.

Notice the affinity between the indicator theory and the causal-role account of functions discussed above. They are both dispositionalist and have analogous problems: one cannot account for misrepresenting and the other cannot account for malfunctioning. Indeed, the indicator theory may be regarded as adopting the causal-role account of functions and identifying meanings with a part of those functions. The causal-role of a representation in a cognitive system includes not only its causes, but also its effects—which are crucial in the explanation of the
contribution of the representation to the systems’ overall capacities. Clearly, the indicator theory does not identify meanings with the complete causal-roles of representations. However, the indicator theory does identify meanings with the (distally characterized) causes of representations, which are part of their causal-roles: the part that allegedly fixes their meanings or contents. When seen this way, the indicator theory’s problem of error turns out to be an instance of the broader problem of malfunctioning that undermines causal-role accounts of functions. On the other hand, the etiological account of functions is tailored to make room for malfunctioning and it is precisely an alternative to dispositionalism. This suggests that the problem of error that undermines the indicator theory may be overcome by adopting an etiological account of functions.

5.3 Teleosemantics and Indication

The simplest teleosemantic proposal is to combine indication with an etiological account of functions. Dretske (1988), for example, modifies his previous indicator theory precisely in this way. According to his amended version, “the meaning or representational content of an element is what it has the function of indicating rather than what it actually succeeds in indicating” (Dretske 1988: 151). The idea is that the meaning of an item depends on what it is supposed to indicate—in the teleological sense of “supposed”. This combination of teleology and indication is meant to allow for indication relations fixing the meanings of representations while making room for representations sometimes failing to indicate. I will discuss this form of teleosemantics and the problems it faces before exploring other versions of teleosemantics that depart further from the indicator theory.
Dretske offers a historical explanation of why the function of a representation is to indicate something. His explanation takes into account both the causes and the effects of representations. The crucial effects of representations, he reasonably argues, are their behavioral effects. Representations allow organisms to behave in ways that are appropriate responses to their surroundings. According to Dretske, representations are able to do this because they are able to carry information about those surroundings. The causal relation between a representation $C$ and the behavior $M$ it elicits, Dretske argues, must be explained by the *meaning* of $C$ which in turn, he argues, “will have to be explained by the fact that $C$ indicates, or has the function of indicating, how things stand elsewhere in the world” (Dretske 1988: 84). But why is the *function* of $C$ to indicate something? Dretske argues that $C$ has this function because of what it has been *recruited* to do: “Once $C$ is recruited as a cause of $M$—and recruited as a cause of $M$ *because of what it indicates about $F$*—$C$ acquires, thereby, the function of indicating $F$.” (Dretske 1988: 84).

Whether “indicating”—in Dretske’s sense—can indeed be a *function* is something I will discuss after presenting the view in more detail.

Dretske’s account of recruitment is historical. A simple example is provided by artifacts. Consider a thermostat that turns on the heating in a room when the temperature drops. Here $M$ is the heating turning on, $C$ is the state of the thermostat that causes $M$, and $F$ is the drop in temperature that causes $C$. The function of $C$ is to indicate $F$, Dretske suggests, because it has been recruited for this job: the system has been intentionally designed for $F$ to reliably cause $C$ and $C$ to reliably cause $M$ (Dretske 1988: 86-89). But how can we extend this kind of account beyond the case of intentional design?

Consider the hair triggers used by the Venus flytrap (a carnivorous plant) to detect that an
insect has landed on it. When an insect touches twice any of the hair triggers in a leaf within a short time interval, the plant snaps together the two halves of the leaf, trapping the insect between them (when successful). Dretske comments on this example:

[Here] leaf movement \((M)\) is caused by an internal state \((C)\) that signals the occurrence of a particular kind of movement, the kind of movement that is \emph{normally} produced by some digestible prey. And there is every reason to think that this internal trigger was \emph{selected} for its job because of what it \emph{indicated}...

\cite{Dretske1988:90} (emphasis added)

Dretske suggests that the same kind of the explanation applies whenever a representation is innately linked to an instinctive behavior. For example, when noctuid moths detect the high-frequency sounds \emph{normally} produced by echolocating bats approaching to prey on them, they swerve away. Moths’ bat-detectors \((C)\) were \emph{selected} for triggering the avoidance behavior \((M)\), Dretske argues, because they normally \emph{indicated} that bats were approaching \((F)\) \cite{Dretske1988:91}. Dretske’s suggestion is that whenever the instinctive behaviors of an organism are innately triggered by perceptual mechanisms or states, \emph{natural selection} is responsible for recruiting such mechanisms or states as indicators: “What the theory of evolution has to tell us about these cases... is that...\(M\) is produced by an indicator of \(F\) because such an arrangement confers a competitive advantage on its possessor.” \cite{Dretske1988:92}.

Once an item has been recruited for this job by natural selection, Dretske argues, it has the biological function of indicating:

If we suppose that, through selection, an internal indicator acquired (over many

\footnote{Dretske does not discuss the distinction between mechanisms and states. So it is not entirely clear what is his precise account of what is selected. See Godfrey-Smith (1989) for a critical discussion of this issue.}
generations) a biological function, the function to indicate something about the animal’s surroundings, then we can say that this internal structure represents (or misrepresents, as the case may be) external affairs. (Dretske 1988: 94)

If indicating is a teleonomic function, then items that have this function can fail to perform it. As Dretske points out, an indicator may represent or misrepresent. However, Dretske’s account does not make enough room for misrepresentation in spite of its reliance on teleology. I will get back to this in a moment.

What about acquired representations? When a representation is acquired rather than innate, Dretske argues, its recruitment as an indicator is done instead by the learning process. In operant conditioning, he proposes, reinforcement recruits indicators as causes of behavior: “By the timely reinforcement of certain output—by rewarding this output when, and generally only when, it occurs in certain conditions—internal indicators of these conditions are recruited as causes of this output.” (Dretske 1988: 98). Consider a pigeon that learns in a lab to get food by pecking on a screen if and only if pictures of trucks are displayed (Dretske 1988: 153). Here the learned behavior would be pecking on truck pictures and the indicator recruited would be a detector of truck pictures. Or consider the following example: when a blue jay becomes ill after eating a monarch butterfly, it avoids eating monarch butterflies again (Dretske 1988: 111). Here the learned behavior would be refusing to eat monarch butterflies and the indicator recruited would be a detector of monarch butterflies.

Operant conditioning is a general kind of learning: it does not have a specific target, which makes it particularly plastic. The case of the pigeon is paradigmatic. But perhaps the blue jay learns not to eat monarch butterflies because it has instead a specialized innate mechanism. In that case, the function of the learned behavior and the representation that causes it would be best characterized as a derived function. In any case, there are various forms of learning and surely something along the lines of operant conditioning has a place among them. Also, notice that the complex mechanisms responsible for operant conditioning must themselves be the result of the evolutionary history of organisms that have them. So the functions of responses to stimuli learned in this way
and the natural selection of innate perceptual representations. In both cases, some perceptual representation $C$ is recruited to detect some $F$ in the surroundings and cause some behavior $M$. In both cases, the recruitment happens because of the benefits of the effects of responding in certain ways to certain stimuli. Dretske indeed suggests that the learning process may itself be regarded as selection process: “Learning... is a way of shaping a structure’s causal properties in accordance with its indicator properties. $C$ is, so to speak, selected as a cause of $M$ because of what it indicates about $F$.” (Dretske 1988: 101).

The three kinds of historical process discussed by Dretske (artificial design, natural selection and learning) are alleged to recruit items to have the function of indicating conditions in the world—a function that items may fail to perform. What kind of account of misrepresentation is available to this approach? If the thermostat breaks and gets stuck in $C$ even though the temperature has not dropped, we can say that $C$ is supposed to indicate $F$—if this is what it was recruited for—but that it is failing to do so. Consider the Venus flytrap. If its hair triggers have been naturally selected for indicating that an insect has landed, they would be failing to indicate what they are supposed to when they are triggered by a botanist touching them twice with a stick. The same may be said of indicators selected by learning processes. The blue jay’s detector of monarch butterflies would be failing to indicate what it is supposed to when triggered by a fake monarch placed by a researcher.

Unlike the standard indicator theory, the combination of indication and teleology does make room for misrepresentation. The idea is that if item $C$ was selected for indicating $F$ because it did indicate $F$ during its recruitment, then it is failing to do what it is supposed to when it does...
not indicate \( F \) after its recruitment. Sometimes misrepresentation may happen because the mechanism itself is or becomes internally defective, like in the case of the thermostat stuck in \( C \) or a moth that is born unable to detect bats. But the interesting cases are those where the internal mechanisms are not defective and the misrepresentation occurs because “the world is deceptive” (Godfrey-Smith 1989: 546). In such cases, the teleological version of the indicator theory can claim that the conditions in the world that cause the misrepresentation are abnormal, since they differ from the conditions under which the indicator was recruited. The teleological component is intended to offer a naturalistically respectable characterization of the distinction between “normally and abnormally caused representation tokens” (Godfrey-Smith 1989: 546). When the Venus flytrap’s detectors are triggered by insects or the blue jay’s detectors are triggered by monarch butterflies, we may say that the detectors are operating under normal conditions. When the same detectors are triggered instead by researchers, the detectors are still supposed to indicate insects and monarch butterflies, respectively, but they fail to do so because they are operating under abnormal conditions.

What does the teleological version of the indicator theory regard specifically as a normal condition? An etiologically characterized normal condition must be a condition that obtains during selection. But this leaves room for more than one teleonomic account of normal conditions. As we will see in the next section, Millikan offers a very different account than Dretske’s. According to Godfrey-Smith’s interpretation, the teleological version of the indicator theory relies on the “ecologically normal” environment were the indicator evolved or was selected (Godfrey-Smith 1989: 546). And this is indeed what Dretske proposes. His idea becomes clear in his discussion of cases where misrepresentation occurs because of the
abnormality of the circumstances. Here is a revealing (albeit long) passage:

In many cases of biological interest, a sign—some internal indicator on which an animal relies to locate and identify, say, food—will... be a reliable indicator only in the animal’s natural habitat or in conditions that approximate that habitat. Flies, for instance, when given a choice between nutritionally worthless sugar fructose and some nutritive substance like sorbitrol, will invariably choose the nutritionally worthless substance and starve to death. Surprising? Not really. Under natural conditions... the substances that stimulate the receptors are nutritional. Under natural conditions, in a fly’s normal habitat, then, receptor activity indicates a nutritional substance. Furthermore, the correlation between receptor activity and nutritional value of its activator is no accident... Flies would not have developed (or maintained without modification) such a receptor system in environments where such a correlation did not exist. (Dretske 1988: 57-58)

According to Dretske, what a detector is supposed to indicate is what it actually indicates in the natural habitat or environment where the detector was recruited. The fly’s detector is supposed to indicate a nutritional substance because this is what it has been naturally selected for indicating and still indicates in ecologically normal conditions. When the detector misrepresents, Dretske suggests, it is because the conditions are ecologically abnormal. Notice also that Dretske is still thinking of indication in terms of reliable causation or correlation: a detector has the function of indicating X because it is (or at least was) a reliable indicator of X in the environment where it was recruited or selected. Recall that the inability of the original indicator theory to make room

---

72 Here is another passage where Dretske clearly makes these points:

Put a frog in a laboratory where carefully produced shadows simulate edible bugs. In these
for misrepresentation resulted from its requirement of reliable correlation. Dretske’s teleological modification no longer requires that a representation is reliably correlated to what it represents under all conditions. Nevertheless, the modification still requires reliable correlation in ecologically normal conditions.

Dretske’s teleological version of the indicator theory does make some room for misrepresentation, but it is not enough. The problem is that misrepresentation can occur more frequently than correct representation even in ecologically normal conditions, as Godfrey-Smith (1989 & 1992) points out. Consider the following example: “many birds have a hair-trigger flight response to dark fluttering shapes that could be predators” (Godfrey-Smith 1989: 547). The benefit of the flight response comes from helping birds to escape predators. This is why the predator-detectors have been favored by natural selection. Ancestor birds managed to escape predators by responding in this way and the trait was inherited by their descendants. The perception that triggers this response is clearly supposed to represent predators, even when it fails to do so. But we should distinguish two kinds of misrepresentation. On one hand, a bird’s predator-detector may fail to be activated when a predator is actually present. On the other hand, a bird’s predator-detector may be activated when no predator is actually present. The first case is a “false negative”, while the second one is a “false positive” (Godfrey-Smith 1992: 299). It is safer for birds to have predator-detectors that generate many false alarms than detectors that...

unnatural circumstances the frog’s neural detectors—those that have, for good reason, been called «bug detectors»—will no longer indicate the presence or the location of bugs... If we suppose... that it is the function of the frog’s neural detectors to indicate the presence of edible bugs, then, in the laboratory, shadows are misrepresented as edible bugs... Occasionally, when an edible bug flies by, the frog will correctly represent it as an edible bug, but this is dumb luck... The frog has... no reliable representation. (Dretske 1988: 68-69)
Again, what a detector is supposed to detect is determined by what it indicates in ecologically normal conditions—and misrepresentation can occur in abnormal conditions. But here Dretske clarifies that correct representation can also occur in abnormal conditions—even though it is not reliable in such conditions. The assumption again is that representation would be reliable in ecologically normal conditions.
sometimes fail to be triggered when a predator is actually menacing. As Godfrey-Smith points out, whenever the benefit of avoiding false negatives outweighs the cost of avoiding false positives, natural selection will favor detecting mechanisms that produce representations that are not reliably correlated to what they are supposed to represent—namely, mechanisms that produce many false positives (Godfrey-Smith 1992: 302-303 & 1989: 547-548). Dretske’s account ascribes the wrong meanings to the representations produced by such mechanisms. The activated state of a bird’s predator-detector is reliably correlated to dark fluttering things in ecologically normal circumstances, but presumably it represents predators.

The same problem arises in the case of representations selected by learning processes. Consider the case of the blue jay that learns to avoid eating monarch butterflies again after becoming ill. When the blue jay avoids eating a fake monarch placed by a researcher, this can be reasonably characterized as an ecologically abnormal situation. However, there are naturally occurring butterflies, such as viceroy butterflies, that mimic the appearance of monarchs. Viceroy butterflies are part of the normal ecosystem of blue jays. As Dretske himself points out, blue jays actually avoid eating viceroys after learning that monarchs are poisonous (Dretske 1988: 111). What the blue jay’s detector is reliably correlated to—what it indicates—in ecologically normal circumstances is monarchs or viceroys. Indeed, Dretske claims that, in this case, the detector represents “monarch-like butterflies” (Dretske 1988: 111). However, the purpose of the detector is specifically to help the blue jay to avoid indigestion by not eating monarchs. The avoidance of viceroys results from blue jays mistaking them for monarchs—i.e., false positives. But Dretske’s theory cannot characterize such false positives as misrepresentations. The problem arises even for detection systems that are produced by intelligent design. A house alarm system may be
triggered not only by burglars but also by insects or rodents. But it may be better to have an alarm system that produces false positives than one that may fail to detect burglars. The system may be purposefully chosen despite its known unreliability: perhaps it is less expensive than a more reliable one. The detectors are reliably correlated in normal conditions to the presence of moving objects in the house. What they are supposed to represent is the presence of burglars.

Another objection that has been raised against Dretske’s combination of indication and teleology is that indicating cannot be an etiological function. Talking about the combination of teleonomy with informational semantics, Godfrey-Smith warns: “A thing cannot have the function to have come from somewhere, to have certain normal antecedents. A thing’s function is always something that it does; a function is always something like a power.” (Godfrey-Smith 1989: 542). Similarly, Millikan claims that “it cannot be the function of a state to have been produced in response to something” (Millikan 1993: 85). If we think of indication in terms of reliable causation, the objection then can be simply expressed as follows: $R$ cannot have the function of indicating $X$ because $R$ indicating $X$ involves $R$ being caused reliably by $X$, while a function is a selected effect. If we are talking about the mechanism that produces $R$, the same consideration applies: the mechanism can indeed have the function of producing $R$, but it cannot have the function of being caused by $X$ to produce $R$. This objection works against Dretske’s way of combining indication with teleology—where indication is simply identified as the function of a representation. Neander (2013) proposes an updated combination of causal or informational semantics with teleonomy—which I discuss in Section 4.5. In reply to the objection that functions are selected effects, she suggests that functions can sometimes include causes in addition to effects. A causal version of teleosemantics, however, does not need to identify what is
supposed to cause a representation with its function. I will propose in Section 4.6 a causal variant of teleosemantics that is compatible with the standard account of functions as selected effects.

Let me recapitulate. According to Dretske’s teleological version of the indicator theory, the activated state of a detector represents what it is *supposed to indicate*. This means—given his account of “normal” conditions—that the state represents whatever it *indicates or is reliably correlated to in ecologically normal circumstances*. The problem is that this does not make enough room for misrepresentation. In particular, the account cannot characterize false positives as misrepresentations when they occur frequently in a creature’s natural habitat. Additionally, Dretske’s way of combining indication with teleology has the problem of being at odds with the etiological account of functions as selected effects. Despite these problems, I think that the idea of appealing to teleonomy in order to draw a distinction between normally and abnormally caused representation tokens is a crucial insight which is on the right track.

Other teleosemantic theories—even those that are causal or informational—differ considerably from Dretske’s. The orthodox teleosemantic approach—originally proposed by Millikan (1984) and Papineau (1984)—does not only reject the requirement of reliable causation under ecologically normal conditions, but also the requirement that tokens of representations must be caused by what they represent. An alternative approach—proposed by Neander (1995)—rejects only the requirement of reliable causation under ecologically normal conditions, but still attempts to explain the meanings of representations in terms of teleonomically characterized causal links to reality. The orthodox approach relates the meanings or contents of representations instead to the *effects* they are supposed to have. The general motivation for this is that, under an etiological account, functions are selected effects. Papineau (1984) proposes a teleosemantics
that focuses primarily on the contents of the complex mental representations involved in human belief-desire psychology. Millikan (1984) proposes a teleosemantics that starts with an account of the most basic forms of non-human representation, which she extends to cover the mental and linguistic representations used by human beings. I will discuss these applications of teleosemantics to sophisticated human representations in Chapter 6—where I borrow ideas from Papineau and Millikan, but I argue that a proper teleosemantics for human representations should be articulated in terms of basic sentence acceptance. In the following sections of this chapter, I will examine Millikan’s account of the content of basic representations (Section 5.4) and Neander’s alternative (Section 5.5). Afterwards, I will propose a hybrid account that combines some insights from these two approaches (Section 5.6). The hybrid account is intended to apply to basic perceptual representations, not the complex human representations covered by the acceptance-based teleosemantics proposed in Chapter 6.

5.4 Consumer-Based Teleosemantics

Millikan draws a distinction between the mechanisms that produce representations and the mechanisms that consume or use representations, and suggests that the latter are responsible for something being a representation and having a specific meaning or content: “What we need to look at is the consumer part, at what it is to use a thing as a representation. Indeed, a good look at the consumer part... ought to be all that is needed to determine not only representational status but representational content.” (Millikan 1993: 88). In this section I will examine the core features of Millikan’s consumer-based teleosemantics, focusing on basic non-human representations. I will argue that her account has the advantage of making ample room for misrepresentation but is
susceptible to a serious objection because it disregards the causes of representations as irrelevant to their content.

Let me illustrate Millikan’s distinction between producers and consumers of representations with a couple of simple examples. In the case of inner representations, the producers and consumers are located within a single organism. For example, the moth’s perceptual mechanisms responsible for tokening a representation when a bat is approaching are the producers, while the mechanisms responsible for triggering the bat-avoidance behavior when the representation is tokened are the consumers. In the case of outer representations used for communication, the producers and consumers are located in different organisms. When a beaver perceives danger, for example, it makes a loud sound by slapping the water with its tail as it dives for safety. The splashing sound causes other beavers in the colony to also dive under water for safety. Here the producers of the alarm signal are the mechanisms that cause the original beaver to slap its tail, while the consumers are the mechanisms that cause the other beavers to hide when hearing the sound. The consumers are the mechanisms responsible for representations having the behavioral effects they are supposed to have.

Why should we focus on the consumer mechanisms in order to explain representational status and representational content? The explanatory role of consumers in determining representational status becomes clear when we consider cases like the following:

The devices in me that produce calluses are supposed to vary their placement according to where the friction is, but calluses are not representations. The pigment arrangers in the skin of a chameleon, the function of which is to vary the
As Millikan points out, “not every device whose job description includes producing items that vary with the world is a representation producer” (Millikan 1993: 85). The mechanisms responsible for calluses and chameleons’ camouflage colors, just like the mechanisms responsible for moths detecting bats and beavers slapping their tails, produce items that covary with conditions in the external world. Yet calluses and chameleons’ camouflage colors are not representations. What sets apart the inner states produced by moths’ bat-detectors and the tail splashes of beavers from calluses and chameleons’ camouflage colors is that they are *used* as representations by *consumer* mechanisms: such mechanisms modulate *behavior in accordance with the conditions in the world that representations are supposed to covary with*. Notice that the *benefit* of having representations comes precisely from their behavioral effects. While the producer and consuming mechanisms must have evolved to work in coordination with each other—i.e., they must be co-adapted—the function of the producers must be to provide what consumers need to perform their functions (Millikan 1993: 88).

What the consumers need is representations to be tokened whenever certain conditions in the world obtain. The function of the consumer mechanisms that trigger the moth’s bat-avoidance behavior, for example, is to prevent moths from being captured by bats. That is what these mechanisms have been selected for. What they need to perform this function is that the representations they consume are tokened specifically when bats are approaching. Millikan calls this the “normal condition” for the performance of the consumers’ function. Her consumer-based teleosemantics does not *identify* meanings with the teleonomic functions of representations or
their consumers, but with the relations between representations and the world that obtain in “normal conditions”—namely, the relations that the consumers need to obtain in order to successfully perform their teleonomic functions:

Note that the proposal is not that the content of the representation rests on the function of the representation or of the consumer, on what these do... The content hangs only on there being a certain condition that would be normal for performance of the consumer’s functions, namely, that a certain correspondence relation hold between sign and world, whatever those functions may happen to be.

(Millikan 1993: 89)

Millikan’s talk of “normal conditions” differs from common usage. She does not mean the conditions that obtain on average or that are ecologically normal. Instead, Millikan clarifies, her term ‘normal’ should be understood “normatively, historically, and relative to specific function” (Millikan 1993: 86). What is “normal” in this sense is what has historically happened during the proper performance of a function. Whatever happened when an item failed to perform its function is not “normal” in this sense, even if the item failed to perform its function more often than it succeeded. Millikan uses the term ‘normal’ in this normative-historical sense to characterize the etiological explanation of the performance of a function, as well as the conditions that must be invoked in such an explanation:

A “normal explanation” explains the performance of a particular function, telling how it was (typically) historically performed on those (perhaps rare) occasions when it was properly performed. Normal explanations do not tell, say, why it has
been common for a function to be performed; they are not statistical explanations. They cover only past times of actual performance... A “normal condition for performance of a function” is a condition, the presence of which must be mentioned in giving a full normal explanation for performance of that function...

It follows that “normal conditions” must not be read as having anything to do with what is typical or average... (Millikan 1993: 86-87)

In order to avoid confusion with the common usage of the word ‘normal’, I will use instead the uppercased word ‘Normal’ to talk about this normative-historical notion, as Millikan does herself in some of her writings (e.g., Millikan 1984). Consider again beavers’ alarm signals. The function of the consumers of these representations is to protect beavers from danger. The Normal explanation of how this function was historically performed, when it was performed properly, must mention that perception of the signal made beavers instinctively dive under water (a complete explanation of this will include an account of how the relevant mechanisms worked). The Normal explanation must also mention that a crucial Normal condition for the consumers performing properly their function was that the representations were tokened when there was actual danger, since only on those occasions diving under water was a beneficial behavioral response to the representation. In evolutionary terms, the cooperating producer and consumer mechanisms of beaver tail-splashes were favored by natural selection because occasionally they actually helped to protect beavers from danger. Since the crucial condition that must obtain in the world for the consumers of beaver tail-splashes to perform their function is that there is danger, it follows—according to Millikan—that tail-splashes mean danger:

The splash means danger, because only when it corresponds to danger does the
instinctive response to the splash on the part of the interpreter beavers, the consumers, serve a purpose. If there is no danger present, the interpreter beavers interrupt their activities uselessly. Hence, that the splash corresponds to danger is a normal condition for proper functioning of the interpreter beavers’ instinctive reaction to the splash. (Millikan 1993: 90)

Like the standard indicator theory, Millikan’s teleosemantics explains the meaning of a representation in terms of its direct links to what it represents. Like Dretske’s teleological version of the indicator theory, it explains more specifically what a representation is supposed—but may fail—to represent in terms of an etiological account of what it is supposed to covary with, be correlated to, or correspond to in the world. But Millikan’s consumer-based teleosemantics does not place any requirement of reliability under ecologically normal conditions on the production of a representation. Consequently, it can make ample room for misrepresentation. Suppose, as it is likely, that the majority of beaver tail-splashes turn out to be false positives—produced when there is no real danger around. This does not impact at all what the tail-splashes are supposed to represent according to Millikan’s account. As long as the trait has been selected and maintained in the population because it aided the survival of beavers by making them dive under water for safety when there was danger—and even if this happened only on rare occasions—the splashes will mean danger. When there is no real danger around, a tail-splash is a misrepresentation. Millikan’s teleosemantics can account for most tail-splashes in ecologically normal conditions being misrepresentations. Notice that, for many biological devices, successful performance of their functions is something that rarely occurs but explains why the devices exist. Most sperm cells, for example, do not succeed in fertilizing egg cells. Nonetheless, this is precisely what they
have been selected for doing (Millikan 1984: 4). That an item may fail to perform its function more often than succeeding is not peculiar to representations.

Another advantage of Millikan’s consumer-based teleosemantics is that it offers a plausible account of how representations can represent distal affairs rather than the proximal stimuli involved in their deployment. Consider male hoverflies, which wait for females to mate with by hovering in a single place for long periods of time. When they perceive a female, they instantly dart towards it. This behavior conforms to a very precise rule that specifies the angle of the chasing flight in relation to the vector angular velocity of the target’s image moving across the hoverfly’s retina (Collett and Land 1978). Since this rule specifies how the hoverfly is supposed to react to a proximal stimulus, rather than a distal object, Millikan calls it the “proximal hoverfly rule” (Millikan 1993: 218-219). As Millikan points out, while it is unlikely that the hoverfly “calculates over any inner representation of this rule in order to follow it”, it still has a “genetically determined mechanism of a kind that has historically proliferated in part because it was responsible for producing conformity to the proximal hoverfly rule, hence for getting male and female hoverflies together” (Millikan 1993: 219). Given the selective history of these mechanisms, they are supposed to conform to the proximal hoverfly rule. Why not regard the representations consumed by these mechanisms then as representations of the proximal stimuli—i.e., images—rather than of distal female hoverflies? Images produced by all sorts of distal objects, such as midges or distant birds, trigger the chasing behavior in conformity

73 Following Devitt, we can say that this is a structure rule “governing the outputs of a competence”, rather than a processing rule “governing the exercise of the competence” (Devitt 2006: 18). Whether a rule is a structure rule or a processing rule depends on whether the rule governs the exercise of the competence: whether “following the rule” involves “calculating over its representation”. This distinction should not be conflated with the one between proximal and distal rules. As I point out below, the crucial rule governing the hoverfly’s behavior is one that relates the hoverfly to distal affairs. This is a distal structure rule.
to the rule more often than actual female hoverflies (Collett and Land 1978). Yet the mechanisms seem to be working in accordance with their evolutionary “design”, rather than malfunctioning, whenever they follow the proximal rule. Millikan argues that the representations consumed by these mechanisms are nevertheless representations of distal female hoverflies, because the mechanisms are also supposed to make the male hoverfly follow the crucial rule: “If you see a female, catch it.”—which she calls the “distal hoverfly rule” (Millikan 1993: 222). The etiological explanation of why the mechanisms exist relies on the occasional past conformity of the behavior they produce to the distal rule, since “it was only when ancestor hoverflies conformed to the distal hoverfly rule that they became ancestors” (Millikan 1993: 223). The proximal rule alone cannot explain why the mechanisms exist. What the mechanisms needed to perform the function that explains why they were selected is that the representations they consumed covaried with actual female hoverflies. That is why the representations are representations of distal female hoverflies rather than of proximal images. But what is the role of conformity to the proximal rule which, Millikan acknowledges, the mechanisms are also supposed to follow? Millikan suggests that “conformity to the proximal rule is a means to conformity to the distal rule” or, in other words, “the mechanism that has historically accounted for... ancestor hoverflies’ successes at conformity to the distal hoverfly rule begins with conformity to the proximal rule” (Millikan 1993: 222-223). The idea is that the function of conforming to a proximal rule is part of the Normal explanation of how the mechanism that consumes an inner representation properly performs a function that involves conformity to a distal rule. Whenever an inner representation is supposed to be used in conformity with certain proximal rule in order to—or as a means to—be used in conformity with certain distal rule, what
it represents is what it must be related to in the external world for its use to succeed in conforming to the distal rule.

Millikan’s teleosemantics defines representational content in terms of the historically Normal conditions for the proper performance of the functions of representation consumers. This allows the approach to make ample room for error by not imposing any reliability requirement on the production of representations—not even under ecologically normal circumstances. Millikan goes a step further, arguing that her account does not even require representations to be caused by what they represent. She offers the following example—originally discussed by Dretske (1984)—to clarify this point. There is “a certain species of Northern Hemisphere [anaerobic] bacteria which orient themselves away from toxic oxygen-rich surface water by attending to their magnetosomes, tiny inner magnets, which pull toward the magnetic north pole, hence pull down” (Millikan 1993: 92-93). The function of the consumers of the representation produced by the magnetosomes is to guide the bacteria towards safe oxygen-free water—this is the beneficial effect that they were selected for. While the Normal explanation of how the function is performed will mention magnetosomes pulling towards the magnetic north pole and bacteria swimming in that direction, the crucial Normal condition for the consumers to perform properly their function is that the magnetosomes pull towards oxygen-free water. According to Millikan, then, what the magnetosomes represent is “the direction of oxygen-free water”, even though this “is not... a factor in causing the direction of pull of the magnetosome” (Millikan 1993: 92-93). The “Normal mechanisms” (Millikan 2004a: 76-79) by which producers perform their functions—i.e., the mechanisms invoked in a Normal explanation of these performances—may often rely on causal links between representations and what they represent. But, according
to Millikan, this is not a requirement and anyway does not determine the meanings of representations: what matters is *that* the producers Normally token representations that are correlated to what they represent, not *how* they manage to do so. This is a strong and controversial claim that can be reasonably contested.

An influential objection against Millikan’s teleosemantics shows that her view that the causes of representations are irrelevant to their content is problematic. Paul Pietroski (1992) proposes a thought experiment involving two imaginary animal species: kimus and snorfs. Kimus live in a valley where snorfs, their main predators, hunt during the day. Kimus are utterly unable to recognize snorfs, but they have evolved both a mechanism that detects red things and the instinctive behavior of moving towards them, which makes kimus climb a nearby hill every morning, attracted by “something red on the hilltop”—a trait that has been favored by natural selection because climbing the hill has helped kimus to avoid being hunted by snorfs (Pietroski 1992: 273). According to Millikan’s theory, the activated state of the detectors represents snorf-free areas, since the behavior produced by its consumers was beneficial only because it led kimus to such areas. But Pietroski argues that this content ascription is unacceptable. He points out that we ascribe “intentionality” (i.e., aboutness) or representational content to mental states in order to explain behavior: “For what are intentional states, if not those states that figure in (correct) intentional explanations?” (Pietroski 1992: 276). While a relation to snorfs is a crucial part of the evolutionary or selection explanation of why kimus have the trait, Pietroski argues, it is not part of a plausible intentional explanation of kimu-behavior (Pietroski 1992: 274-276).

It does seem counter-intuitive to claim that kimus climb the hill because they represent it as snorf-free. The content ascribed by Millikan’s theory seems to conflict in this case with the
one we would ordinarily ascribe—based on the perceptual causes and behavioral effects of the representation (see Chapter 2)—and likely also with what a scientific intentional psychology would ascribe. This is because, as Pietroski points out, kimus cannot “discriminate snorfs from non-snorfs” (Pietroski 1992: 276) and their representations are tokened “in the absence of any causal interaction” with snorfs (Pietroski 1992: 277). Pietroski acknowledges that the correct theory of content may turn out to be “somewhat revisionary of our common-sense practice of providing intentional explanations”, but he warns that Millikan’s theory is “radically revisionary” and that she does not provide any “independent motivation”—apart from her own theory—to support such revision and to regard the evolutionary relation between kimus’ states and snorfs as constitutive of the states’ content (Pietroski 1992: 277). Notice that this last concern of Pietroski is in tune with the methodological point I made in Chapter 2—following Devitt (1996)—about the task of semantics: we need an independent and uncontroversial identification of the meanings that need explanation. Otherwise, it is unclear what would count as a successful explanation and whether allegedly competing semantic theories are actually attempting to explain the same phenomena. I will get back to this methodological issue in Section 4.6.

Millikan agrees that her theory implies that kimus represent snorffree areas, but rejects the intuition that this is problematic. She points out that certain desert tortoises find their food (edible vegetation) by its green color, which is merely correlated to, and not caused by, the nutritional characteristics that make it food—characteristics that tortoises are unable to discriminate (Millikan 2000: 236). Millikan’s reply is that, according to Pietroski’s criteria, the representations that tortoises use as guides to find food cannot be representations of where there is food. If we accept that tortoises have representations of where there is food, Millikan argues,
we should revise the intuition that supports Pietroski’s objection.

It is not clear that the tortoises case can settle the issue in favor of Millikan’s view, since these representations can be beneficial because they direct tortoises to take in nutrients without being representations about the nutritious characteristics of what tortoises eat. Perhaps they are representations of “tortoise food”—in the sense of “what tortoises eat” rather than of “what is nutritious for tortoises”—or perhaps they are simply representations of green stuff. I prefer the former option because it fits well with the eating behavior guided by the states and because the latter option would imply that a tortoise eating a green plastic bag—left behind in the desert by a hiker—would not be misguided by a false positive. There is nothing peculiar about a representation of “food” in the proposed sense. To put it metaphorically, the representation can tell the tortoise something like “this is what you can eat” rather than “this is nutritious”. It is doubtful that animal representations of food represent it qua nutritious stuff.

Another important point is that there would be no causal interaction between tortoises’ representations and what they represent only under Millikan’s assumption that they represent food qua nutritious stuff. If we take the tortoises’ states to represent green stuff, the causal connection is clear. But there is also a causal connection if we take the states to represent “tortoise food”. In the sense suggested above, “tortoise food” is green (even though its nutrients are not green and are not responsible for its color). So, the presence of “tortoise food” causes the tokening of tortoises’ representations—because it is green. In contrast, neither snorfs nor their absence are causally involved in the tokening of kimus’ representations.\footnote{There is considerable philosophical debate about what is the nature and what are the relata of causal relations. Perhaps under some views it is conceivable that snorf-free areas cause kimus’ states. But the view of causation that Millikan relies on in her reply to Pietroski is not one of them. Millikan relies on the notion that $A$ can cause $B$ because $A$ is $F$ rather than because $A$ is $G$—even if $A$ is both $F$ and $G$. My reply simply points out that in such a case $A$ still causes $B$.} Millikan is counting on
her readers agreeing that tortoises have perceptual states that represent food but are not caused by food. But the sense of ‘food’ in which we can correctly say that tortoises represent food is not necessarily the same as that in which we can correctly say that tortoises’ representational states are not caused by food.

Another reply to Pietroski offered by Millikan relies on rejecting altogether the assumption that creatures can only have representations of what they are able to discriminate. Millikan complains that “a creature can perfectly well have a representation of Xs without being able to discriminate Xs from Ys” (Millikan 2000: 237). This reply has an important kernel of truth, but I do not think that it suffices to undermine Pietroski’s objection. What is true in Millikan’s claim is that the strong discrimination view it opposes is very problematic. The problematic view is that a creature cannot have a perceptual representation of Xs if it is not able to discriminate Xs from some (perhaps many) non-Xs. Such a view does not make enough room for misrepresentation. False positives occur precisely because of the limited discriminatory capacities of the perceptual mechanisms involved. However, the problematic view should not be confused with the reasonable view that a creature cannot have a perceptual representation of Xs if it is not able to discriminate Xs from any non-Xs whatsoever. Now consider Pietroski’s claim that kimus cannot represent snorfs because they cannot discriminate snorfs from non-snorfs. If the claim is based on kimus not being able to discriminate snorfs from some non-snorfs, then the view is implausible because it does not make room for false positives. Pietroski does talk about kimus not being able to “reliably discriminate” snorfs from non-snorfs (Pietroski 1992: 276), but his argument does not depend on this. Much more relevant is the fact that kimus cannot discriminate snorfs from any non-snorfs whatsoever. So Millikan’s fair point against an
unreasonably strong discrimination view does not defend her theory from Pietroski’s objection. The content ascription of Millikan’s theory in the kimus case violates even a minimal and seemingly reasonable discrimination requirement.

Millikan’s focus on the role of representation consumers allows her to offer an account that makes plenty of room for misrepresentation. But her account is susceptible to Pietroski’s objection because it disregards the role of representation producers as irrelevant to content. It should be noticed that Millikan (1984) originally emphasized equally the functions of consumers and producers and that she has later returned to a more balanced focus, insisting that consumers and producers have been “designed” or selected for cooperating with each other: “What the consumer does helps the producer, and what the producer does helps the consumer... The presence of each is part of the normal mechanism by which the other fulfills its functions.” (Millikan 2004a: 79). Beyond these matters of emphasis, Millikan has not changed her mind about the irrelevance of the causes of representations. Her motivation seems to rely on her view that functions are a matter of effects: “Taking the... example of the firing of a fly detector in a frog’s eye, it couldn’t be a function of that firing to have been caused by a fly.” (Millikan 2004a: 68). In contrast, Neander has proposed that perceptual states represent what the mechanisms that produce them have the function of detecting or responding to, when analyzed at the lowest or most “mechanical” teleo-functional level (Neander 1995: 129-136). Neander’s view is that the mechanism involved in a representation’s production—in how the representation is caused—is what is relevant to the representation’s content. Unlike Dretske’s teleological version of the indicator theory, Neander’s teleosemantics does not depend on what typically causes representations in ecologically normal conditions. Unlike Millikan’s teleosemantics, it does
explain meaning in terms of causation.

5.5 Producer-Based Teleosemantics

Neander points out that biological traits have a hierarchical array of functions and argues that, in the case of systems that produce representations, the lowest-level functions—involving causal links between representations and what they represent—are the ones that determine representational content. In this section, I will examine the core features of Neander’s producer-based teleosemantics and argue that it has the advantage not being susceptible to Pietroski’s objection, due to the role it assigns to the causes of representations, but also that it has the disadvantage of not making enough room for misrepresentation, due to its treatment of the effects of representations as irrelevant to their content.

Biological traits typically have various effects. Let us start with a non-representational example: “Suppose a trait in an antelope population altered the structure of the hemoglobin, which caused higher oxygen uptake, which allowed the antelope to survive at higher ground to which they were forced to retreat. Suppose also that, as a result, this trait was selected.” (Neander 1995: 114-115). Which among all of these effects of the trait is its etiological function? According to Neander they all are, because the explanation of why the trait exists must mention them all. As Neander points out, however, the various functions are not independent of each other, since the trait (1) contributed to gene replication by (2) allowing the antelope to survive at higher ground by (3) increasing oxygen uptake by (4) altering the structure of the hemoglobin (Neander 1995: 114-115). The various functions are hierarchically related, each of them being performed by means of the performance of the function in the following “lower” level—where
the lower level specifies the *underlying mechanism* responsible for *how* the higher-level function is performed and the higher level specifies the *purpose* for which the lower-level function is performed (Neander 1995: 116). While *all* the levels describe functions of the trait, Neander argues that the lowest level describes the function that is *specific* to the trait, since the trait itself is *malfunctioning* only if it fails to perform this function (Neander 1995: 119-120). A failure in the increase of oxygen uptake may occur in a certain antelope because its lungs are malfunctioning, even if the structure of hemoglobin has been properly altered (Papineau 1998: 4). But the trait in question is itself malfunctioning if it fails to properly alter the structure of hemoglobin. So, according to Neander’s account, altering the structure of hemoglobin is the specific function of the trait.

Neander applies these considerations to the functions of traits involved in the *production* of representations. Consider the magnetosomes which are part of the phenotypes of the anaerobic bacteria discussed above. What are their functions? What did the magnetosomes of ancestral bacteria did that explains why they exist today? The magnetosomes *(1) contributed to gene replication by *(2) helping bacteria to survive by *(3) guiding bacteria towards oxygen-free water by *(4) aligning themselves with the magnetic north pole. According to Neander, all of these are functions of the magnetosomes, but only the last one is their specific function, since the performance of higher-level functions—e.g., guiding bacteria towards oxygen-free water—depends on the magnetosomes’ “collaboration with other components” (Neander 1995: 118). The *other* components would include, most crucially, the *consumers* of the representation.

Now consider the case of frogs, which respond to the visual perception of their prey by darting their sticky tongues to capture and eat it. Their prey includes insects, spiders, worms and
all sorts of living animals small enough to fit into their mouths. Frogs respond in this way not only to their prey, but also to any small and dark thing moving in front of them. According to Millikan’s teleosemantics, the frog’s perceptions represent *frog food*, since it was only when the representations of ancestor frogs corresponded to edible things that their consumers performed the functions that explain why both the consumers and their cooperating producers where favored by natural selection. But according to Neander’s teleosemantics, the frog’s perceptions represent instead *small and dark moving things*, because detecting such things is the lowest-level and specific function of the perceptual mechanism involved: “it is by detecting small dark moving things that the frog detects frog-food” (Neander 1995: 130). Suppose that a frog snaps at a small and dark leaf carried by the wind. According to Neander, the detectors are not *themselves* malfunctioning in this case, even though they are failing to perform the higher-level function of detecting frog food. The detectors have not been “designed” to tell apart prey from small and dark moving leaves. In this case there is *no misrepresentation* according to Neander’s teleosemantics (Neander 1995: 129). This is a potentially problematic feature of the account. I will I get back to this later in this section.

A crucial virtue of Neander’s theory, she argues, is that it ascribes contents which are related to the cognitive capacities of the representation-producing mechanisms—to how they work when they function properly: “Low Teleology [i.e., Neander’s theory] has something to say about the criteria by which a cognitive system identifies that which it represents, and how it represents it, whereas High Teleology [e.g., Millikan’s theory] has nothing to say about this” (Neander 1995: 134). Recall that in Millikan’s theory what matters regarding the producer mechanisms is merely *that* they token representations when the Normal conditions for the
performance of the consumers’ functions obtain—and not how the producer mechanisms are supposed and able to do it. Because of this, Millikan’s theory regards the causes of representations and the discriminatory capacities of perceptual mechanisms as irrelevant to representational content. But this is precisely why her theory can sometimes ascribe implausible contents. As Pietroski’s kimus’ case illustrates, it can ascribe perceptual representations about things that the perceptual mechanisms are causally insensitive to and unable to discriminate from anything else. Pietroski suggests that his objection undermines teleosemantics in general (Pietroski 1992: 278). But Neander points out that this is not the case, since the objection applies only to the “High Church” variety (Neander 1995: 127). Indeed, Neander’s teleosemantics does not ascribe counter-intuitive contents in the kimus case. The kimus’ trait did evolve because it helped kimus to survive by avoiding snorfs. While this is a crucial biological function of the kimus’ trait, it is not the function that determines the content of kimus’ representation, according to Neander’s theory. What the perceptual mechanism is designed to causally respond to or detect is specifically red stimuli. So the resulting perceptual states represent red things, not snorf-free areas. This indeed is an advantage of Neander’s account over Millikan’s.

Neander’s analysis of the relation between the higher-level and lower-level functions of traits is illuminating. Her teleosemantic account—which focuses on the functions of the mechanisms that produce representations—has the advantage of ascribing contents that are related to the capacities of such mechanisms. But there are some important questions or concerns that may be raised about her account. Does it truly respect the requirement that etiological functions must be selected effects? Does it actually allow for the representation of distal affairs? And does it make enough room for misrepresentation? I will address each of these questions,
starting with the first one. Regarding this issue, Papineau objects:

Neander has taken insufficient note of Millikan’s point that representational content hinges on how the representation is used, not on what causes it. In her general discussion of teleology, Neander focuses, quite rightly, on the effects of biological traits. But as soon as she turns to representation she shifts to the question of what is supposed to cause the frog’s state (what it is supposed to detect). (Papineau 1998: 7)

Neander’s account is committed to the view that the content of a representation is whatever is supposed to cause it—which, in Papineau’s view, is at odds with the etiological account of functions as selected effects. This is similar to the objection raised by Godfrey-Smith (1989) and Millikan (1993) against earlier attempts to articulate an indicator or informational version of teleosemantics, including Dretske’s: an item cannot have the function of being caused by—or being produced in response to—something (see Section 4.3). Neander replies that while it is true that functions must involve effects—since this is what is required for natural selection to occur—they can also involve causes: “Biological mechanisms are selected for their causal roles, which can include dispositions to respond to specific types of causes.” (Neander 2013: 24). Going against teleosemantic orthodoxy, Neander suggests that sensory systems have response functions: functions to produce states in response to certain input conditions—where ‘response’ is understood in “purely causal” terms (Neander 2013: 23). Her proposal is that a sensory state indeed represents what its producer mechanism has the function of causally responding to, given its evolutionary “design”.
Neander’s proposal that sensory systems have response functions has some intuitive appeal. Nonetheless, the proposal is indeed at odds with the standard etiological account of functions as selected effects—which Neander (1991) herself helped to articulate (see Section 3.5). It is no surprise that Millikan disagrees. She objects that a sensory system cannot be selected for producing states in response to certain causal inputs—since this cannot be a selected effect—and proposes that the causal relations that Neander has in mind are instead part of the Normal explanation of how sensory systems often are supposed to work to perform their functions (Millikan 2013a: 37-40). The idea here is that a sensory system may be selected because it produced states that were caused by certain causal inputs, but this is not what it has been selected for doing: it is not its purpose or function. Millikan’s suggestion is that we can account for the explanatory role of the relevant causal relations—when a sensory system relies on them to perform its function—without abandoning the standard etiological account of functions as selected effects. But let us consider an example that Neander provides to support her view that items can have response functions:

Now think about those biological functions that depend on natural selection. One type of mechanism might secrete melatonin in response to the dimming of light, while another type of mechanism might secrete melatonin in response to light brightening and a third more or less randomly. Since melatonin makes us sleepy, the first mechanism might be more adaptive in creatures like us who have poor night vision. So the first type of mechanism might be selected in preference to the second and third because the first differs in its input cause, in what triggers its production of melatonin. (Neander 2013: 24)
It sounds initially reasonable to claim that the selected mechanism in this example was selected for secreting melatonin in response to the dimming of light. Such a function would be a selected disposition or causal role, rather than merely a selected effect. There is no doubt that there is an important difference between the three mechanisms in this example and that the difference is responsible for one of them being favored by natural selection. The way Neander describes it, the difference is merely about input causes, since the three mechanisms have the same effect: secreting melatonin. However, the standard account of functions as selected effects can handle this case without problems. The effects of the three mechanisms can be described differently: secreting melatonin at night, during the day or randomly. So there is no problem omitting the cause in the characterization of the function of the selected mechanism. The cause can be accommodated instead as part of the Normal explanation of how the selected mechanism manages to secret melatonin at night: the mechanism performs this function by causally responding to the dimming of light.

It looks like the selected effects view can account for the phenomena that may intuitively seem to have response functions. We should remain open-minded, but the selected effects view is the standard etiological account of functions—the status quo—so we would need compelling reasons to abandon it and adopt Neander’s revisionary proposal. Furthermore, the selected effects view presumably has an advantage in those cases that may seem to involve response functions. Suppose that the mechanism discussed by Neander happens to secret melatonin certain night by accident—i.e., not in response to the dimming of light. There is a sense in which the mechanism did perform its function, albeit accidentally. As Millikan points out: “functions are sometimes performed yet performed ab Normally” (Millikan 2013a: 40). The selected effects
view has no problem characterizing this situation as I did. The function of secreting melatonin at night was performed, but Normally because the dimming of light is what Normally causes the secretion—i.e., what is supposed to cause it. The response function view, however, is unable to characterize the situation this way: the function of secreting melatonin in response to the dimming of light was simply not performed. So this view has a problem allowing for alleged response functions to be performed accidentally. While I am skeptical about the need for response functions, I am not skeptical about Neander’s view that the content of a representation depends on what is supposed to cause it. I will get back to this in the next section.

Moving on to another issue, Millikan has also objected that Neander’s teleosemantics treats all representations as representations of proximal stimuli (Millikan 2000: 231). But Neander has recently offered her own solution to the distality problem. Her proposal is that a representation $R$ refers to something distal $D$ rather than something proximal $P$ (when $P$ is an intermediary link in the causal chain between $D$ and $R$) if and only if the sensory system has the etiological function of producing tokens of $R$ in response to $D$ by means of producing tokens of $R$ in response to $P$ (Neander 2013: 34). Notice that this suggestion is analogous to Millikan’s solution to the distality problem, but articulated in terms of response functions. In both cases, a relation between a representation and proximal stimuli is regarded merely as a means to perform a function involving a relation between the representation and something distal. However, Neander’s solution seems to be incompatible with her own claim that the content of a representation is determined by the lowest-level mechanical function of the system that produces it. Consider Neander’s claim that the frog’s perceptual states represent small and dark moving things, rather than prey or frog food. With her new solution to the distality problem, she is able to
argue that the frog’s states represent distal things, since they are the products of sensory systems that have (1) the function of responding to *distal* small and dark moving things, which is performed by means of performing (2) the function of responding to *proximal* small and dark moving images. But if we must accept that a higher-level function determines content in order to solve the problem of distality, why not regard this higher-level function as the one performed when the frog darts its tongue towards its *prey* or *food*? Neander’s solution to the distality problem requires abandoning her “lowest level of analysis” account of representational content. So that analysis can no longer provide a reason for favoring her claim that frogs’ perceptions represent small and dark moving things rather prey or food.

In fairness, Neander can still appeal to the fact that the frog does not have the capacity to distinguish its prey from other small and dark moving things, even when its visual systems are working properly: the perceptual systems are not themselves malfunctioning when the frog darts its tongue towards a moving leaf. But this brings us to the remaining question I raised about Neander’s teleosemantics: can it make enough room for misrepresentation? Neander’s causal or informational teleosemantics, unlike Dretske’s combination of indication and teleology, does not depend on reliable causation or correlation under ecologically normal circumstances. Yet her account faces similar problems. Presumably the representation that causes the frog to dart its tongue towards a moving leaf is a *false positive*—just like the beaver’s tail splash produced in the absence of danger or the bird’s representation of predators tokened in the absence of predators. For frogs, the benefit of not missing actual prey outweighs the cost of darting their tongues towards non-edible targets, so natural selection has favored a perceptual mechanism that cannot discriminate between actual prey and a number of other things, but is quite effective at
allowing frogs to catch their prey, which is its purpose. According to Neander’s teleosemantics, however, the frog’s representation triggered by a small and dark moving leaf is a true representation instead of a false positive: “I suggest that the frog doesn’t misrepresent as long as its representation... is caused by something appropriately small, dark and moving” (Neander 1995: 129). The leaf example is mine, but Neander remarks that according to her theory a frog responding to a “bee bee” (a small metallic projectile shot by an air gun) is not misrepresenting (Neander 1995: 131).

I should clarify that Neander’s teleosemantics does make some room for misrepresentation. The frog does misrepresent when it responds to “anything which reflects onto its retina a pattern that falls outside of the specified parameters” of size, shade and movement—which can happen, she points out, if the frog is sick, neurologically damaged or has a congenital defect (Neander 1995: 131). Neander’s account makes room for misrepresentation in cases where the perceptual mechanism fails to perform its function because it malfunctions. The problem, however, is that misrepresentation often occurs when the mechanism fails because “the world is deceptive”—to use Godfrey-Smith’s expression. As Millikan reasonably argues against Neander (2013), there can be misrepresentation without malfunctioning: this occurs when the Normal condition in the external world required for the representation producer to perform its function does not obtain (Millikan 2013a: 40). Neander’s teleosemantics does not make room for this kind of misrepresentation and this is a serious drawback. Given the limited discriminatory capacities of male hoverfly mate-detectors, of beaver danger-detectors, of bird and moth predator-detectors—just to mention some other examples discussed above—it turns out that the states produced by these devices do not represent mates, danger or predators, according to Neander’s
teleosemantics, despite their guiding mating, hiding and avoidance behavior.

Dretske’s original attempt to combine informational semantics with teleology was intended to make room for misrepresentation. It was a nice try. But it did not make enough room. Neander’s updated attempt has virtues that I have pointed out. However, it shares with Dretske’s attempt the problem of not making enough room for misrepresentation. Perhaps it makes even less room, since Dretske does not take such a restrictive approach regarding the discriminatory capacities of creatures. Frogs, he says, represent bugs (Dretske 1988: 68). And I already quoted him claiming that flies represent nutritional substances by means of detecting their sweet flavor—because in their natural habitat “the [sweet tasting] substances that stimulate the receptors are nutritional” (Dretske 1988: 58). Under Neander’s account, these receptors would only represent sweet substances—and Millikan seems to agree that a causal theory would be forced to adopt such a restrictive view of causal connections. Perhaps Dretske is mistaken to claim that flies represent nutritional substances as such. Still, the representations do tell flies to eat these substances—not just that they are sweet. So a more plausible content ascription may be that they represent edible stuff or “food”—in the sense of something to get a hold of and eat, rather than something nutritious. In the analogous case of carnivorous hunters or predators, we have a better word for this: ‘prey’.

Back to Dretske, it is worth mentioning that—even in his early version of the indicator theory—he allows for more complex informational groundings than Neander. Notice that Neander does not limit herself to simple causal links: $R$ can represent $X$ even when they are not directly connected by a single causal link, but rather by a chain of causal links: this is what her solution to the distality problem is meant to allow for. But Dretske allows for other variants of
what we may call “causal groundings”. Suppose, for example, that something \( X \) causes both \( Y \) and \( Z \), but \( Y \) and \( Z \) are not causally linked otherwise. In this case, Dretske proposes, \( Y \) and \( Z \) carry information about each other due to their common cause (Dretske 1981: 38-39). It follows that a perceptual system can manage to indicate \( Z \) by detecting \( Y \) even if it is not causally sensitive to \( Z \) itself. If these more complex forms of causal grounding do work, then it may be the case that flies do have representations of nutritious stuff. The presence of nutrients and of sweet tasting molecules in a fruit, for example, surely are connected by a common and complex causal process. The same would apply to the green color and nutrients of the edible vegetation that tortoises eat—recall Millikan’s reply to Pietroski. These are complicated matters, but it certainly looks like adopting a restrictive view of causal connections is not the only reasonable option for a naturalized semantics. It should be noticed, however, that Dretske goes beyond “causally grounded” informational relations. His definition of indication relies on the conditional probability of \( X \) happening if a representation \( R \) is tokened (Dretske 1981: 65). Such correlations may often be the result of causal connections, which can be found in most of Dretske’s examples, just like in most of Millikan’s. But they are not strictly required. That is why Dretske suggests that the magnetosomes of northern anaerobic bacteria may represent oxygen-free water, even though there is no causal connection between the magnetic north they detect and oxygen-free water (Dretske 1986: 26). On this issue Dretske and Millikan coincide. But this sort of view leads to the problem highlighted by Pietroski’s though experiment.

---

75 Dretske argues that it is indeterminate whether the magnetosomes represent the magnetic north or oxygen-free water (Dretske 1986: 28-32). The lack of a causal connection in one of the two alternatives does not motivate him to prefer the other.
5.6 A Hybrid Account

Neander’s producer-based teleosemantics avoids Pietroski’s objection because it does relate content to the cognitive capacities of the mechanisms that produce representations, but cannot characterize many false positives as genuine misrepresentations: those that occur because “the world is deceptive” rather than because of internal malfunctioning. Millikan’s consumer-based teleosemantics can properly characterize such false positives as misrepresentations because it relates content to the Normal conditions for the performance of the functions of the mechanisms that consume representations—i.e., that guide behavior based on them—but cannot avoid Pietroski’s objection. In this section I will propose a hybrid account that relates content to the roles of both producers and consumers of representations. The hybrid account promises to make ample room for misrepresentation while making plausible ascriptions of content that take into account what the perceptual mechanisms of a creature do causally interact with—thereby avoiding Pietroski’s objection. While the hybrid account makes ascriptions that are in tune with Pietroski’s plausible intuitions, it should be noticed that my case for the approach rests mainly on methodological considerations about what phenomena it is the task of semantics to explain.

The hybrid account is meant to apply primarily to “basic representations” that are caused by perceptual mechanisms and directly guide behavior: representations of the kind that have been the focus of this chapter. The approach also has an application for the simplest perceptual representations of cognitively complex creatures such as human beings. But those

---

76 My hybrid account may look similar at first sight to Nicholas Shea’s suggestion that Millikan’s teleosemantics should be supplemented with an input condition (Shea 2007). Shea adds an input condition to Millikan’s output-based account. I agree that we need a teleosemantics that takes into account both inputs and outputs. But the hybrid account I am proposing differs in a crucial respect from Shea’s: while his input condition involves mere informational correlations between representations and what they represent, my hybrid account relies on causal links—like Neander’s. Since Shea’s input condition is not causal (Shea 2007: 16), his proposal shares problems with Millikan’s. In the kimus’ case, it also entails that their states represent snorl-free areas.
representations differ in crucial ways from what I am treating here as basic representations, since they typically do not directly govern behavior and their consumers are likely to include a variety of mechanisms, including those that token more complex representations such as thoughts and concepts. I will focus here on “basic representations”. My proposal in a nutshell is the following:

**Hybrid Producer-Consumer Teleosemantics:** The content of a basic representation is determined by what is supposed to cause tokens of it—what its producer mechanisms are supposed to respond to—in order for its consumer mechanisms to perform their function in accordance with a Normal explanation.

Like Neander’s account, the hybrid account relates the content of a basic representation to what is supposed to cause it: the links involve connections to things that the perceptual system of a creature does causally interact with. Like Millikan’s account, the hybrid account relates the content of a basic representation only to those links to reality that the representation tokens have on those occasions when the consumers succeed performing the functions that they have been selected for. These functions are selected effects, as conceived by the standard etiological account of functions. In the case of basic representations, the functions are to produce behaviors that are appropriate to—and vary accordingly with—the creatures’ circumstances.

I pointed out in the last section that I agree with Neander’s view that the content of a basic representation depends on what is supposed to cause it, but I explained why I am skeptical about the need for response functions. How can we get the “normative” notion of being “supposed to be caused” by something without identifying this as a function of a mechanism? Millikan’s distinction between what is part of a function and what is part of the Normal
explanation of how the function is performed provides the answer. It allows us to extend the “normativity” conferred by the teleonomic account of functions to the mechanisms and conditions involved in their performance. Kidneys are *supposed* to filter blood because this is their function. They are also *supposed* to be connected in certain ways to the circulatory and renal systems. But this is not something they *do*, not an *effect* that they have been selected for. Instead, it is a Normal condition for their managing to filter blood in the way that historically accounts for their existence. The same applies to the causes of perceptual representations under Millikan’s analysis: such representations are *supposed* to be caused by certain distal events in order for their producers and consumers to perform their functions. Millikan, of course, thinks that such causal relations are not relevant to the representations’ contents. But notice that what Millikan does identify as the contents of representations—namely, certain crucial Normal conditions for the performance of the consumer’s functions—are not part of a function either, but of a Normal explanation. The fact that causes are not a part of the functions of the producers or consumers of representations does not rule them out as candidates for content determination in a teleosemantic theory.

What kinds of content ascriptions are provided by the hybrid account? I will discuss the cases of frogs and kimus below. Let me start with other two examples. According to the hybrid account, the states produced by male hoverfly mate-detectors and by moth predator-detectors represent mates (i.e., female hoverflies) and predators (i.e., bats), respectively, since the consumers of the representations perform Normally their behavior-governing function of guiding these creatures to mate or avoid predators when the representations are caused by mates or predators. The states are *supposed to be caused* by mates and predators, respectively, so they
misrepresent when triggered by things that are not mates or predators. Notice that the content ascriptions are plausible given the roles of these representations in the lives of hoverflies and moths. Also, the approach makes plenty of room for misrepresentation. In these examples, the content ascriptions are similar to those provided by Millikan’s theory, but the hybrid account relies on causal connections between representations and reality rather than on mere correlations.

An interesting point made by Pietroski is that Millikan’s teleosemantics looks appealing because there are cases where the contents it ascribes coincide with those ascribed by a plausible intentional explanation of behavior: “selection explanations and intentional explanations sometimes make reference to the same properties; and this is probably what makes teleological accounts of content like Millikan’s look so plausible” (Pietroski 1992: 278). For example, he points out, the plausible intentional explanation of frogs’ tongue-darting behavior and the evolutionary explanation of why frogs have such a trait both make reference to bugs (Pietroski 1992: 278). Pietroski’s kimus case is meant to show that intentional and evolutionary explanations do not always coincide. This happens when the crucial Normal condition that Millikan would regard as determining the content of a representation does not coincide with what causes the representation: “The kimu example is designed to be just like that of the frog, with one exception: Whereas bugs at least sometimes cause neural firings in frogs, snorfs never cause [neural firings] in kimus.” (Pietroski 1992: 274). As Pietroski figured out, when the causes and the crucial Normal conditions associated with a representation come apart, Millikan’s ascriptions of content look considerably less plausible. While Pietroski takes his objection to apply to any teleosemantic theory, the situation is different for causal versions of teleosemantics. Presumably, a causal version can ascribe contents that are plausible by Pietroski’s standards. As discussed
above, Neander’s teleosemantics does ascribe to kimus the content that Pietroski argues to be the plausible one. However, the ascriptions of Neander’s teleosemantics do not always coincide with what Pietroski regards as plausible ascriptions. In the frog’s case, for example, Pietroski thinks that the ascription that best explains the tongue-darting behavior is that the frog represents bugs. This is closer to Millikan’s ascription than Neander’s. So neither Millikan’s theory nor Neander’s delivers content ascriptions that consistently match the ascriptions which, according to Pietroski, would be part of a plausible intentional explanation of behavior. The hybrid teleosemantic approach I am proposing, in contrast, does deliver ascriptions that are more consistently in tune with Pietroski’s plausible ascriptions and, I dare say, with the ordinary and scientific ascriptions made to successfully explain animal behavior that is governed by representational states.

In the imaginary kimus case, the hybrid account entails that kimus represent *red things*, since red things are the cause of the representations in those occasions where the consumers managed to perform their function—mainly guiding the kimus up the hill, which was indeed beneficial because kimus avoided snorfs by going to the hilltop. The causal part of the hybrid account ensures that—unlike what happens with Millikan’s theory—the contents ascribed are properly related to creatures’ perceptual capacities. In the frog’s case, the hybrid account entails that the representations are about *prey* or *frog food*, since the presence of prey or frog food was causally responsible for the tokening of the representation in those occasions where the consumers succeeded in performing their function of making frogs catch and eat their prey. The consumer part of the approach ensures that—unlike what happens with Neander’s theory—the contents ascribed are properly related to creatures’ behavioral repertoires. Unlike Millikan’s consumer-based and Neander’s producer-based account, the hybrid account ascribes contents
that are in tune in both cases with what Pietroski argues to be plausible ascriptions, based on ordinary non-teleosemantic standards. I do think that this is an advantage of the hybrid account, because the contents suggested as plausible by Pietroski are indeed the ones we would ascribe based on the causal roles of representations—that is, on their causes and behavioral effects. While these ascriptions are in principle subject to revision, the identifications of meanings they provide are overall likely to be correct because—as I argued in Chapter 2—we are largely successful explaining behavior based on them.

While I agree with Pietroski’s take on what would count as plausible ascriptions, my case for the hybrid account does not rely merely on Pietroski’s *intuitions*. Instead, it is based on methodological considerations about what phenomena it is the task of semantics to explain. As I pointed out in Chapter 2, following Devitt (1996), we need a relatively uncontroversial identification of meanings in order to define a common explanatory goal for semantic theories. It is troubling that the disagreements between semantic theories are not just about what is the best explanation of the nature of meanings, but also about what counts as a meaning that needs explanation (Devitt 2012: 62). Given this situation, it is hard to establish that different theories are genuinely competing with each other to explain the same phenomena. Consider the disagreement between Millikan and Neander regarding frog’s representations. Their theories disagree about what non-semantic phenomena are responsible for the content of these representations; but they also disagree about what is the content to be explained: whether frogs represent *frog food* or *small and dark moving things*. Why should we regard their theories as competing with each other? We need a common and theory-neutral identification of what needs explanation to regard them as rivals.\footnote{Otherwise, it may look like Millikan and Neander are simply interested in different aspects of frogs’} Furthermore, it is problematic that the identification of the
putative content (the *explanandum*) seems to depend exclusively on the very theories that offer
the explanation of its nature (the *explanans*).

Perhaps the main point made by Pietroski in his objection to Millikan is that we need a
theory-independent identification of contents in order to be able to judge whether Millikan’s
theory succeeds or fails at explaining *them* (Pietroski 1992: 277). I argued in Chapter 2,
following Devitt (1996), that folk and scientific content ascriptions based on representations’
causal roles provide the needed independent identifications. This happens to be also the sort of
theory-independent ascription that Pietroski has in mind. As his discussion makes clear, he takes
proper content ascriptions to be those that contribute to the explanation of behavior and take into
consideration the causes of representations. He also regards discriminatory capacities as relevant
to (perceptual) content. I made a distinction above (see Section 5.4) between a reasonable and a
problematic discrimination requirement: the problematic one is that a creature cannot have a
perceptual representation of Xs if it is not able to discriminate Xs from some non-Xs—which
does not make room for false positives—while the reasonable one is that a creature cannot have a
perceptual representation of Xs if it is not able to discriminate Xs from any non-Xs.

Consider the kimus case. It is reasonable to conjecture that this case looks *prima facie*
like a counter-example to Millikan’s theory precisely because our ordinary standards for content
ascriptions would rule against kimus having representations of snorfs or snorf-free areas. Based
on the behavior of kimus and the stimuli that causes it we would likely ascribe to them instead
representations of—and an attraction towards—red things. Cognitive ethologists or comparative

representational activity, which they are *honing* with the label of “meaning”. Recall Lycan’s humorous
“Double Indexical” theory of meaning—“MEANING =_{def} Whatever aspect of linguistic activity happens to
interest me now” (Lycan 1984: 272)—which we could paraphrase more broadly here as follows: “MEANING
=_{def} Whatever aspect of representational activity happens to interest me now”.
psychologists studying kimus’ behavior would do the same. The discovery that kimus’ fondness for red things evolved because the nearby red-looking hilltop happened to be snorf-free is not likely to prompt a revision of this ascription. Since kimus’ cannot discriminate a snorf from anything else and their perceptual systems are not causally sensitive to snorfs or their absence, the best explanation of kimus’ behavior remains the one in terms of perception of—and fondness for—red things.

Now let us consider the frog’s case. Neander complains that it has been an assumption in the philosophical debate that the frog’s state represents “frog food or fly or some such thing” and that it misrepresents when triggered by a bee bee (Neander 1995: 131). But this hypothesis is based on ascriptions made after witnessing frogs’ hunting/feeding behavior and the stimuli that causes it—ascriptions that provide a very plausible explanation of frogs’ behavior. Our ordinary ascriptions are not perfect. People tend to ascribe more complex and fine-grained contents than animals’ representations are likely to have. Some people may think that frogs can represent flies qua flies, crickets qua crickets, etc. Also, some people may ascribe beliefs and desires to frogs: “It thinks there is a fly passing by and wants to eat it”. But even such anthropomorphizing ascriptions often seem to contain a grain of truth. In any case, misconceptions can be easily corrected by studying further frogs’ behavior: we can go scientific. Consider the revolutionary research on frogs’ vision by Jerry Lettvin and his colleagues. Lettvin et al. (1959) start the presentation of their discoveries with some crucial observations about frogs’ behavior:

A frog hunts on land by vision. He escapes enemies mainly by seeing them... The frog does not seem to see... the detail of stationary parts of the world around him. He will starve to death surrounded by food if it is not moving. He will leap to
capture any object the size of an insect or worm, providing it moves like one. He
can be fooled easily… by any moving small object. (Lettvin et al. 1959: 1940)

After exposing frogs to various stimuli—fooling them in various ways—and measuring with
electrodes the responses in their nervous systems, they discovered that frogs have different
feature detectors—i.e., neurons that respond to specific stimuli such as moving edges or the net
dimming of light. The most interesting ones were “convexity detectors”:

We have been tempted... to call the convexity detectors “bug perceivers.” Such a
fiber... responds best when a dark object, smaller than a receptive field, enters that
field, stops, and moves about intermittently thereafter... Could one better describe
a system for detecting an accessible bug? (Lettvin et al. 1959: 1951)

Lettvin et al. realize that these neurons play the role of bug detectors, even though they cannot
discriminate bugs from some other things—i.e., small and dark moving things.78 This is because
of the role of the detectors in the explanation of frogs’ behavior: “the purpose of a frog’s vision is
to get him food and allow him to evade predators” (Lettvin et al. 1959: 1950). Regarding them
merely as detectors of small and dark moving things would not make room for characterizing the
frog as being “fooled” in the lab when engaging in hunting behavior towards a metal pellet.

The fact that the ascriptions entailed by the hybrid account are in tune with ordinary and
scientific ascriptions provides evidence in favor of the approach: independent evidence that the
phenomena explained are indeed meanings. I have focused on the role of ascriptions in the
explanation of behavior, which is one of their main purposes (Devitt 1996). But ascriptions also

---

78 Notice that this is consistent with the reasonable discrimination requirement, but not with the problematic one,
which would not allow for a characterization of the neurons as bug detectors.
serve the purpose of informing us about reality (Devitt 1996). In the case of scientific ascriptions of meaning to the representations used in non-human languages, the main purpose is that of explaining how animal utterances serve as guides to reality to their intraspecific audiences (Devitt 2013: Sec. 3). Consider the cases I discussed of vervet monkeys and domestic chickens (see Chapter 2). Vervets use three distinct vocalizations to warn of the presence of leopards, eagles and snakes (Cheney and Seyfarth 1990). Chickens use two distinct vocalizations to warn of the presence of aerial predators and land predators (Evans and Marler 1995). Each vocalization triggers an avoidance behavior in the audience that is appropriate to the particular kind of predator being announced. What these alarm calls represent, according to the hybrid account, are the specific kind of predators that cause their utterance when the consumers in the audience manage to perform Normally their avoidance behavior functions. When a chicken or a young vervet makes an aerial or eagle call in response to a non-predatory bird—as they often do—this is a misrepresentation because the running-for-cover behavior it triggers does not perform the function of protecting from an aerial predator, at least not Normally (it may on occasion do it abNormally, when the alarm is triggered by a non-predator but a yet unseen aerial predator is also approaching). In cases of animal communication, the hybrid account also makes ascriptions that are in tune with those made by scientists studying what “information” animals communicate to their conspecifics.

The hybrid account looks at what the consumers of representations need to perform their functions in order to determine which among all the causes that the producers respond to are the ones that they are supposed to respond to. By relying on the actual causal interactions of perceptual mechanisms, the approach avoids ascribing implausible contents. By relying on the
functions of the consumers to separate Normal from abNormal causes, the approach is able to make plenty of room for misrepresentation.

5.7 Conclusion

In this chapter I have explored the core motivation for teleosemantics and the main teleosemantic approaches. On one hand, I argued that a consumer- or output-based approach makes ample room for misrepresentation but can ascribe implausible contents given the cognitive capacities of the perceptual mechanisms involved. On the other hand, I argued that a producer- or input-based approach has the advantage of explaining content in terms of what perceptual mechanisms can causally interact with, but has the drawback of not making enough room for misrepresentation. I proposed a hybrid producer-consumer approach that promises to make ascriptions that take into account the capacities of perceptual mechanisms while also making enough room for misrepresentation. The focus of this chapter has been mainly on basic representations. In the next chapter I discuss how teleosemantics has been extended to complex human representations and I argue that a proper teleosemantics for such representations should be articulated in terms of basic sentence acceptance. The resulting account combines insights from Horwich’s use-theory of meaning and available teleosemantic theories.
Chapter 6:
Basic-Acceptance Teleosemantics

6.1 Introduction

In Chapter 5, I argued that a hybrid producer-consumer teleosemantics ascribes the right contents to representations—taking into account the capacities of the (perceptual) mechanisms that produce them, while also making enough room for misrepresentation. The proposal is that the content of a representation is determined by its *normal cause*: what is supposed to cause tokens in order for them to produce the effects they have the function of bringing about. The focus in Chapter 5 was on the contents of the most basic representations, which are the least controversial candidates for a teleosemantic account, since the existence of the mechanisms that produce and consume them is surely explained by natural selection operating over innate traits. But how can a teleosemantic approach be extended to cover also the meanings of the sophisticated mental and linguistic representations used by human beings? I suggested in Chapter 3 that a combination of basic-acceptance semantics with a teleonomic account of the functions of words and concepts can overcome the problems of ignorance and error that undermine Horwich’s dispositionalist version. In this chapter, I will elaborate on what sets apart human representations and what are the challenges for a teleosemantic account of their contents, I will then discuss available attempts by Papineau and Millikan to offer a teleosemantics for human representations and critically assess them. Finally, I will discuss how a basic-acceptance teleosemantics can be articulated and what advantages does it have over standard teleosemantic accounts and over Horwich’s theory.
6.2 Etiological Functions and Human Representations

The sophisticated mental and linguistic representations that humans use for thinking and for communicating with each other differ in a number of ways from the basic representations I discussed in Chapter 5. Notice that plenty of non-human representations also differ from such basic representations in at least some of these ways and it is likely that some human representations are basic. But let us focus on the stark contrast between sophisticated human representations—which I will simply call “human representations”—and basic representations:

i. Many human representations are *acquired* rather than innate. Unlike beaver tail-splashes and the inner states produced by moth’s bat-detectors, which are directly the result of genetic selection, all the linguistic symbols that humans use to communicate with each other are the result of linguistic conventions and many—perhaps most—of the mental symbols that humans use for thinking are the result of learning. Humans coin new concepts as a result of their interactions with the environment and thought processes, as well as new words to communicate thoughts involving these new concepts, but they also acquire concepts from learning the conventional meanings of words in their linguistic communities. So there is a complex interrelation between human language and thought.

ii. Human representations are *compositional*. Beaver tail-splashes and the states produced by moth’s bat-detectors are not combinations of simpler representations and do not combine into more complex ones. Human sentences and thoughts, in contrast, are combinations of words and concepts, respectively, and their meanings depend on the meanings of these parts and the ways they are combined. The recursively applicable rules
of syntax explain the productivity and systematicity of human language: how a finite stock of words can be combined into innumerable sentences and how the words of some sentences can be recombined into other sentences. The hypothesis that thoughts are syntactically structured combinations of concepts offers a very plausible explanation of the analogous productivity and systematicity of human thought (Fodor 1987: 147-154).

iii. Most inner human representations only have behavioral effects when combined with other representations. The activated inner states of moth’s bat-detectors automatically make moths engage in bat-avoidance behavior without the mediation or collaboration of other inner representations. Human concepts can only have behavioral effects when combined into thoughts. But most human thoughts do not have behavioral effects on their own either: beliefs and desires only cause behaviors when suitably combined with each other to produce decisions to act—e.g., neither the belief that there is food in the fridge nor the desire to get food will independently make someone open the fridge to get food.

iv. Human thoughts—as well as the sentences used to express them—are inferentially linked. The inferential processes involved in theoretical and practical reasoning connect beliefs and desires to produce new beliefs and desires and, eventually, to produce decisions to act. The inferential links between thoughts parallel those between sentences. Just like ‘Dolphins are warm blooded’ can be inferred from ‘Dolphins are mammals’ and ‘All mammals are warm blooded’, someone can come to believe that Dolphins are warm blooded because of believing that they are mammals and that all mammals are warm blooded. Since our best explanation of inferential links relies on syntactic structure, the account of inferential processes as computational operations performed over syntactically
structured inner representations arguably offers the best explanation of thinking (Fodor 1987: 143-147). While some non-syntactic structures—such as map-like structures—could in principle account for the productivity and systematicity of thought, it is unclear how could they account for thinking processes (Devitt 1996: 154-156).

v. Human words as well as concepts have different kinds of meanings. Beaver tail-splashes and the activated states of month’s bat-detectors have meanings determined by direct links to reality. While many human words and, more fundamentally, the concepts they express also have meanings determined by direct links to reality (e.g., ‘tiger’), it is very plausibly the case that others have meanings determined by links to other words/concepts (e.g., ‘bachelor’) and yet others have meanings determined by their links to rules of inference (e.g., ‘and’).

vi. The overall deployment of concepts—and, consequently, of the words used to express them—is partly a function of inferential processes. While the activated states of moth’s bat-detectors are directly produced by perceptual mechanisms, many tokens of human concepts are produced instead by inferential mechanisms that token thoughts containing them as a result of other thoughts containing them—even when the causal-inferential chain may begin, for concepts linked to reality, with tokens produced by perception.

Teleosemantics can offer a plausible account of the meanings of basic representations which are innate, non-compositional, directly connected to perceptual causes and behavioral effects and, consequently, directly linked to what they represent. But an ambitious teleosemantic theory that expects to explain the nature of the meanings of mental and linguistic human representations
must: (1) offer an account of how acquired representations which do not have a history of genetic selection can have etiological functions, (2) offer an account of the meanings of thoughts/sentences based on an account of the meanings of concepts/words and how they are combined, (3) offer and account of the different functions of beliefs and desires and (4) offer an account of the explanatory role of inference in the performance of the functions of beliefs and desires. Additionally, as I argued in Chapter 3, a proper theory of meaning must (5) offer an account of how the meanings of various concepts and words are related either to their links to reality, their links to other concepts and words or their links to rules of inference, and also (6) explain how the meanings of linguistic and mental words (i.e., concepts) govern their overall use in inference. I will argue that none of the available versions of teleosemantics properly satisfies requirements (5) and (6), but that a teleosemantic theory articulated in terms of basic sentence acceptance—a “basic-acceptance teleosemantics”—can do so. Recall that I argued in Chapter 3 that we need a basic-acceptance semantics precisely in order to account for the variety of kinds of meaning and, more importantly, for how meanings govern the uses of words. But I also argued that Horwich’s dispositionalist version is undermined by problems of ignorance and error. Basic-acceptance teleosemantics is meant to overcome this problem by virtue of being based on a teleonomic account of how words are supposed to be used—instead of a dispositionalist account of how they are regularly used.

Before elaborating on my proposal, I will discuss the teleosemantic theories that Papineau and Millikan offer to explain the meanings of human representations. Their accounts attempt to satisfy requirements (1) to (4). Among these, (1) is the crucial requirement. An ambitious teleosemantic theory can borrow available (non-teleonomic) accounts of belief-desire
psychology, inference and compositionality—giving a teleonomic twist to them—as long as human representations and/or the mechanisms responsible for their deployment have etiological functions. The most pressing question is how acquired human representations and/or the mechanisms responsible for their use can have etiological functions. The version of teleosemantics I will propose is new in regarding basic acceptance as crucial to meaning determination. But it will borrow plenty from the work of Papineau and Millikan, particularly regarding how acquired representations can have etiological functions.

How can acquired representations have etiological functions? One option, proposed by Papineau, is to regard the learning processes responsible for the acquisition of new mental representations as selective processes (Papineau 1984: 557). Another option, proposed by Millikan, is to regard acquired mental representations as having functions derived from the functions of the innate mechanisms responsible for their acquisition and deployment (Millikan 1984: 41-43). Regarding linguistic representations, Millikan suggests that they have etiological functions because they are socio-culturally reproduced items with their own histories of selection. I will start discussing mental representations. If Papineau is right, acquired representations have direct functions. This allows for a fairly straightforward application of teleosemantics to human mental representations. I will first elaborate on Papineau’s account of mental content. Afterwards I will contrast it with Millikan’s alternative—and elaborate on her teleosemantic account of linguistic meaning. It is worth pointing out that despite differences in their earlier work, the views of Millikan and Papineau have to a large extent converged over the years. This includes their views of what are the sources of teleonomy for acquired representations, which happen to be complementary rather than mutually exclusive.
6.3 Papineau on Human Mental Representations

Papineau offers accounts of the contents of human beliefs, desires and the concepts that compose them, which are based on the functions these representations have been allegedly selected for performing—whether they are innate or acquired. In this section, I will elaborate on his proposal.

Papineau expects his teleosemantic account to cover both innate and acquired representations, since he conceives the acquisition of new representations as a selective process that is analogous to the intergenerational selection of innate traits: “Natural selection takes place in learning as well as in intergenerational evolution (though then it is natural selection of cognitive components rather than genes).” (Papineau 1984: 557). Dretske’s (1988) suggestion that the recruitment of representations by operant conditioning may be regarded as a form of selection is a variant of the same general idea (see Chapter 5, Section 5.3). The idea that learning processes are selective has an important precedent in twentieth-century psychology. The influential psychologist B. F. Skinner insisted on the similarity between operant conditioning and natural selection:

In both operant conditioning and the evolutionary selection of behavioral characteristics, consequences alter future probability. Reflexes and other innate patterns of behavior evolve because they increase the chances of survival of the *species*. Operants grow strong because they are followed by important consequences in the life of the *individual*. (Skinner 1953: 90)

Skinner’s claim is that both innate and learned behaviors are shaped by their *effects*. While the phylogenetic shaping of behavior takes long periods of time, working only across multiple
generations, the ontogenetic shaping of behavior due to operant conditioning takes place in short periods of time, during the life of individuals. Skinner suggested that both processes “change the organism so that it adjusts to its environment in the sense of behaving in it more effectively... Successful responses are selected in both cases, and the result is adaptation” (Skinner 1966: 1211-1212; emphasis added). The idea is that learning is a fast-paced selective process that, just like its slower phylogenetic counterpart, produces adaptive behaviors. Furthermore, Skinner pointed out that both the phylogeny and the ontogeny of behavior “seem to «build purpose into» an organism” so that “an ongoing system acts as if it had a purpose” (Skinner 1966: 1211; emphasis added). Notice that Skinner here is getting very close to suggesting that the products of operant conditioning, just like those of natural selection, have teleonomic functions or purposes. He seems to be warning—just like those biologists that cautiously distinguish teleonomy from classical teleology—that the sense in which the products of selection can be said to have “purposes” is not the robust sense of in which intentionally designed items have purposes.

Skinner’s selectionist view of operant conditioning is very plausible—since the analogies between operant conditioning and natural selection are striking—and it constitutes an important precedent of the selectionist views of learning proposed by Papineau and Dretske. But Skinner’s selectionist view is entangled with two problematic features of his thought. First, Skinner was a radical behaviorist who disregarded inner or mental phenomena as irrelevant to the explanation of behavior. Second, he expected conditioning to explain the ontogeny of all acquired behaviors, from the simplest to the most sophisticated ones. This model turned out to be inadequate and has been largely replaced by a cognitivist model in which inner structures, 79 See Hull, Langman and Glenn (2001) for a detailed discussion of the analogies between operant conditioning and natural selection—as well as a general characterization of selection that applies to biological evolution, operant behavior and the workings of the immune system.
processes and representations play crucial explanatory roles. One of the triggers of this cognitive
revolution was, incidentally, the criticism by Noam Chomsky (1959) of Skinner’s attempt to
explain the acquisition of human verbal behavior in terms of reinforcement (Skinner 1957).
Operant conditioning is now considered just one among many forms of learning—not well suited
to explain verbal or other sophisticated behaviors—and it is often itself taken to involve
cognitive components. Dretske’s suggestion that learning may be regarded as selecting inner
representations, for example, resurrects Skinner’s selective account while disregarding his
behaviorism. However, Dretske only presents the idea as a tentative suggestion, when talking
specifically about representations acquired by means of operant learning (Dretske 1988: 101),
and refrains from fully endorsing a selective account of the functions of all acquired
representations.80 Papineau, in contrast, is confident that the kinds of learning involved in the
acquisition of all inner representations—including human beliefs, desires and concepts—are
function-conferring selective processes. He offers the following hypothetical illustration of how
this may come about in the case of beliefs:

Suppose our individual psychological developments throw up new possible belief
types, new ways of responding mentally to circumstances, at random, analogously
to the way that our genetic history throws up mutations at random. Then we
would expect such new dispositions to become ‘fixed’ just in case belief tokens
give rise to advantageous (that is psychologically rewarding) actions, analogously
to the way that genetic mutations become fixed just in case they have

80 Dretske prefers to talk about the “recruitment” of innate and acquired representations. He is only committed to
the recruitment of innate representations being a selective process. But he does not elaborate much on how the
recruitment of both innate and acquired representations is a function-conferring process. Godfrey-Smith argues
that this “leaves a gap in [Dretske’s] theory” that should be filled in by a selective account which “can apply
equally to genetic and learning-theoretic cases” (Godfrey-Smith 1992: 291).
Papineau does not articulate a detailed account of how human beings actually acquire representations—which is something better left for scientific psychology to figure out. But he warns that the link between sophisticated representations and behaviors in human belief-desire psychology is not fixed merely by conditioning which, he points out, is a “relatively primitive method of generating actions” (Papineau 1993: 79-80). Nonetheless, he expects the acquisition of the relevant representations to have—just like operant conditioning—crucial features in common with genetic selection. This is a very interesting idea. It is very plausible that some inner representations are acquired by means of some selective process. It is doubtful, however, that the analogy with genetic selection can be properly extended from operant conditioning to the kinds of learning involved in the acquisition of all novel representations. But the account can be combined with the account Millikan offers (see Section 6.4) in order to provide a solid teleonomic account of the acquisition of human mental representations. I will elaborate on how Papineau makes use of this idea to offer a teleosemantic account that applies equally to the meanings of innate and acquired human representations.

Papineau offers teleonomic accounts of the contents of beliefs, desires and the concepts that compose them. Regarding beliefs, his suggestion is the following:

The disposition to form a given type of belief is explained by the fact that that belief has typically arisen in certain circumstances, and in those circumstances the actions that it has directed have been selectively advantageous. The typical circumstances in question are the belief’s truth conditions. (Papineau 1984: 557)
Since Papineau regards learning as a selective process, this account is supposed to cover all beliefs. The idea is that a belief (whether acquired or innate) has been selected because it has had beneficial behavioral effects (either for the individual or for its ancestors) when tokened under specific conditions. Given its selective history, the belief is *supposed* to be tokened in such conditions, which are what it represents—its truth-conditional content. This account has the virtue of making room for misrepresentation: belief tokens are false when they occur in “untypical” circumstances that differ from those under which those types of tokens where selected (Papineau 1984: 558). Suppose that I have acquired a disposition to token certain belief because past tokens of this belief had beneficial behavioral effects specifically when it was raining—such as helping me stay dry by making me seek cover or pick up an umbrella. My belief is consequently *supposed* to be co-present with the occurrence of rain and it is false when tokened in the absence of rain.

Notice that Papineau’s account is somewhat similar to Millikan’s account of the content of basic innate representations—although extended to cover acquired representations: in both cases the content is the condition that the representation, given its selective history, is supposed to covary with. (A crucial difference is that Papineau relies on “typical” conditions, while Millikan relies on “Normal” conditions—and the difference is not merely terminological, since he does regard such conditions as those that were average during the selection period. This leads to a problem that I will point out in the next section.)

Beliefs, however, differ from basic representations not only because they are often not innate, but also because they only cause actions when combined with desires. My belief that it is raining will only make me pick up an umbrella or seek cover if I want to stay dry. Also, the
contents of beliefs and desires depend on the contents of concepts that compose them and the way they are combined. The content of the belief that elephants swim depends on the content of the concepts expressed by ‘elephants’ and ‘swim’—and presumably on its predicational structure. Papineau attempts to accommodate these complexities by offering teleonomic accounts of the content desires, of the content concepts and of how the different pieces of the puzzle fit together.

The “biological purpose” of desires, Papineau proposes, is to cause behaviors that bring about conditions which meet the varying needs of the organism. He offers the following account of their content:

[The] satisfaction condition for a given desire is that characteristic result of the actions it directs which has been selectively advantageous, and the production of which therefore explains the disposition to form that desire. (Papineau 1984: 562)

The idea is that a desire has been selected because the behaviors prompted by its tokens produced certain beneficial effects. Given its selective history, the desire is supposed to produce such effects, to bring about certain conditions in the world, which are what it represents—its satisfaction conditions. Again, Papineau intends this account to cover innately determined and acquired states. The explanation of why I have an innate disposition to desire to eat (whenever I need nutrition) involves intergenerational natural selection: my ancestors benefited from having inner states that made them get food into their stomachs. According to Papineau, the acquisition of novel desires is also a selective process (Papineau 1984: 563). The explanation of why I have a desire to drink coffee in the morning, for example, may be that drinking coffee in the morning
has been psychologically rewarding in the past, so a disposition to token the desire has been selected for producing this behavior.

In Papineau’s account, desires and beliefs are both selected because of their behavioral effects, but their contents depend on different aspects of their roles. The contents of desires are the conditions they are supposed to produce, while the contents of beliefs are the conditions they are supposed to be co-present with. Notice that desires “fluctuate… in response to environmental circumstances”—e.g., the desire to eat is supposed to be triggered when the level of blood sugar is low—yet they do not represent such circumstances (Papineau 1984: 564). Conversely, beliefs are supposed to direct behavior, but they do not represent their behavioral effects. How are the contents of desires and beliefs related to their etiological functions or purposes in Papineau’s account? In the case of desires, the answer is simple: their contents are the conditions they have the purpose of bringing about. Regarding beliefs, Papineau’s initial suggestion is that their purpose is “to be present when certain states of affairs obtain” (Papineau 1984: 559). So their contents would be the conditions they have the purpose of being co-present with. But this characterization of the purpose of beliefs is problematic. Etiological functions are selected effects. Being co-present with certain states of affairs is not even an effect of beliefs. A later suggestion of Papineau is that the purpose of beliefs is “to produce whichever results will fulfil the purposes of the desires they are acting in concert with… to generate actions which will produce desired results in such-and-such conditions” (Papineau 1998: 8). This second suggestion is preferable. Directing behavior to satisfy the desires they are combined with is an effect that beliefs could be selected for producing (including acquired beliefs, if Papineau’s selective account of learning were right). In order to explain their contents in terms of co-presence, then,
we should not say that beliefs have the purpose of being co-present with certain states of affairs, but rather that beliefs perform their function of contributing to the satisfaction of desires by being co-present with certain states of affairs, which are what they represent.

Papineau refines his characterization of the contents of beliefs, taking into account their collaboration with desires: “we should count as the truth conditions of beliefs not simply circumstances in which the resulting actions have advantageous effects, but rather circumstances in which those actions lead to the satisfaction of desires” (Papineau 1984: 564-65). Consider, for example, my disposition to form the belief that the artifact I am perceiving is my television’s remote control. According to Papineau’s account, the explanation of why this belief has been selected is that its tokens have led in “typical circumstances” to the satisfaction of the desires they were combined with. The “typical circumstances” were specifically those in which I was perceiving my television’s remote control, so those circumstances are the belief’s truth conditions. Once the truth conditions of the belief have been fixed by its selective history, tokens will still cause similar behaviors when combined with similar desires, but they may be false and, consequently, fail to perform their function. When combined with the desire to turn on the television, the belief will make me press the artifact’s power button. When combined with the desire to turn off the sound of the television, it will make me press the artifact’s mute button. Such behaviors may succeed or fail to satisfy the relevant desires depending on whether the belief is tokened in “typical” or “untypical” circumstances—e.g., when I mistake my stereo’s remote control for my television’s, I fail to turn on the television when pressing the artifact’s power button. Basically, Papineau explains the current success or failure of an already selected belief in performing its function in terms of its truth or falsity and the truth conditions of the
belief in terms of its past success in contributing to the satisfaction of desires. (Papineau further characterizes truth conditions of a belief as those that guarantee the success of actions based on them. This is another problematic feature that I will briefly discuss in the next section.)

A belief will produce different behaviors in combination with different desires. Beliefs do not have specific behavioral effects of their own—however, they do have specific circumstances under which they are supposed to be tokened. Similarly, a desire will produce different behaviors in combination with different beliefs. My desire to eat may make me open my fridge or go to the grocery store, depending on what beliefs it is combined with. But Papineau’s account does require some specific effect of all successful behaviors directed by a desire to be its satisfaction condition, since this is the condition the desire has the purpose of bringing about. So he suggests that the content of the desire is the effect it is always supposed to produce, regardless of what beliefs it is combined with (Papineau 1998: 12). Whether it makes me open my fridge or go the grocery store, my desire to eat has the purpose of getting food into my stomach. Still, in order to direct behaviors that fulfill their purposes, desires must be combined with appropriate beliefs: beliefs that are true and relevant to their satisfaction. Papineau suggests that part of the purpose of the “inferential mechanism” in the “human decision-making system” is to generate true beliefs about means to satisfy desires—out of true background beliefs—and that, consequently, there is an interdependence between the purposes of beliefs, desires and the inferential mechanism (Papineau 1993: 76-77). The inferential mechanism, he proposes, has been selected for making valid inferences: “Habits of thought which correspond to... deductively valid steps will tend to be preserved by natural selection” (Papineau 1987: 77). The advantage of this etiological account is that it is able to explain why the mechanism is supposed to produce certain conclusions rather
than others. While “humans often draw invalid conclusions”, Papineau points out, this “doesn’t show that validity isn’t the inferential mechanism’s purpose, any more than heart failures show that blood circulation isn’t the heart’s purpose” (Papineau 1993: 77). Notice that a non-historical account of the function of the inferential mechanism in terms of its current dispositions—of the sort that a Cummins-style functional analysis would provide—cannot explain inferential error for the same reason that it cannot explain why a failing heart is malfunctioning (see Chapter 4, Section 4.7).

Papineau’s discussion of the inferential mechanism focuses on deductive inferences, including those that have beliefs about means to satisfy desires as conclusions. But a full account should be extended in two ways. First, it should cover inductive and abductive theoretical inferences—which also play important roles in how human beings form new beliefs out of old ones. Second, it should cover practical inferences that have beliefs and desires as premises and intentions or decisions to act as conclusions—which are crucial to explain how beliefs and desires are combined to direct behavior. Consider my belief that drinking the liquid in my glass will quench my thirst. This seems to be the kind of belief about means that Papineau has in mind. I may have inferred it from my beliefs that lemonade quenches thirst and that the liquid in my glass is lemonade. But the belief will not make me drink the liquid unless it is combined with my desire to quench my thirst to produce a decision to act. This form of practical inference provides a clear example of the interdependence between the purposes of beliefs, desires and the inferential mechanism. Also, it is in tune with Papineau’s claims that beliefs and desires cannot have behavioral effects on their own and that the inferential mechanism is part of the decision-making system. Once the etiological account has been extended to cover these other forms of
inference, however, the claim that validity is the purpose of the inferential mechanism must also be revised. Good inductive and abductive inferences are not valid. Intentions or decisions to act have satisfaction-conditions rather than truth-conditions, so a good practical inference cannot be characterized as truth-preserving. What the etiological account must say is that the purpose of the inferential mechanism (or mechanisms) is to produce conclusions that are well-supported by the premises—thereby producing (1) beliefs that are guaranteed or likely to be true when the premises of theoretical inferences are true and (2) decisions to act in ways that satisfy desires when the beliefs about means in the premises of practical inferences are true.

A crucial piece of the puzzle is still missing. Human beliefs and desires are structured combinations of meaningful parts. The structures of thoughts are crucial to their participation in inferential processes. If the purposes of beliefs and desires are interdependent with the purpose of the inferential mechanism in the decision-making system, then beliefs and desires must have appropriate structures in order to perform their functions. Additionally, the meanings of thoughts depend on their structures and the meanings of their parts. A teleonomic account of the functions and contents of thoughts must explain how they are related to the functions and contents of the concepts that compose them. Papineau proposes that the function of a concept depends on its selective history:

A concept will get selected because in combination with other concepts it forms beliefs which, in ‘typical’ circumstances, lead to advantageous behavior. At bottom it is the concepts that get selected... And so we can think of concepts themselves as having purposes – namely, to ‘refer’ to certain objects – which then contribute to the purposes of beliefs. (Papineau 1987: 76-77)
Papineau intends a full account to include desires: concepts get selected because they form beliefs and desires which, in concert with each other, lead to advantageous behavior. Like him, I will set aside desires for the sake of simplicity. Papineau expects the truth-conditions of beliefs to be explained by their structure and the referential values of the concepts they contain: “we need to start with the components of beliefs, such as singular concepts, predicate concepts, ways of combining concepts and so on, and to focus on the referential values of such components” (Papineau 1993: 82). His account of truth-conditions in terms of what states of affairs beliefs are supposed to be co-present with is reductive: it explains truth-conditions non-semantically, in terms of teleonomic links between beliefs and states of affairs. But he intends the complete explanation to be compositional: what state of affairs a belief is supposed to be co-present with, he says, “will depend in turn on the concepts and structure making up the causal role of the belief” (Papineau 1987: 76). His basic idea is clear: what a concept contributes to the truth-condition of a belief is its referential value, which is reductively explained by its teleonomic links to certain objects.

Papineau insists that a concept can only get selected because it contributes to beliefs performing their functions (Papineau 1987: 77). So the teleonomic links between a concept and certain objects are those links that it had when it was selected because the beliefs it was tokened in directed advantageous behavior. While Papineau says that the purpose of concepts is to “refer” to certain objects, this semantic description is a place-holder for a non-semantic one in terms of teleonomic links. But notice that this characterization of the function of concepts is problematic. Referring to certain objects—even if this is spelled out in non-semantic terms—is not an effect of a concept. The effects that concepts could be selected for producing (if Papineau’s selective
Suppose that beliefs are mental sentences. According to Papineau’s theory, then, the name ‘Harry’ (i.e., a singular concept) in my mental lexicon was selected because it was tokened in mental sentences that were co-present with states of affairs involving Harry. The link between ‘Harry’ and Harry was what the word contributed in those occasions to the link between the full sentences and the complete states of affairs. Similarly, my mental predicate ‘is talking’ (i.e., a predicate concept) was selected when tokened in mental sentences that were co-present with talking things. Given their selective history, ‘Harry’ and ‘is talking’ are supposed to be tokened in sentences that are co-present with states of affairs involving Harry and talking things, respectively. Now suppose that I combine the name and the predicate, forming a mental token of ‘Harry is talking’. Given its structure and the teleonomic links of its parts, this sentence is supposed to be co-present with the state of affairs that Harry is talking. If I made a mistake and the one who is talking is instead Harry’s twin brother, my belief is false because ‘Harry’ has a teleonomic link to Harry and not to his brother. If I made a mistake and Harry is babbling rather than talking, my belief is false because ‘is talking’ has a teleonomic link to talking things and not to things that are making meaningless vocal sounds. Reference is not determined by the current causal links to reality of tokens of a mental word. So the account makes room for misrepresentation, for errors in the deployment of mental words.

As we have seen, Papineau adopts the common characterization of the contents of beliefs
and desires in terms of truth and satisfaction conditions, as well as a compositional characterization of such conditions in terms of the structures of thoughts and the referential values of the concepts that compose them. Additionally, he adopts the standard characterization of the roles of beliefs, desires and inference in the production of behavior. But Papineau explains each of these elements in terms of etiological functions—specifically, of functions that beliefs, desires, concepts and even inference patterns have been selected for performing. The main advantage of this etiological twist is that it promises to make room for representational error. I have tried to offer a sympathetic summary of Papineau’s views. But the way he articulates his teleosemantic account does face some difficulties—the most serious of which stems from his bold suggestion that the learning processes responsible for all acquired mental representations are selective processes. These problems are avoided by combining his account with Millikan’s alternative one. In the next section, I will explain how Millikan’s account of the functions and contents of mental representations differs from Papineau’s, discuss how it avoids some problems faced by Papineau’s account and elaborate on how Millikan extends teleosemantics to cover linguistic representations. Afterwards, I will discuss the problems shared by both accounts and argue that a basic-acceptance teleosemantics can overcome them.

6.4 Millikan on Human Mental Representations

Millikan’s teleosemantic account of the contents of human mental representations has many features in common with Papineau’s. Like Papineau, she proposes that the main etiological function of desires is to bring about certain conditions in the world, which constitute their contents or satisfaction conditions, and that beliefs—given their etiological functions—are
supposed to be correlated with certain conditions in the world, which constitute their contents or
truth conditions. Millikan also offers a teleonomic account of the inferential mechanisms involved in the decision-making system and a compositional explanation of the contents of thoughts in terms of their structure and the teleonomic links to reality of the concepts they contain. So described, Millikan’s proposal seems identical to Papineau’s. But this general description glosses over some differences between the proposals, which I will discuss in this section. I will first discuss the similarities and some minor differences between the proposals. Afterwards I will focus on the most significant difference: that Millikan’s account of the functions of acquired mental representations does not rely on their own selective history, but on the selective history of the innate mechanisms responsible for their acquisition and use: novel mental representations have derived rather than direct functions. I shall argue that this feature of Millikan’s account is a crucial contribution to the project of extending teleosemantics to cover acquired mental representations. While it is indeed very plausible that some acquired representations have selective histories of their own, as Papineau suggests, there are reasonable doubts about all of them having such histories. However, there are good reasons for expecting that the underlying innate mechanisms that enable human beings to coin and deploy novel representations been selected for doing so. So a combination of both accounts of the source of teleonomy for human mental representations is preferable. After discussing Millikan’s account of acquired mental representations, I will present her proposal for how to extend teleosemantics to cover acquired linguistic representations.

Let us start with desires. Like Papineau, Millikan proposes that the contents of desires are individuated by their purposes: “the descriptions that we give of desires are descriptions of their
most obvious proper functions” (Millikan 1993: 67). The content we ascribe to a desire, according to Millikan, is the condition it has the purpose of bringing about if the ascription is correct. Millikan suggests that desires have an additional purpose related to their role in the decision-making system: “desires have as a subsidiary proper function to participate in practical inferences along with beliefs” (Millikan 1993: 71). The idea is that desires perform their main function—bringing about their satisfaction conditions—by means of performing the function of participating in inferences that combine them with relevant beliefs to produce decisions to act. Like Papineau, Millikan suggests that the inferential mechanisms themselves have etiological functions. One such function is “to produce true beliefs from prior beliefs... by conforming to logical principles... that move us to true beliefs reliably, or relatively reliably (inductive inference), from other truths” (Millikan 1993: 245-46).\(^8\) She also stresses that the inferential mechanisms have the function of combining beliefs and desires to produce decisions to act (Millikan 1993: 99).

In contrast with her view on desires, Millikan argues that the contents of beliefs are not individuated by their purposes. One function of a belief, she proposes, is “to participate in inferences in such a manner as to help produce fulfillment of desires”, while another function is “to participate in inferences to yield other beliefs, true ones” (Millikan 1993: 71). Millikan’s account of the functions and contents of beliefs differs in some respects from Papineau’s. She does identify, like Papineau, the truth-conditional content of a belief with the state of affairs it is supposed to be correlated with—or, as she prefers to put it, “correspond” or “map onto”. However, she does not regard mapping onto a state of affairs as the function of a belief. So why

---

8\(^1\) The inferential mechanisms can also function properly when producing true beliefs from prior false beliefs—as it is the case when they perform correct reductio inferences (Millikan 1993: 246n).
is the belief *supposed* to map onto a specific state of affairs? Millikan’s general answer is the same as the one she gives in the case of basic representations: mapping onto a specific state of affairs is the crucial Normal condition for the performance of the belief’s function (Millikan 1993: 71-74). When I presented Millikan’s account of the content of basic representations (Chapter 5, Section 5.4), I pointed out that the notion of a “Normal condition” is a normative and historical notion: it is a condition that must obtain for an item to perform its function in accordance with a Normal explanation—which in turn is an explanation of what has historically happened during the *proper* performance of the function. This simple model must be adjusted in order to apply to the performance of the *derived* functions that novel human beliefs have according to Millikan. I will set this aside to simplify matters, but I will get back to it shortly. For the moment, I will simply assume that Millikan has a way to account for the functions of novel representations and the Normal conditions for their performance. Another complexity is that the mapping in the case of human beliefs is not between a simple representation and a state of affairs, but involves instead some sort of compositional principles according to which a belief is supposed to map onto a certain portion of reality based on its structure and the referential links to reality of the concepts it contains—just like in Papineau’s full account. I will also set these issues aside for a moment.

Millikan’s account of the functions of beliefs, as well as her notion of a “Normal condition”, differs from Papineau’s account of the functions of beliefs and his notion of a “typical circumstance”. As I pointed out above, Papineau’s proposal that the function of a belief is “to be present when certain states of affairs obtain” (Papineau 1984: 559) does not fit well with the view that functions involve *effects*. Millikan, in contrast, characterizes the functions of a

---

82 Here is another passage where Papineau expresses this view: “the biological purpose [is] to occur in
belief clearly in terms of the effects it is supposed to have: one function of a belief is “to participate in inferences in such a manner as to help produce fulfillment of desires”, while another function is “to participate in inferences to yield other beliefs, true ones” (Millikan 1993: 71). In fairness, Papineau’s later proposal—i.e., that the function of a belief is “to produce whichever results will fulfil the purposes of the desires they are acting in concert with” (Papineau 1998: 8)—is equivalent to Millikan’s and does not have the problem of his initial one, as I pointed out in the last section. The views of Millikan and Papineau have converged on this matter. Still, there are a couple of other significant differences on which the accounts differ.

Recall that, for Papineau, a belief represents those “typical circumstances” under which it has lead to advantageous behavior. For Papineau, the “typical” circumstances under which a belief is supposed to be tokened are the average or ecologically normal conditions that obtained during the period when it was selected. This allows for beliefs being misrepresentations when tokened in “untypical” circumstances after selection has taken place: “it is only the past predominance of true belief over false that is required” (Papineau 1984: 558). But this does not make enough room for misrepresentation happening during the selection period. As Devitt objects in his review of Papineau (1987):

The insistence that a belief must have had a history of usually being true runs into the problem that was so disastrous for the... hybrid theory consisting of informationalism with a dash of teleology… This is the problem of false

---

the presence of certain states of affairs, which states of affairs therefore [count] as their truth conditions” (Papineau 1987: xvi). Papineau (1993) still makes a similar claim: “we can pick out the real truth condition of a belief as that condition which it is the biological purpose of the belief to be co-present with” (Papineau 1993: 58). But he clarifies in a footnote: “biological purposes are always a matter of results… the purposes of beliefs are to get the organism to behave in a way appropriate to certain circumstances” (Papineau 1993: 59n). This latter claim avoids the mentioned problem.
positives. Many species have survived only by mostly misrepresenting the world. What got selected, *from the beginning*, were beliefs whose tokens were true *when it mattered for survival*, even though they may have been mostly false. In this crucial respect, Papineau’s full-blown teleological theory is no advance over the partly teleological hybrid theory. (Devitt 1991: 439)

The hybrid theory that Devitt is alluding to here is the combination of indicator semantics with an etiological account of functions offered by Dretske (1988). Papineau’s account based on “typical conditions” suffers from the same problem with false positives (*see* Chapter 5, Section 5.3). Additionally, Papineau characterizes typical circumstances as *guaranteeing* the success of behaviors: “truth conditions are circumstances which ensure the success of action” (Papineau 1984: 565). As Devitt remarks, this is simply wrong: “Truth does not guarantee success. The satisfaction of many desires depends on things that are quite outside the control of the desirer: on the natural order; on the socio-economic situation; on the personal foibles of others; and so on.” (Devitt 1991: 439). Based on my true belief that it is raining I may carry an umbrella and succeed in staying dry. But there may be no umbrella around. My true belief obviously does not guarantee that I will succeed.

Millikan’s characterization of content-fixing circumstances avoids the two problems just mentioned: her Normal conditions are neither typical conditions nor success-guaranteeing conditions. Instead, they are the conditions that obtained on those special historical circumstances under which the representation successfully performed its function. Such conditions may have *rarely* obtained when the representation was tokened. As I showed in Chapter 5, Section 5.4, Millikan’s account in terms of Normal conditions makes room for
frequent false positives. Additionally, the conditions may obtain on plenty of occasions when the representation fails to perform its function. That a frog darts its tongue in the presence of a fly does not guarantee that the fly will not manage to escape. Millikan’s account has no problem with this because her “Normal conditions” are not sufficient conditions for the performance of a function. An output-oriented teleosemantic theory like Papineau’s, which attempts to explain the truth-conditions of beliefs in terms of circumstances that they are supposed to be co-present or covary with in order to bring about beneficial effects, is better off relying on a Millikanian notion of Normal conditions. I take it that this is a small but helpful amendment for a theory like Papineau’s. As I argued in Chapter 5, however, while an output-oriented teleosemantics based on Normal conditions makes plenty of room for misrepresentation (including false positives), it also faces problems that stem from not assigning any role whatsoever to the causes of representations in the determination of their meaning. This is a shared—and in my opinion serious—concern about Millikan’s and Papineau’s accounts. I will get back to it in the next section.

It is time to discuss the main difference between the proposals of Millikan and Papineau—at least as originally articulated—which concerns their accounts of the source of teleology for acquired mental representations. While Papineau proposes that learning processes are selective, Millikan suggests that acquired representations have functions derived from the functions of the innate mechanisms that are ultimately responsible for their acquisition and deployment. I discussed Millikan’s account of derived functions in Chapter 4, Section 4.6. Recall that,

83 As Justine Kingsbury points out, Millikan’s Normal conditions are neither necessary nor sufficient conditions for a representation performing its function (Kingsbury 2006). The reason they are not necessary is that the token of a representation may perform its function abnormally in situations when the Normal condition does not obtain. For example, a beaver may mistakenly token an alarm call, causing all the family to dive for safety, after hearing a branch fall from a tree, but by pure accident save the family from an undetected predator that was lurking around. It is easy to imagine analogous situations involving human beliefs.
according to Millikan, an item $A$ has a *derived function* to do $F$ if and only if $A$ is the product of a device $M$ that has the *direct function* of producing items like $A$ in order to do $F$ (Millikan 1993: 13-14). (Strictly speaking, the device $M$ may also have a *derived* function to produce items like $A$, but then such a function will have to be derived from another device that has the function of producing devices like $M$… Derived functions must be ultimately anchored on the *direct* functions of mechanisms that have been historically built and maintained by a selective process.)

For example, the color pattern of the skin of a Smith’s dwarf chameleon has the function of protecting the chameleon from predators by making it blend with its surroundings. But the pattern has not itself been *selected* for doing this. It is not itself a reproduction of earlier patterns and, in fact, it may be a pattern that has never occurred before in the evolutionary history Smith’s dwarf chameleons. However, the pattern is the product of innate mechanisms that have the function of producing such color patterns in order to protect chameleons from predators. The mechanisms have been built and maintained by natural selection precisely for doing this. Consequently, the color pattern has the *derived* function of protecting the chameleon from predators. According to Millikan, the relationship between the functions of novel representations and the functions of the innate mechanisms responsible for producing them is of the same kind (Millikan 1984: Ch. 2 & 1993: Ch. 4).

Consider the functions ascribed by Millikan’s theory to human desires and beliefs: the function of desires being to bring about their satisfaction conditions by participating in inferences that combine them with relevant beliefs to produce decisions to act; and the function of beliefs being to participate in inferences to produce other beliefs and to help produce the fulfillment of desires. As far as *acquired* beliefs and desires are concerned, these are, according
to Millikan, derived functions that these states have because the mechanisms that produce them have the function of producing such states as means to perform their ultimate functions—which we may characterize, very roughly, as modulating behavior in ways that are appropriate, given varying conditions in the world, to satisfy our varying needs. The machinery involved in the deployment of human mental representations is orders of magnitude more complicated than the machinery involved in the deployment of the basic internal representations discussed in Chapter 5—such as the frog’s “bug detector”. It involves a whole system of beliefs and desires composed out of concepts and interlinked with each other in inferential processes. This complexity makes the system considerably more flexible. But at bottom it works, just like the most basic systems, by mediating between inputs from perception and outputs to behavior, according to certain principles of operation:

The human’s mainframe takes, roughly, stimulations of the afferent nerves as input… It responds, in part, by developing concepts, by acquiring beliefs and desires in accordance with these concepts, by engaging in practical inference leading ultimately to action… When conditions are optimal, all this aids survival and proliferation in accordance with a historically normal explanation, one of high generality, of course. (Millikan 1993: 95)

Millikan’s very reasonable contention is that the innate human mechanisms responsible for our development of concepts and our acquisition of beliefs and desires have been built and maintained by natural selection—i.e., “designed”—precisely for doing this, so that producing novel concepts, beliefs and desires are among their direct functions. After all, the extremely complex “design” of the human brain responsible for our acquisition of concepts, beliefs and
desires calls for an explanation. And the only viable explanation is the Darwinian one.

As I said above, the account of belief content in terms of Normal conditions needs to be adjusted in the case of acquired beliefs. Since an acquired belief has no selective history of its own on Millikan’s model, the crucial Normal condition for the performance of its derived functions cannot be one that figures in a Normal explanation of how token beliefs of the same type historically managed to perform their functions. How can there be Normal conditions for a device performing a derived function? Millikan’s answer relies on the Normal explanation of how the mechanisms that produce the device are supposed to work. In the case of the chameleon, for example, the pigment rearrangement mechanisms are supposed to perform their function—of protecting it from predator detection—by means of producing patterns of skin color that bear the relation of having the same-pattern-of-color-as the surface the chameleon is sitting on—whatever it may be. The Normal condition for a particular and novel pattern of skin color performing its derived function is, consequently, that is does bear that relation to the surface the chameleon is sitting on (Millikan 1984: 42-44). For example, if the chameleon’s skin is currently yellow with black dots, then its Normal condition for performing its derived function—of protecting the chameleon from predators—is simply that the chameleon is sitting on a surface that actually is yellow with black dots.

Even though the chameleon’s skin is not a representation, the situation of acquired beliefs is analogous in Millikan’s account: the crucial Normal condition to perform their derived functions—of participating in inferences to produce other beliefs and to help in the fulfillment of desires—is that they bear a certain general type of relation to the world. Millikan proposes that the relevant relation in the case of beliefs is that of “corresponding to” or “mapping onto”
portions of reality—because the Normal explanation of how mechanisms that produce beliefs historically managed to perform their functions is that they produced states that bore such relations to reality. Millikan’s account of the relation between true beliefs and reality is along the same lines as Papineau’s. I will say a bit more about it in a moment.

While Millikan relies on functions derived from innate biological mechanisms to extend her teleosemantics to cover acquired mental representations, she is by no means against postulating selective processes that work outside the bounds of biological intergenerational natural selection. Actually, Millikan relies on non-biological selection to offer a teleosemantic account of the meanings of linguistic representations. Her basic idea is that the elements that compose a language, including words and syntactic forms, are themselves reproduced items that spread across linguistic communities by being copied from user to user—and from generation to generation—forming causal chains of use that constitute lineages or, as Millikan calls them, “reproductively established families” (Millikan 1984: 19):

The selection of language forms takes place on the social level. Language survives when it serves cooperative functions often enough, functions that reward at once both speakers and hearers… Language forms proliferate when aiding speaker and hearer cooperation on common projects—typically, the sharing of information… or the coordinating of projects and activities… (Millikan 2005: 85)

Basically, language forms proliferate in this historical process because they are socially selected to serve “cooperative functions” involving communication. As reproduced items that get historically selected in order to serve certain purposes, linguistic representations have direct
functions. Recall that Millikan’s account of a direct function is formulated in general enough terms to cover the functions of any reproduced items that exist because of being favored by one or another selective process: an item $A$ has a *direct function* to perform $F$ if and only if $A$ is a reproduction of prior items that performed $F$ in the past and $A$ exists because such prior items performed $F$ (Millikan 1993: 13).

Millikan suggests that the main function of “descriptive” language forms—i.e., the utterance of indicative sentences—is to produce true beliefs in hearers, while that of “directive” language forms—i.e., the utterance of imperative sentences—is to produce compliance by hearers (to make them behave in particular ways). I will focus on indicative sentences. Millikan relates their function and meaning as follows. Their truth-conditional meaning is determined by the crucial *Normal condition* for such sentences to perform their function—of inducing true beliefs in hearers—in accordance with a Normal explanation: namely, that they bear a relation of correspondence to reality in accordance with certain “mapping rules” (Millikan 2005: 63).

Even though the functions of beliefs and indicative sentences are different, their truth-conditional meanings are alleged by Millikan to be determined in both cases by the Normal conditions that are supposed to obtain for these representations to perform their functions. In both cases, the relevant Normal conditions are determined by “mapping rules” that relate the complete representation to specific portions of reality based on its structure and the referential properties of its parts—referential properties that are reductively explained in terms of the *teleonomic links* between concepts or words and portions of reality such as individuals, stuffs, kinds and properties (Millikan 2000: 1 & 2005: 66). One of the features that sets apart mental and linguistic *human* representations, Millikan suggests, is their logico-syntactic structure: they
have subject-predicate form and can be negated (Millikan 1984: 308-309 & 1993: 100-101).
Notice that, while expanded to cover also linguistic representations, this is roughly along the
same lines as Papineau’s account of the truth-referential contents of thoughts. Both theories can
be regarded as contributing to the project of reductively explaining reference and truth-
conditions along the lines of a contemporary version of the correspondence theory of truth—
which combines a Tarskian characterization of truth-conditions based on logico-syntactic
structure and referential properties with a reductive account of reference (see Chapter 2, Section
2.2). Surprisingly, however, Millikan often characterizes her theory of truth-conditions more
ambitiously: as one that involves isomorphisms or picturing relations. I will critically assess such
characterization of her theory in Appendix 1—where I will argue that, despite her rhetoric,
Millikan’s theory does not actually rely on picturing relations.

Millikan’s account of the meanings of linguistic representations, based on their direct
functions qua reproduced devices, is intended to explain the literal meanings of linguistic
representations. It is worth comparing briefly Millikan’s teleonomic account with what may
deserve to be called the “standard” account of the source of literal meanings. The standard
account is that the literal meanings of utterances depend on conventions that link them to the
meanings of speakers’ thoughts. There is a lot to say in favor of this view. On one hand, prima
facie it looks like we use language to communicate what we think: to express our thoughts. On
the other hand, the arbitrary linguistic signs we use to achieve this goal are obviously a matter of
social convention. As David Lewis put it: “It is a platitude that language is ruled by convention”
(Lewis 2002: 1). In English, for example, we use sentences containing the word ‘dog’ to
communicate thoughts about dogs. Other languages use different words with this same purpose.
And in English any other arbitrary sign could have been used for this purpose. To explain how this comes about, we need explanations of how we can use utterances to communicate thoughts and of how linguistic conventions can emerge from our communicative practices. The work of Paul Grice (1989) is often regarded as helping to solve at least the first part of the puzzle and the work of Lewis (2002) as helping to solve the second part.

Grice proposed an influential account of what speakers mean by their utterances, based on complex communicative intentions. The basic idea is that a speaker $A$ means that $p$ by an utterance $x$ on if “$A$ uttered $x$ with the intention of inducing [in the hearer/s] a belief [that $p$] by means of the recognition [by the hearer/s] of this intention.” (Grice 1989: 219). Grice proposed that the literal meanings of expressions could be explained as regularities that emerge from what speakers mean based on their communicative intentions: “[to] say what a word means in a language is to say what it is in general optimal for speakers of that language to do with that word… what particular intentions on particular occasions it is… optimal for them to have” (Grice 1989: 299). But this does not seem like a plausible mechanism for literal meanings to emerge from speakers’ meanings: it seems to over-intellectualize linguistic competence (Millikan 1984: Ch. 3; Horwich 2005: 57-62).

Lewis work on linguistic conventions offers an interesting alternative to Grice. Lewis suggests that conventions are arbitrary regularities of behavior that are perpetuated because they solve recurring coordination problems among people. For example, the convention of driving on the right side of the road arises from the coordination problem faced by drivers wanting to avoid colliding with each other, and it is arbitrary because driving on the left would have worked equally well. Once a solution to the coordination problem arises, it serves as a precedent that
people continue relying on when the problem recurs, so a behavioral \textit{regularity} emerges. The particular coordination problems that language conventions solve are related to communication. People want to share thoughts with each other, so certain regularities in the use of words emerge to facilitate this task. Lewis idea is that speakers conform to certain regularities in their use of words because they have managed to truthfully communicate what they think by conforming to those regularities in the past (Lewis 2002: 179). Just like in the case of driving on the right side of the road, other regularities would have worked equally well—e.g., the word ‘dog’ could have been used in English to share thoughts about cats. But once a solution arises—e.g., using the word ‘dog’ to share thoughts about dogs—, it will serve as a precedent that people will continue to rely on to solve the same coordination problem when it arises again—e.g., sharing thoughts about dogs—; so a linguistic convention will emerge. The crucial advantage of Lewis’ account over Grice’s is that it relies on \textit{precedent}—so that speakers and hearers do not need to keep relying on complex nested intentions on every communicative interaction. Once a convention has emerged, speakers and hearers can simply rely on the available recipe.

Millikan’s account of the linguistic conventions responsible for the literal meanings of words is similar in some respects to Lewis’. First, she also proposes that linguistic conventions emerge as solutions to \textit{coordination problems} related to communication. Second, she also regards linguistic conventions as patterns emerging due to the weight of \textit{precedent}—e.g., the word ‘dog’ is used to share information about dogs because it has been successfully used in the past with that purpose. But the crucial difference with Lewis’s account is that Millikan does not characterize linguistic conventions as mere regularities, but as reproduced patterns that proliferate because they perform coordinating functions. Lewis initially proposed that a pattern
of behavior is conventional if everyone conforms to it (Lewis 2002: 42). This is too strong. He later relaxed this requirement, suggesting that a behavior is conventional if almost everyone conforms to it (Lewis 2002: 78). But even this relaxed requirement turns out to be too demanding. Speakers may often fail to conform to the conventional pattern, because they misspeak, they misuse words due ignorance or error, they commit grammatical mistakes, and so forth. Surely the literal meanings of words are a matter of conventions governing their use. But an account of such conventions as regularities cannot make enough room for misuses of words.

One alternative is to treat linguistic conventions as involving norms instead of regularities. A clear example is Robert Brandom’s account, according to which meanings are engendered by norms governing the use of words. This approach has the advantage of making room for misuses of words: the existence of norms, after all, does not require almost everyone to conform to them. But the approach also has a serious drawback: while Brandom offers a “reductive story about norms as instituted by social practices”, he insists on “the irreducibly normative character of the metalanguage in which norm-instituting social practices are specified” (Brandom 1994: 626)—which entails that meanings cannot be reduced to natural facts. As Brandom acknowledges, his theory is “opposed to naturalism” (Brandom 2000: 26).

In Millikan’s account, like in Brandom’s, the conventions governing the use of words transcend mere regularities and provide a basis for distinguishing correct uses from misuses. But Millikan’s account, unlike in Brandom’s, is fully consistent with naturalism. Instead of relying on prescriptive norms that are irreducible to natural facts, it relies on what we might call “teleonomic norms”. As Millikan puts it:
[The] central norms applying to language are… like those norms of function and behavior that account for the survival and proliferation of biological species… Specific linguistic forms survive and are reproduced together with cooperative hearer responses because often enough these patterns of production and response benefit both speakers and hearers. Like conformity to… biological norms, conformity to these patterns need not be uniform or even average. In some cases conformity may not even be particularly common. Conformity is needed only in a critical mass of cases, enough to insure [sic] that the cooperative use constituting the norm—the convention—continues to be copied… (Millikan 2005: vi)

If the conventions that govern the use of words are *teleonomic norms*, then they establish how words are *supposed* to be used. This has the advantage of making plenty if room for misuse while anchoring the literal meanings of words on natural facts. Notice that Millikan’s account of the literal meanings of words in human languages is at its core analogous to her account of the meaning of simple animal languages. The main differences are that in the case of human languages the reproduction is cultural rather than biological and the *teleonomic links* are complex: the teleonomic links between sentences and states of affairs are determined by the logico-syntactic structure of sentences and the reference-determining teleonomic links between their words and the individuals, stuffs, kinds or properties that (allegedly) make up those states of affairs. The same compositional “mapping rules” apply to her account for the contents of thoughts. Millikan tends to present her theory as a teleosemantic version of a “picture” theory of truth-conditions. This is a problematic aspect of her work that I will leave to discuss in Appendix 1. Setting that issue aside, the main point is that on Millikan’s account the meanings of concepts
and words are basically their referential properties, which Millikan explains in terms of their *direct teleonomic links* to entities in the world.

### 6.5 Assessing Papineau’s and Millikan’s Teleosemantics

In the previous two sections I elaborated on the attempts by Papineau and Millikan to offer teleosemantic accounts of the meanings of sophisticated human representations. In this section, I will briefly assess their proposals regarding how to extend the etiological account of functions to human representations and discuss some problems with—and limitations of—their theories of meaning. I will argue that there is good reason to expect a teleonomic account of human cognition and communication to succeed. But I will argue that the specific accounts of meaning that Millikan and Papineau offer are inadequate because (1) they rely on reference-determining *teleonomic links* that abstract away from the causes of representations, (2) they do not offer adequate accounts of the meanings of non-primitive symbols and (3) they do not ascribe meanings that could plausibly govern the overall use of symbols in inference. The first problem is related to the concerns of Pietroski (1992) and Neander (1995) about output-based teleosemantics that I discussed in Chapter 5. The second and third problems are specific to human symbols: they are related to the generality and explanatory power considerations that favor—as I argued in Chapter 3—a basic-acceptance approach over other approaches.

Papineau and Millikan make interesting proposals regarding the source of teleonomic functions for acquired mental representations. Millikan even suggests a bold account of the source of teleonomic functions for linguistic representations. Are their proposals plausible?

Let us begin with Papineau’s suggestion that learning processes are selective in nature. As
I pointed out in Section 6.3, learning by operant conditioning exhibits striking similarities to natural selection. As Skinner observed, it is able to “select” behaviors based on their effects, just like it happens in the natural selection of hereditary traits, but during the life of individual creatures (Skinner 1953 & 1966). But does the similarity between operant conditioning and natural selection extend to other forms of learning and, in particular, those involved in the development of human concepts and the acquisition of human beliefs and desires? It is very plausible that processes involved in the acquisition of some mental representations are selective. Recall Dretske’s example of a pigeon that learns to get food by pecking on a screen if and only if pictures of trucks are displayed (Dretske 1988: 153). If conditioning can achieve such results in a humble pigeon, it is not hard to imagine that humans could develop perceptual and perhaps conceptual skills through conditioning in similar ways. Another form of learning that is shaped by effects and can plausibly be seen as a selective process is trial and error learning. Perhaps some acquired human beliefs are fixed by some form of trial and error. Nevertheless, the processes involved in the acquisition of some representations do not seem to be selective. Suppose that someone learns that elephants can swim after a single encounter with a swimming elephant. This example illustrates a standard way in which we form many of our beliefs: we combine concepts we already possess in novel ways as a result of our interactions with the world. Yet, there is nothing here that resembles natural selection. The belief does not seem to be selected from competing variations and no beneficial effect is responsible for its fixation. Acquiring a true belief is, of course, something that is potentially beneficial. But the benefit may come after the belief has been fixed or never. Another example of a belief that is not acquired through a selective process comes from Papineau himself, who proposes that a function of the
inferential mechanisms is to generate new beliefs about means to satisfy desires out of background beliefs. Clearly, such inferential processes are not selective.

Because some acquired representations are not selected by a learning process, they cannot have *direct* functions. Papineau’s proposal has the advantage of being simpler than Millikan’s, but it cannot account for *all* acquired representations. That is why Millikan’s proposal that acquired representations have functions *derived* from the mechanisms responsible for their acquisition is crucial to extend teleosemantics to human mental representations. And there are compelling reasons to support Millikan’s proposal. Human beings are only able to develop new concepts and acquire new beliefs and desires during their lifetimes because they are born with complex brains that come equipped with innate mechanisms that make this possible. This is beyond dispute. But once we acknowledge this, it becomes clear that there must be some explanation of why we have such mechanisms. Pure chance is not a good explanation. But even if some complex trait happened to be built by pure chance, that would not account for why it was *maintained*. But there is the reasonable concern that what look like complex traits that constitute adaptations—that have been selected for—may instead be *byproducts* of the selection for other traits (Gould & Lewontin 1979; Gould & Vrba 1982). To assess this concern, it is crucial to distinguish between two very different kinds of byproducts: (1) those that originated as byproducts of other traits but have been subsequently maintained by natural selection because they turned out to be adaptive or beneficial for survival, and (2) those that continue to exist without any selection for them—so-called “spandrels”—because they continue to be mere byproducts of other beneficial traits. If the complex mechanisms responsible for the acquisition of new human representations were of the former kind, then their maintenance by natural
selection would be enough to confer teleonomic functions to them—and derived functions to their products. If instead they were of the latter kind, then they would not have teleonomic functions.

A point that often goes unnoticed is that the claim that a complex trait is merely a “spandrel” is a very bold hypothesis—as bold and in as much need of evidence as the hypothesis that the trait is an adaptation. The mere abstract possibility of some traits being “spandrels” does not seriously undermine teleosemantics, because it is prima facie implausible that having acquired concepts, beliefs and desires in our cognitive repertoires does not confer an advantage for survival: “To suspect that the brain has not been preserved for thinking with—… in the absence of any alternative hypothesis—would be totally irresponsible” (Millikan 1993: 96). The best available hypothesis we have by far—about why our brain comes equipped with the innate mechanisms it does—is that such mechanisms have been selected for producing concepts, beliefs and desires that help us to meet our needs in varying circumstances.

Analogous considerations apply to acquired linguistic representations, because having words and ways of combining them into sentences to communicate with each other surely confers as much of an advantage for survival as the simpler innate languages of other animals do. Additionally, even if linguistic representations are themselves culturally reproduced items, it is important to also take into account that our capacity to learn a language is innate. As much as linguistic representations may be said to have direct functions that stem from their own selective history, they can also be said to have derived functions from our innate language faculty. I consider the view that there are non-biological selective processes—such as some selective learning processes (Papineau) and the socio-cultural reproduction of linguistic symbols
Granting that innate and acquired human representations have teleonomic functions, it remains to be seen what is the proper account of their meanings. While Papineau and Millikan have made an extremely valuable contribution to semantics, I think that their theories face some very serious problems that show that a different kind of teleosemantics for human representations is needed. The first problem stems from the purely output-based character of their accounts. In Chapter 5, I argued that Millikan’s “consumer-based” teleosemantics for basic representations has the problem of ascribing implausible contents that are detached from the cognitive capacities of the mechanisms that produce them. As Pietroski (1992) showed, the ascriptions of content made by Millikan’s theory can be implausible, because they are unrelated to the capacities of the mechanisms that produce them. In his example, the kimus’ state that is caused by red surfaces is adaptive because it led kimus to snorf-free areas, so the states mean “snorf-free area”, even though kimus are utterly unable to tell apart a snorf from anything else. I argued that the ascriptions of Millikan’s theory—based exclusively on the “Normal conditions” for the performance of the representation’s evolutionary function—are indeed implausible because the causes of perceptual representations do not play any role in them. I proposed in its place a hybrid input-output theory according to which the content of a basic representation is determined by its Normal cause—namely, what causes tokens when they manage to produce the effects the representation has the function of producing. These are also teleonomic links, but they are causal.

As it turns out, Papineau’s and Millikan’s output-based teleosemantics for human representations face the same problem with causes playing no role in meaning determination.
Millikan’s “Normal conditions” link words/concepts to their referents abstracting away from any causal connection. Papineau’s theory has the same problem. While discussing what his theory and Millikan’s have in common, Papineau (writing with Graham Macdonald) says:

[The] approach… dissociates the determination of content from input conditions… content depends on how consumer mechanisms interpret representations… on the behavioural output, not the informational input. The content is that condition under which the resulting behaviour would be appropriate, whether or not the actual circumstances that caused the representation are of that type. (Macdonald & Papineau 2006: 6)

What Millikan and Papineau tell us is that the teleonomic links that determine meanings are links to what concepts are supposed to be correlated to in order to perform their functions, regardless of whether those correlations are causally based or not. But this is precisely the kind of view that leads to the problem pointed out by Pietroski and Neander: ascribing meanings that are implausible because they are detached from the cognitive capacities of the mechanisms responsible for producing the representations. In the case of human representations linked to inputs from perception, the problem is clear: the theory detaches their meanings from what the perceptual mechanisms causally interact with. Consider the concept of darkness. Suppose, as it is likely to be the case, that many human infants have an innate tendency to fear being alone in

84 As far as I can tell, the views of Millikan and Papineau were output-oriented from the beginning—unlike, say, the clearly input-oriented view of Neander. However, Millikan seems to have been the one that originally rejected most adamantly the idea that causes may play a role in determining content. Interestingly, at the same time that Papineau (1987) explained truth-conditional content in terms of co-presence rather than causation, we can find remarks where he seems to assign a crucial role to causation. In a passage he remarks, for example, that the explanation in terms of natural selection “gives us a substantial distinction between ‘normal’ and ‘abnormal’ causes” which are related to circumstances when a belief “had advantageous behavioral effects” (Papineau 1987: 65). This is very much in tune with the core idea of my hybrid input-output approach! However, these passing remarks did not seem to be articulated into his official doctrine.
open dark places. Suppose that such tendency is meant to protect us from night predators, by making us avoid staying in places where we would be vulnerable to them. Now, such tendency will play a role in the development of the concept of a dark place and of thoughts such as the belief that I am in a dark place and the desire to leave such a place as soon as possible—which combined in inferential processes will on various occasions make us leave dark places. Perhaps this sort of effect saved the lives of many of our ancestors. (We may assume that the learned concepts and acquired beliefs have derived functions or direct ones—following Millikan or Papineau.) But according to the output-based theory it looks like the concept of darkness, or of a dark place, led to beneficial behaviors because it was correlated with night predators. So it looks like the meaning-determining teleonomic link is with night predators. But this is an extremely implausible ascription. Our pre-theoretical ascriptions relate the concept to darkness. We can perfectly explain the behavior in terms of the fear of dark places. Of course, we can also explain the evolution of such behavior based on the avoidance of night predators. But what we explain then is not that the concept of darkness refers to predators, but rather that the evolutionary explanation for our desire to avoid darkness is that it prompted behaviors that were adaptive because they helped us to avoid night predators. An output-based teleosemantics cannot distinguish cognitively relevant links from other evolutionarily relevant ones. Consider another example: the desire for soda, derived from an innate desire for sweetness together with a belief that soda is sweet. Surely the contents of the thoughts that make us drink soda are not about nutritiveness—even if the evolutionary explanation of why we have those desires is one where the goal of getting nutrients plays a crucial role. Papineau is willing to go along with Millikan and bite the bullet in the case of basic representations, but he argues (together with Graham
Macdonald) that Pietroski’s objection does not apply to representations that belong to sophisticated systems of representation: if kimus had a color-perception system serving various functions, then their states would mean “red” anyway (Macdonald & Papineau 2006: 9). But this reply does not suffice to explain why the kimus’ states do not also mean “snorf-free area”. If there is a history of selection related to the absence of snorfs, the states presumably acquire that meaning regardless of also having other contents and functions.

The problem I just discussed already affected output-based teleosemantics for basic representations. The following two problems I will point out are specifically about human representations. In Chapter 3, I argued that a theory of meaning articulated in terms of basic sentence acceptance has two advantages over other theories. First, it is able to account for different kinds of words/concepts having meanings that relate them either to the external world, to other words/concepts or to rules of inference. Second, it is able to explain how the meanings of words/concepts govern their overall uses in inferential processes. As it turns, the teleosemantic theories of Papineau and Millikan are at a disadvantage: they do not offer plausible accounts of non-primitive symbols and the meanings they ascribe even to primitives do not govern their use in inference.

The views of Papineau and Millikan are very different regarding the issue of non-primitives. Millikan seems to defend an extreme form of semantic externalism according to which content is just reference and all referring terms are primitives. (It will become clear below why I say “seems”.) Millikan (2000), for instance, argues that “substance” concepts—i.e., concepts of individuals, stuffs and natural kinds—have the purpose of enabling us to re-identify things, and that this purpose is independent of the various ways in which we may achieve it. It is
independent of any particular causal links through perception and from any associated beliefs or concepts. She argues that reference is not determined by any modes of presentation. All the causal and conceptual links of a substance concept, she claims, are part of a non-reference determining conception. (See Millikan 2000: Ch. 11 & 12.) When talking about the literal meanings of words, she candidly claims: “[The] public meaning of a referential term… is essentially just reference. I intend this sweeping assertion to include terms for properties, kinds, stuffs, and so forth…” (Millikan 2005: 66). These views seem extremely radical. What about the word ‘bachelor’ and the concept it expresses? Is Millikan willing to say that ‘unmarried man’ is merely part of a non-reference determining associated conception? And what about the terms of logic such as ‘and’? As I argued in Chapter 3, a general theory of meaning must cover all of these sorts of words/concepts. Surprisingly, Millikan herself comments on an endnote: “We should also note that there are some terms—‘bachelor’ is the favorite example—that do seem simply to be shorthand for their standard public intensions.” and speculates that “terms of this kind are not very common” (Millikan 1984: 340n). Regardless of how common they may be or not, we need a theory of meaning to accommodate them. But Millikan does not comment on how to do so within her theory. Given her views, there should be a teleosemantic story to be told here. After all, the word ‘bachelor’ is a reproduced device presumably with a literal meaning governed by a teleonomic norm, while the concept it expresses is tokened by mechanisms with teleonomic functions. But the story cannot be the same story that Millikan offers for primitives.

Papineau, in contrast, offers a different theory for “non-observational concepts”, based on their inferential links to other concepts (Papineau 1987: 78-80). What he proposes is in fact a kind of description theory. Interestingly, he suggests it is a teleosemantic theory. The purpose of
non-observational concepts, he suggests, is to “enable us to respond to features of the world to which we only have indirect inferential access” (Papineau 1987: 80). This sounds promising, surely there is such a thing as the misapplication of concepts covered by a description theory. An explanation of their meanings based on what inferential links they are *supposed* to have would be interesting. But, unfortunately, the version of the description theory that Papineau adopts is hopeless. He surprisingly claims that “a non-observational concept refers to that entity whose role in the causal structure of the world mirrors the role of the concept itself in the cognitive structure” (Papineau 1987: 93). Why not go instead with something closer to a classical description theory? Papineau’s version does not seem to work for many non-observational concepts. As Devitt objects, ‘bachelor’ “probably depends for its reference on its causal links to ‘unmarried’ and yet that causal link is not ‘mirrored’ by one between bachelors and unmarried objects” (Devitt 1991: 432). Additionally, Papineau’s version of the description theory is *holistic* (see Papineau 1987: 97-98). As Devitt complains: “Papineau… is committed to an extremely individualistic version of holism: if the functional role of a concept in your head, or in my head last week, differs in the slightest degree from one in my head now, then the concepts must differ” (Devitt 1991: 432). These problems, however, stem from the version of the description theory adopted by Papineau, rather than the teleosemantic twist he suggests for it.

Setting aside non-primitives, the final problem with the theories of Papineau and Millikan is that the meanings they ascribe are not causally responsible for governing the use of concepts in inference. That this is the case is already clear given the fact that Papineau and Millikan regard the teleonomic links that determine meaning as causally detached from tokens, as I discussed above. But even if their direct links were causally anchored, we would still have a problem.
Imagine, for example, that we modified Millikan’s theory replacing her teleonomic links with the teleonomic causal links of my hybrid account. These would still be links between isolated tokens of the concept and its perceptual causes—e.g., between tokens of ‘dog’ in the belief-box and dogs. But the overall use of a word/concept in inferential processes—as I argued following Horwich in Chapter 3—can only be governed or shaped by certain sentences containing it. For example, the use of ‘dog’ in inference may be governed by tokens of ‘This is a dog’ in the belief box—tokens that may in turn be caused by the perception of dogs.

6.6 Basic-Acceptance Teleosemantics

This has been a long journey, in which I have assessed the virtues and problems of Horwich’s use-theory and available versions of teleosemantics. It is time to tie up loose ends. I will start with a review of the preliminary conclusions of the previous analyses, articulate the combination of basic-acceptance semantics and teleosemantics, and point out some advantages of the view over Horwich’s theory and existing teleosemantic theories. I will end with some reflections on the roles of reference and truth in the resulting account.

In Chapter 1, I argued that meanings are natural worldly phenomena because they play causal roles in the explanation of behavior. In Chapter 2, I argued that the task of semantics is to explain the underlying natures of the meanings identified pre-theoretically by the causal roles of mental and linguistic representations. As folk semanticists, we identify samples of meanings in attitude ascriptions that have the purpose of explaining other people’s behaviors and using their thoughts and utterances as guides to reality. I argued that an explanation of the underlying nature of these meanings should account for the causal roles used to identify them. In the particular case
of words and concepts, I argued that we identify their meanings by the way they are used and as what is causally responsible for their uses. This sets a clear target for an explanation of the underlying nature of their meanings: it must account for what non-semantic phenomena explain the overall uses of words and concepts and, consequently, constitute their meanings. Notice that all of these considerations regarding the identification of meanings and the task of semantics do not assume any views about the nature of truth and reference. While the mainstream view in semantics is truth-referentialism, which requires truth and reference to play substantial explanatory roles, the identification of the task of semantics should not presuppose any such view, which would beg the question against deflationists, who argue that truth and reference are not properties that could play such roles and usually advocate for “use” theories of meaning where truth and reference allegedly do not play any substantial roles. I argued that we are likely to find out about the nature of truth and reference precisely by finding out if we need to appeal to them in the explanation of meanings.

In Chapter 3, I argued that Horwich’s version of the “use” theory of meaning has crucial advantages over other available accounts. Horwich proposes that the meanings of words are constituted by their roles in certain basic or core sentences containing them, which are causally responsible for the rest of their uses—or, as Horwich likes to put it, they govern the overall use of words. Horwich’s theory is meant to apply to words in natural languages and also to words in the language of thought. He largely treats language and thought as a “seamless whole” which is appropriate as a simplification for some explanatory purposes, because language and thought are indeed closely interlinked. But I agree with Devitt’s point that in an ultimate analysis an acceptance-based theory should treat thought as more fundamental (Devitt 2002: 109-111). I will
Horwich’s main reason for suggesting that the meanings of words must be constituted by the acceptance of some fundamental sentences containing them is that most uses of words are a result of their participation in inferential processes, so that what constitutes their meanings must be something capable of shaping such use in inference. The crucial point is that only sentences can participate in inferences. Horwich’s insight is that a few sentences containing a word can play the role of something like axioms or premises in inferential processes, thereby governing the overall use of the word. How this works is rather clear in the case of words traditionally regarded as covered by a description theory, such as ‘bachelor’. The sentence ‘The bachelors are the unmarried men’ can serve as a premise in inferential processes if it is accepted by the user: tokened in his or her belief-box. Other uses of the word ‘bachelor’ can subsequently be explained in terms of this fundamental use, together with other factors. But what makes Horwich’s proposal interesting is that it can account for a variety of words. As he suggests, the meaning of words can be shaped by links to the external world. For example, the acceptance of ‘This is red’ when perceiving a red surface is plausibly what explains the overall use of ‘red’. A connection between ‘red’ and red surfaces, by itself, cannot play a role in inference. So it cannot govern the use of the word. Finally, Horwich suggests that the overall use of logical words can also be explained in terms of basic acceptance. For example, the use of ‘and’ may be governed by the acceptance of the two-way schema “p, q // p and q”.

What I just described is the core insight underlying Horwich’s use-theory: that the acceptance of some basic or fundamental sentences containing a word constitutes its meaning. I call this view “basic-acceptance semantics”. The view is recommended by its promise to account
for the overall uses of words—to explain how meanings govern use—and because of its generality—it can account for meanings of words that depend variously on their links to other words, to the external world and to rules of inference.

Horwich’s use-theory has extra commitments. First, it is proposed as a companion to Horwich’s deflationism about truth and reference. So Horwich expects the acceptance of basic sentences containing a word *not* to determine its reference. Second, it is articulated in terms of dispositions and regularities of use. Horwich claims that our disposition to accept basic sentences containing a word is the fundamental law-like regularity governing its use. In Chapter 3, I argued that these two commitments are problematic. Regarding deflationism, Devitt points out that the theory seems to provide the same links that a moderate truth-referentialism would count as reference-determining. So the theory risks collapsing into truth-referentialism. I happen to believe that a collapse is likely, but I argued that it does not undermine basic-acceptance semantics. I also argued that Horwich’s articulation of the theory is undermined by serious problems of ignorance and error that stem from its dispositionalism. People often have dispositions to make mistakes (error) or lack the relevant dispositions to use a word (ignorance), so a theory that reduces meanings to use-dispositions cannot make enough room for misrepresentation. For instance, if someone confuses wolfs and coyotes with dogs, the law-like regularity underlying his use of ‘dog’ will be his disposition to accept ‘This is a dog’ when perceiving dogs, wolfs or coyotes—with the unwelcome consequence that he is not misrepresenting when he applies ‘dog’ to a coyote. Horwich proposes a version of the division of linguistic labor that may be thought to help with this problem. But his account is dispositionalist as well: a speaker’s use of a word can be characterized as a misuse provided that he or she is
disposed to defer to experts. But as Devitt replies, people may make mistakes about when or whom to defer to, and may not be disposed to defer at all.

I suggested in Chapter 3 that a combination of basic-acceptance semantics with explanatory frameworks other than “dispositionalism/regularism” could avoid the problems of ignorance and error. One option would be to adopt a “normativism” according to which meanings do not depend on regularities, but on prescriptive norms governing the uses of words. If the meaning of ‘dog’ is constituted by the norm that we ought to accept ‘This is a dog’ in the presence of dogs, then basic-acceptance semantics could draw the line between correct uses and misuses. But this alternative is unattractive because it involves abandoning the project of naturalized meanings—and meanings play causal roles in the world, so they cannot be irreducibly normative. Another alternative that promises to overcome the problem of misrepresentation, but within the confines of naturalism, I suggested, is to combine basic-acceptance semantics with a teleonomic or teleological account of the functions of symbols.

The suggestion to combine these views led me to explore the teleonomic account of functions and available theories that rely on it to explain meanings. The teleonomic notion of function discussed in Chapter 4 has its home in biology, where it is explained in terms of what traits have been naturally selected for doing. Hearts have the function of pumping blood because that is what the hearts of ancestors did that explains why there are creatures with hearts today. What is of interest to semantics is that this notion of function allows for “quasi-normative” ascriptions of failure to perform a function. If a heart does not pump blood, it is not doing what it is supposed to, given its evolutionary history. Applying this notion to offer a theory of meaning promises to make room for misrepresentation. The simplest cases are those of basic innate
There are two main kinds of teleosemantic approaches that I explored in Chapter 5. On one end of the spectrum, there is Millikan’s account, according to which meanings are determined by the conditions in the world that must obtain for the consumers of a representation to perform their function. For example, the frog’s so-called “bug detectors” represent frog food because it was when frogs captured nutritious bugs by darting their tongue that the consumers performed the function for which the trait was selected. While very attractive at first sight, it turns out that causal links to reality through perception play no role in this account, which can lead to implausible content ascriptions and, more fundamentally, the account fails to achieve a fundamental goal of semantics, which is to account for the causes of representations being tokened. On the other end of the spectrum, there is Neander’s producer-based teleosemantics, which has the advantage of tying meanings to the causal links responsible for the tokening of representations, but which does not make enough room for misrepresentation because it disregards the consumption of representations as irrelevant to their meanings. So the frog’s state, according to Neander, means “small, black, moving thing”, since the frog’s perceptual mechanisms cannot tell apart a flying bug from any other small black thing—even though what the consumers need is to catch bugs. To overcome this dilemma, I proposed a hybrid account according to which basic representations represent what I called their “Normal causes”: namely, what is supposed to cause their tokens in order for their consumers to perform their functions. The hybrid account makes plenty of room for misrepresentation, like a consumer-based one, but ascribes meanings that do play a causal role in the tokening of representations.

The remaining challenge was to explore how teleosemantics can be extended to the
sophisticated mental and linguistic representations of human beings. In this chapter, I presented the proposals of Papineau and Millikan, who agree that desires have the function of producing behaviors that satisfy them and that beliefs have the function of collaborating with desires to produce such behaviors—and are supposed to covary with conditions in the world. But they both ultimately offer compositional accounts according to which the truth-conditions of beliefs (and the satisfaction conditions of desires) are determined by their structure and the referential properties of their parts. At bottom, their theories are about *teleonomic links* between concepts and their referents. The main difference is that Papineau proposes that the source of teleonomic functions for acquired representations is that learning processes are *selective*, while Millikan proposes that the functions of acquired representations are *derived* from the functions of the mechanisms responsible for their acquisition and deployment—like the varying colors of the skin of a chameleon, which have not been selected themselves, but are the product of innate mechanisms that have been selected to protect the chameleon by producing patterns of color that make it less visible to predators. Millikan also offers a teleonomic account of the functions of words, based on the fact that they are culturally reproduced items with a history of social selection. Millikan plausibly argues that the conventions that determine the literal meanings of words are teleonomic norms. I argued that some but not all learning processes are selective, so we need to rely on Millikan’s account of derived functions, although it can be combined with Papineau’s account. There is very good reason to expect the innate mechanisms responsible for the acquisition of concepts, beliefs and desires have the function of producing such states, since human cognition is obviously adaptive. So the most plausible hypothesis is that there are function-conferring selective processes underlying the maintenance of such mechanisms.
Similarly, reproduced patterns of linguistic behavior are at least as adaptive as simpler innate systems of animal communication. So the most plausible hypothesis is that they have teleonomic functions.

On a critical assessment of Papineau’s and Millikan’s accounts of concept and word meanings, I pointed out three serious problems. First, that the reference-determining teleonomic links they rely on are not causal in nature, so their theories have the same problem as Millikan’s consumer-based theory of basic representations. Indeed, those links are not causally responsible for the tokening of concepts/words, and they are detached from the cognitive capacities of the mechanisms that produce them. So, based on the methodological considerations proposed in Chapter 2 and deployed against Millikan’s consumer-based teleosemantics in Chapter 5, I objected that such links cannot plausibly be regarded as meanings. Second, relying on one of the arguments in support of basic-acceptance semantics deployed in Chapter 3, I objected that the theories of Millikan and Papineau fail to be general enough because they do not offer plausible accounts of non-primitives. In Millikan’s case, the theory does not even attempt to offer an account of anything other that primitives. Papineau’s case is different. To his credit, he tries to offer an account for non-observational concepts that is a teleosemantic version of the description theory. Unfortunately, the version of the description theory he adopts fails because it relies on implausible mirroring relations and it is hopelessly holistic. Neither Papineau nor Millikan offer accounts of logical terms. Finally, the theories of Millikan and Papineau do not ascribe meanings that can govern the use of concepts in inferential processes. Most of the uses of concepts are in inference, so direct links to reality—even setting aside the issues mentioned above—are not enough to explain what governs stimulus independent uses.
In summary, the situation is the following. On one hand, a basic-acceptance semantics promises to explain meanings in terms of underlying properties that are causally responsible for the overall uses of words/concepts, which can explain how the meanings of words/concepts govern their overall uses in inference and in a way which is general enough, allowing for meanings to involve links to the world, to other words/concepts or to rules of inference. But the regularist version articulated by Horwich cannot make room for misuses or misrepresentation. On the other hand, teleosemantics promises to offer a naturalist solution to the problem of misrepresentation, but available proposals for human words/concepts are implausible because the links they regard as meaning constitutive are causally detached from the production/tokening of words/concepts, they cannot govern use in inferential processes, and they do not account for the variety of links that a general theory of word/concept meaning should account for.

The solution, I suggest, is to articulate a teleosemantic theory in terms of basic acceptance or—what is equivalent—to articulate a basic-acceptance semantics in terms of teleonomic functions. It is worth mentioning that the idea that a “use” theory of meaning can be coupled with a teleonomic account of the functions of symbols has been suggested before by teleosemanticists, although not by use-theorists. For example, Neander suggests:

[Teleosemantic] theories tend to be versions of causal or informational theories, but teleonomy could also be used to add normativity to a conceptual role theory of mental content. (Neander 1999: 3)

The “normativity” that Neander refers to is, of course, not of the *prescriptive* kind that normative versions of the “use” theory rely on. It is rather the teleonomic normativity involved in there
being, *as a matter of fact*, a distinction between correct and incorrect deployments of concepts. Similarly, Millikan suggests that a functional-role theory of mental content can be coupled with teleology to make room for misrepresentation:

Suppose, for example, that you think of mental representations as items defined in a classical functionalist way, in accordance with patterns of causal/inferential dispositions… Then the teleological theorist, call her Tilly, will come along and point out that surely some of the causal roles of actual representations in actual people’s heads correspond to bad inferences. What you must say, says Tilly, is that what the representation represents is determined by what its causal role *would* be if the head *were* operating correctly, that is, in the way it was designed, by evolution or learning, to operate. (Millikan 2000: 229)

Neither Neander nor Millikan are interested in developing themselves a teleonomic version of functional-role semantics: what we may call a “functional-role teleosemantics”. But they realize that dispositionalist versions of functional-role semantics have problems of misrepresentation. In contrast, theorists that defend naturalistic versions of the “use” theory tend to ignore the problem—or even to dismiss it as a pseudo-problem, like Horwich (*see* Chapter 3, Section 3.6.2). The only “use” theorists that take this problem seriously seem to be those that defend non-naturalist versions that rely on prescriptive normativity, like Brandom (1994 & 2000). It is surprising that “use” theorists themselves have not seriously contemplated the possibility of articulating teleonomic versions of their theories. Some of them even explicitly claim that teleology is incompatible with functional-role semantics. Greenberg and Harman (2006), for example, propose what they consider a broad and inclusive understanding of functional- or conceptual-role
semantics (CRS), which encompasses any theory that regards the content of concepts as determined “by any part of their role or use in thought”, rather than being restricted to those theories according to which content is determined merely by “the role of the mental states or symbols in inference or in other purely internal mental processes” (Greenberg & Harman 2006: 295-296). Under their broad understanding, they point out, indicator or informational theories “count as special versions of CRS” (Greenberg & Harman 2006: 296). However, they go through the trouble of specifically excluding teleological theories from even their broad understanding of functional- or conceptual-role semantics:

Just how inclusive our broad understanding of CRS is depends on how broadly conceptual role or use is understood. For example, *teleological* theories of content give an important role to the evolutionarily determined “function” of symbols or symbol structures, where some such theories understand the notion of the function of a symbol or structure in a way that goes beyond the symbol’s use or role as ordinarily understood… We do not count such theories as versions of CRS. (Greenberg & Harman 2006: 296.)

Greenberg and Harman display a notorious resistance to regard the teleological notion of function as compatible with the kind of naturalist “use” theory they favor. I suspect that their resistance stems from the close relationship between their naturalist “use” theory and classical functionalism, since teleological functions are not merely uses or roles as understood by classical functionalism. According to classical functionalism, mental states are characterized by their functional roles, which consist in their patterns of actual and potential causal relations with other mental states, perception and behavior. Naturalist use theories that regard meaning as determined
by functional roles constitute an application of the general functionalist framework to the explanation of meaning, and they draw much of their plausibility from functionalism. Moreover, “use” theories have traditionally regarded the meanings of expressions as closely related to their inferential roles, and functionalism promises to naturalize inferential roles by reducing them to causal roles. By characterizing functional roles as patterns of actual and potential casual relations, however, classical functionalism and the “use” theories based on it are committed to dispositionalism. It is true that the teleological notion of function goes beyond the one deployed by traditional functional-role theories: the one that has its origin in classical functionalism. But going beyond the dispositionalist notion of function may be a changed that is needed—and not only for the purposes of semantics. As Elliott Sober points out, functionalism about the mind would also benefit from a teleological twist (Sober 1985).

In principle any dispositionalist version of functional-role semantics could benefit from a teleological twist. But as I pointed out in Chapter 3, most of these theories are hopelessly holistic. A teleological twist can only help with problems of misuse and misrepresentation. Any other problems faced by a functional-role theory will presumably be inherited by its teleosemantic counterpart. As I argued in Chapter 3, the best available theory is the non-holistic one offered by Horwich, which relies on the acceptance of basic sentences. It is this version of the “use” theory that I propose to combine with a teleonomic account of functions. Of the many possible forms of “functional-role teleosemantics”, a “basic-acceptance teleosemantics” seems to be the most promising one.

The core feature of a basic-acceptance teleosemantics is that it regards the basic sentences involved in the meanings of words not as what regularly governs their uses, but as
What is supposed to govern their uses. A general statement of the main idea is the following:

The meaning of a word, \( w \), is constituted by the fact that the acceptance of certain basic sentences containing \( w \) is supposed to govern \( w \)’s overall use.

The sense in which some core sentences are supposed to govern the overall use of a word is the same sense in which the heart is supposed to pump blood or the blink reflex is supposed to be triggered by foreign objects touching the cornea. It is a sense in which an item may be supposed do something even if it fails to do it. This is the feature that allows basic-acceptance teleosemantics to make room for misuses of words. Instead of claiming that the meaning of ‘bachelor’ is constituted by the fact that its use is regularly governed by the acceptance of ‘The bachelors are the unmarried men’, basic-acceptance teleosemantics claims that the meaning of ‘bachelor’ is constituted by the fact that its use is supposed to be governed by the acceptance of ‘The bachelors are the unmarried men’. Similarly, basic-acceptance teleosemantics claims that the meaning of ‘horse’ is engendered by the fact that its use is supposed to be governed by the acceptance of ‘This is a horse’ in response to the perception of horses. If someone tends to apply the word also to muddy zebras and odd cows, such uses can be characterized as misuses, because they are not governed by the acceptance property that is supposed to govern them. This account of word meaning promises to maintain the advantages of Horwich’s theory, regarding generality and explanatory power, while overcoming its problems of ignorance and error.

The general statement of the main idea of basic-acceptance teleosemantics I provided above has the advantage of being simple. It provides a convenient way to introduce the idea. But it ignores crucial details, since it does not even use the word ‘function’. As I argued above, there
is very good reason to believe that words do have teleonomic functions based on one or another of the sources of teleonomy for acquired representations proposed by Papineau and Millikan—or, even better, a combination of them. Even our most basic conceptual and linguistic abilities exhibit all the signs of “design” and the only plausible explanation is one that ultimately relies on natural selection and/or other selective processes. But even if we grant that words do have teleonomic functions, this leaves open the question of how such functions are related to the acceptance of some basic sentences containing them. The answer to this question I favor regards the acceptance of basic sentences as the Normal cause of uses of words: as what causes such uses when words perform their functions in accordance with a Normal explanation. Notice that this idea of a “Normal cause” is the same one I deployed in my hybrid account of the meanings of basic representations. The difference is that the Normal causes of basic representations are perceptual causes, while the Normal causes of words deployed in inference are certain sentences in the “belief-box” (although for some words the basic sentences themselves have Normal perceptual causes). I think that reliance on Normal causes provides a good basis to articulate a teleonomic version of basic-acceptance semantics, since basic-acceptance semantics is a theory that is tailored to explain meanings in terms of what causes tokens of words/concepts. Notice that if the acceptance of a basic sentence constitutes the Normal cause of the use of a word, then such sentence indeed is supposed to govern the use of the word. So the general statement I provided above is indeed an accurate, while simplified, representation of the view.

How does a basic-acceptance teleosemantics articulated in terms of Normal causes look like? To simplify exposition, I will first follow Horwich assuming that language and thought form a seamless whole. The appearance that language and thought form a seamless whole may
come from the fact that, on one hand, we use language to express what we think but, on the other hand, we have acquired many of our concepts—or “mental words”—from their linguistic counterparts. But this is not an explanatory circle. It is a “spiral”, since the ultimate origin of newly coined words must be on some thinker that coined the correlated concept and a word to express it (Devitt 1996: 157-158). But the fact that our language and thought are so intimately interconnected means that for certain explanatory purposes we can more or less safely take them as a seamless whole. So in order to articulate a simple version of “basic-acceptance teleosemantics” I will simply talk of sentences, words, and assume that we use the same words to communicate and think with. I will also simply assume that there is a teleonomic story to tell about acquired representations. Let me start with a statement of the account:

The meaning of a word $w$ is constituted by its use in certain core sentences (that link it to the world, other words or rules of inference), because this use is the Normal cause underlying the overall use of the word in virtue of which it performs the functions determined by the history of its use and/or of the underlying cognitive mechanisms responsible for it.

The idea is that certain core sentences are supposed to govern the use of the word because those are the sentences that govern the use of the word when it successfully performs its function in accordance with a Normal explanation. So the quasi-normative teleonomic sense of “supposed” in my previous general statement is explained by how words Normally manage to perform their functions. Notice the difference between the following three statements:

(R) The acceptance of certain basic sentences containing the word $w$ is the
regularity governing the use of $w$.

(P) The acceptance of certain basic sentences containing the word $w$ is the 

*prescriptive norm* governing the use of $w$.

(T) The acceptance of certain basic sentences containing the word $w$ is the 

*teleonomic norm* governing the use of $w$.

(R) states the view adopted by Horwich. But a basic-acceptance semantics can be articulated instead in terms of prescriptive norms, such as (P) or in terms of teleonomic norms, such as (T).

A basic-acceptance semantics articulated as (R) is naturalist, but does not make enough room for misrepresentation. A basic-acceptance semantics articulated as (P) makes room for misrepresentation, but abandons naturalism. While a basic-acceptance semantics articulated as (T) is simultaneously naturalist and makes enough room for misrepresentation.

Let us consider the treatment of a likely primitive, such as the word ‘dog’ by different semantic theories. For a word like this, an indicator theory based on reliable causes would say:

The word ‘dog’ means what it does because its tokens are regularly caused by the perception of dogs.

This theory has, on one hand, a serious problem of error. If someone has a tendency to token ‘dog’ as a result of the perception of dogs, coyotes and wolfs, the result is that, according to the theory, the word means “dog, coyote or wolf”. Additionally, the theory cannot explain stimulus independent uses of the word. Now, this indicator theory can be modified initially in one of two ways. On one hand, we can follow Horwich’s dispositionalist use-theory and articulate the link as one between a basic sentence containing the word and reality:
The word ‘dog’ means what it does because its use is governed by a disposition to token ‘This is a dog’ in the presence of dogs.

This version can explain the stimulus-independent uses of the word, because the meaning ascribed is a property that can causally explain the uses of the word in inferential processes—i.e., in thinking. This is an important advantage of the theory. But Horwich’s articulation does not make room for misrepresentation. Just like with the original indicator theory, if the speaker is unable to distinguish dogs from coyotes and wolves, then the application of the word to coyotes or wolves turns out, implausibly, not to be a misapplication. This is a serious drawback of the theory. On the other hand, we can modify instead the original indicator theory using a version of my hybrid account. I proposed it merely as an account for basic representations. But we could apply it to words assuming, for example, that they have a history of social selection:

The word ‘dog’ means what it does because its tokens are Normally caused by the perception of dogs—i.e., this is the cause of tokens that perform the function that accounts for the proliferation of the word.

This sort of theory promises to make room for misrepresentation, since it states what is supposed to cause tokens of the word. But it has nothing to say about stimulus-independent uses, which is a very serious drawback that makes the theory a non-starter as it is—which is a problem shared with the original indicator theory. Finally, according to basic-acceptance teleosemantics we would get the following:

The word ‘dog’ means what it does because its uses are supposed to be governed by the acceptance of ‘This is a dog’ in response to the presence of dogs—i.e., the
acceptance of ‘This is a dog’ caused by the perception of dogs is the \textit{Normal cause} governing the use of ‘dog’ in thinking.

This view combines the advantages of Horwich’s account and the hybrid teleosemantic account over the simple indicator theory. Like Horwich’s account, it can explain how the stimulus-dependent use governs the stimulus-independent uses in thinking. Like the hybrid teleosemantic account, it makes room for misrepresentation. Surely this articulation of the theory is sketchy. But it is a first attempt in a promising and entirely new direction. When we recount the serious problems faced by available theories, it really looks like exploring further the possibility of combining a basic-acceptance account of meaning with a teleosemantic one is worth it.

Let us assume that words are reproduced items that are culturally transmitted from user to user—and from generation to generation—and that they proliferate because of the benefits they confer to their users as tools for thinking and as tools for communicating with each other. We may then offer the following accounts of the meanings of different kinds of words:

The word ‘dog’ means what it does because the uses in which it performs the functions that account for its proliferation are \textit{Normally governed} by the acceptance of ‘This is a dog’ in response to the perception of dogs.

The word ‘elm’ means what it does because the uses in which it performs the functions that account for its proliferation are \textit{Normally governed} by the acceptance of ‘This is an elm’ in response to the perception of elms.

The word ‘bachelor’ means what it does because the uses in which it performs the functions that account for its proliferation are \textit{Normally governed} by the
acceptance of ‘The bachelors are the unmarried men’.

The word ‘arthritis’ means what it does because the uses in which it performs the functions that account for its proliferation are \textit{Normally governed} by the acceptance of ‘Arthritis is an inflammation of the joints’.

The word ‘and’ means what it does because the uses in which it performs the functions that account for its proliferation are \textit{Normally governed} by the acceptance of the two-way inference schema “\( p, q \iff p \land q \)”.

This articulation is in some regards overly simplistic. It treats the words used for communication and the words used for thinking as the same. And it treats them as reproduced items with their own socio-cultural selective history. The real story is very likely much more complicated than this. However, this simplified story serves the purpose of illustrating how a basic-acceptance teleosemantics promises to explain the uses of words in thinking processes, and the uses of words for communication, all while making plenty of room for misuses of the words and, furthermore, while accounting for a variety of words with meanings involving either links to the world, links to other words or links to rules of inference.

Consider the meaning ascribed to ‘elm’ above. If speakers/thinkers are too ignorant to distinguish elms from beeches—like in Hilary Putnam’s (1975) example—then their application of the word to beeches counts as a misuse because the word has proliferated due to uses that were caused by the perception of elms. Now consider the meaning ascribed to ‘arthritis’ above. Suppose ‘arthritis’ is a term covered by a description theory. This does not prevent users from making mistakes. Yet versions of the description theory that make room for error are difficult to
find. Remember Papineau’s failed attempt at a teleosemantic version of a description theory. While it had problems, the idea of offering a teleosemantic description theory was not itself bad. So, if Bert uses ‘arthritis’ to say something like ‘I have arthritis in my thigh’—like in Tyler Burge’s (1979) example—then he is misusing the word, since the uses that explain the proliferation of the word are Normally governed by the acceptance of ‘Arthritis is an inflammation of the joints’. And it is indeed plausible, I venture to say, that the English word ‘and’ has proliferated because of uses were it was governed by the acceptance of the inferential rule for conjunction. A speaker that mistakenly uses it in some other way—say, as governed by the acceptance of the inferential rule for disjunction—surely is misusing the word.

I have assumed for the sake of simplicity that language and though form a seamless whole and that words simply proliferate as reproduced items. But a better account will have to explain public words as tools for communication and mental words or concepts as tools for thinking. Additionally, while public words are indeed likely candidates for having direct functions, this is not the case for all concepts. These are difficult matters. But I will give an idea of how this may be managed by basic-acceptance teleosemantics.

In the case of acquired concepts, I suggest that basic-acceptance semantics should follow mainly Millikan, proposing that they have particular functions that derive from the general functions of the innate mechanisms responsible for concept acquisition when coupled with experience. Additionally, some concepts may be acquired through selective learning processes, as Papineau suggests—in which case they will have direct functions resulting from their own selective history. But this does not prevent these same concepts from also having functions

85 Devitt and Sterelny point out that the traditional description theory may have problems of ignorance and error even in the case of terms that seem to properly fall under it (Devitt & Sterelny 1999: 94-96).
derived from the very same selective learning mechanisms.

In the case of the arbitrary symbols of natural languages, I suggest that their meanings depend on public conventions that establish them as shared means or tools to express certain concepts, where the linguistic conventions are characterized in terms of the *history* of the use of words instead of regarded as current regularities in their use—following roughly Millikan’s teleonomic account of linguistic conventions. Words spread across linguistic communities by being *copied* from user to user, thus forming lineages or reproduced items. Words proliferate in this historical process because they are socially selected to facilitate communication—the expression of thoughts. Consequently, the process through which a certain use gets standardized as a conventional or public use for a word—e.g., the use to express certain concept—is a process in which the word acquires a historically selected function. Just like with the “learning selection” of some concepts, it is plausible that the “social selection” of words is ultimately possible because people have certain innate mechanisms; in this case, mechanisms that allow people to acquire languages and expand their lexicon, so that words may also plausibly have derived functions from these mechanisms.

The resulting view is, then, that the meanings of words in public languages are determined by historical conventions that link them to words in the language of thought—i.e., concepts—and that the later, in turn, have meanings determined mainly by derived teleonomic functions performed when their deployment in thinking—i.e., inferential—processes is governed by their use in some core sentences that serve as the most fundamental premises in these processes: core sentences that constitute the *Normal causes* of the overall uses of concepts. These core sentences are not themselves the result of inferential processes—in this sense they are *non-
The functions of different words in the language of thought vary considerably. The common function to all words is, I suggest, a very general function—namely, to play certain roles in inference in order to ultimately contribute to the modulation of intentional behavior; since the ultimate effects of inferential processes are decisions to act that lead to the external outputs of the cognitive system: actual behavior. After all, functions are “effects” according to the teleonomic model of functions adopted by basic-acceptance teleosemantics. But my proposal is not to identify the meaning of a word with its function. Nor is it to identify it with a Millikanian “Normal condition” that is detached from the causal-roles of symbols. Instead, I follow roughly the idea I developed in my hybrid account for basic representations, identifying meaning with the Normal causes responsible for words performing their functions—when they do. Borrowing from Horwich’s use-theory of meaning, I propose that the Normal causes responsible for words performing their functions—when they do—involve core sentences governing their use. A word can play a causal role in intentional behavior—and thereby perform its ultimate function—because it is deployed in certain characteristic ways in inference that follow from its use in certain basic sentences that are not themselves the result of any inferences.

I conclude with a provisional consideration about the roles of reference and truth-conditions in basic-acceptance teleosemantics. I pointed out, following Devitt, that Horwich’s version of basic-acceptance semantics risks collapsing into truth-referentialism. I have not articulated basic-acceptance teleosemantics explicitly in truth-referential terms. I am of the opinion that we can continue making plenty of progress in semantics before settling the issue of
deflationism. I very strongly suspect that the direct and indirect links to reality postulated by basic-acceptance teleosemantics will turn out to be reference-determining ones. But I suggest that use-theorists who favor deflationism should explore teleosemantic versions of the use-theory (i.e., versions of “functional-role teleosemantics”), since they promise to overcome the problems of ignorance and error that plague all naturalistic use-theories. I suggest also exploring the possibility of non-reference-determining meanings within the framework of basic-acceptance teleosemantics. But as far as I can see, however, the most reasonable hypotheses about meaning constitutive links seem to be the ones we would expect if truth-conditions and referential relations played substantial explanatory roles in semantics.

A proper truth-referentialist version of basic-acceptance teleosemantics should be moderately truth-referentialist, in order to accommodate words with meanings that link them either to reality, to other words or to rules of inference. It will have the following general shape:

i. The meanings of *primitive words* are constituted by the acceptance of basic sentences containing them with links to reality that are supposed to govern their use and that determine their references.

ii. The meanings of *non-primitives* are constituted by the acceptance of basic sentences with links to other words that are supposed to govern their use and that determine their references.

iii. The meanings of *logical words* are constituted by the acceptance of basic sentences with links to inference rules that are supposed to govern their use and that determine their contribution to the truth-conditions of the sentences they are used in.
Not surprisingly, this is very similar to the shape of a truth-referential basic-acceptance semantics that I offered as an illustration in Chapter 3, Section 3.5.3, but now with a teleonomic twist. The idea is that the very same properties that are responsible for governing the use of a word are what determine its reference. I want to suggest that in a truth-referentialist version of basic-acceptance teleosemantics, the basic sentences that constitute meanings (because they function as the Normal causes of the overall uses of words) should be seen as playing the role of modes of reference. Consider the meaning ascribed to ‘bachelor’ above. The teleonomic link between ‘bachelor’ and the acceptance of ‘The bachelors are the unmarried men’ can be regarded as a descriptive mode of presentation in Fregean style. More interestingly, consider the meaning ascribed to ‘elm’. In this case, the teleonomic link between ‘elm’ and the acceptance of ‘This is an elm’ in response to the perception of elms provides what Devitt (2001b) calls a “causal” mode of reference. Frege regarded the associated description that determines the reference of a word as part of the word’s meaning. So he could explain how co-referential words can differ in meaning: different associated descriptions can determine the same reference. But this only works for non-primitives. Devitt’s idea is that the direct causal link to reality that determines the reference of a primitive can also be regarded as part of the word’s meaning. So we can explain how co-referential primitives can differ in meaning: different causal links can determine the same reference. What I am suggesting is that a truth-referential basic-acceptance teleosemantics as a matter of fact treats the meanings of both primitives and non-primitives as modes of reference.

Basic-acceptance teleosemantics treats words with different basic sentences playing the role of Normal causes as having different meanings, even if they happen to be co-referential. Suppose that the Normal cause of uses of ‘renate’ is the acceptance of ‘A renate is a creature
with a kidney’, while the Normal cause of uses of ‘cordate’ is the acceptance of ‘A cordate is a creature with a heart’. The two words have different meanings but, as it turns out, they have the same reference, since all renates are cordates and vice versa. The old description theory already was able to treat terms such as these as having different meanings or senses, and consequently to explain why the identity statement ‘All renates are cordates and vice versa’ has a different meaning than the trivial statement ‘All renates are renates’. Basic-acceptance teleosemantics offers roughly the same explanation as the description theory in these cases. But now let us focus on primitives: words with meanings involving direct links to reality, which are not covered by a description theory. Suppose the Normal cause of uses of ‘Hesperus’ is the acceptance of ‘This is Hesperus’ in response to the perception of the shiniest star-looking celestial body during the evening, while the Normal cause of uses of ‘Phosphorus’ is the acceptance of ‘This is Phosphorus’ in response to the perception of the shiniest star-looking celestial body during the morning. Since these two words have different Normal causes, they have different meanings according to basic-acceptance teleosemantics, even if it turns out that they both refer to the planet Venus. So basic-acceptance teleosemantics explains why ‘Hesperus is Phosphorus’ has a different meaning than the trivial ‘Hesperus is Hesperus’. A problem faced by “direct reference” theories is that they cannot explain why ‘Hesperus is Phosphorus’ has a different meaning than ‘Hesperus is Hesperus’, since ‘Hesperus’ and ‘Phosphorus’ have the same reference and such theories claim that meaning is exhausted by reference.86 But even a truth-referentialist version of basic-acceptance teleosemantics identifies meanings with the Normal causes of the uses of words, which it takes to be certain basic sentences containing them. If these meanings turn out to

86 Theorists that defend “direct reference” views usually propose that the difference between ‘Hesperus is Phosphorus’ and ‘Hesperus is Hesperus’ is not semantic, but merely pragmatic (e.g., Salmon 1986 & Soames 2002). Devitt (2012: Sec. 3.4 & 3.5) offers compelling arguments against this “pragmatic defense” of the direct reference view.
strongly determine the referential properties of words, they will constitute modes of reference.

6.7 Conclusion

In this chapter, I have discussed how teleosemantics can be extended from the basic representations discussed in Chapter 5 to the sophisticated mental and linguistic representations of human beings. I assessed the valuable attempts by Papineau and Millikan to extend teleosemantics to human representations. While I defended and borrowed their proposals regarding the sources of teleonomic functions in human cognition—as well as the functions of beliefs and desires—I found serious problems with their accounts of concept and word meanings. In their place, I offered a new version of teleosemantics for concept and word meanings that is articulated in terms of basic sentence acceptance. I argued that the resulting view, “basic-acceptance teleosemantics”, overcomes problems of Horwich’s basic-acceptance theory and of the teleosemantic theories of Papineau and Millikan.
Appendix 1:

Millikan’s Pictorial Account

The canonical version of the correspondence theory of truth has been largely abandoned for good reasons. This version relied on alleged relations of resemblance between truth-bearers and facts. One of the boldest attempts to clarify what such relations involved was perhaps Wittgenstein’s *Tractatus Logico-Philosophicus* (1922)—were the correspondence relation was portrayed as involving (1) a relation of structural resemblance between as true sentence and a fact, and (2) referential relations between the words of the sentences and the components of the fact. In Chapter 2, Section 2.2, I discussed the serious difficulties faced by this view—which account for why it has been largely abandoned. However, Tractarian themes have resurfaced in the literature. Wilfrid Sellars, for example, endorsed a picture theory according to which “the manner in which the names occur in the picture is a projection, in accordance with a fantastically complex system of rules of projection, of the manner in which the objects occur in the world” (Sellars 1963: 215). More recently, Ruth Millikan has pursued, as she puts it, “the picturing themes from the *Tractatus* that were carried through in Sellars’ discussions” (Millikan 2005: 77). According to Millikan, “what makes a sentence true is that there is something in the world onto which it maps in accordance with certain mapping functions” (Millikan 1984: 9). She clarifies that what she means by ‘mapping’ is, basically, “as Wittgenstein put it, ‘picturing’.” (Millikan 2005: 63). “The value of a system of representation”, she claims, “depends... on there being some kind of isomorphism... between the domain of the signs and the domain of their signifieds” (Millikan 2004a: 84). Given the insurmountable problems faced by the pictorial account of truth, these confident claims are puzzling. I will argue in this appendix that Millikan wrongly presents her
view as a pictorial one in the tradition of the *Tractatus* and that the relations between true representations and reality that her “mapping functions” actually require are better understood within the framework of a contemporary correlation-based account of truth, of the sort that follows from a Tarski-based characterization of the truth of a sentence in terms of its syntactic structure and the referential properties of its words, combined with a reductive account of reference in terms of substantive relations between words and parts of reality. In this appendix I will abstract away from the other objections I raised against Millikan’s theory in Chapters 5 and 6. The topic of this appendix is merely her claim that her theory involves picturing relations.

Millikan combines what she claims to be a pictorial account of truth with a teleological account of the functions performed by representations, which includes a teleological account of the referential properties of their parts. Her naturalist account of reference is among the main contemporary candidates for reducing referential properties to substantial non-semantic relations. So her “pictorial” view does not suffer from lacking such an account. Millikan’s view is complex and the role of picturing in her account is often overlooked, because most of the interest has been on her theory of reference. Notable exceptions are Peter Godfrey-Smith (1996) and Nicholas Shea (2013), who are sympathetic to some aspects of Millikan’s program but skeptical about the role of isomorphism in her account. Godfrey-Smith and Shea see the teleological part as doing *all* the explanatory work in Millikan’s theory, in such a way that truth-conditions are determined merely by the selective history underlying the mechanisms responsible for producing representations. While their assessments are different, they agree that isomorphism does not play a genuine explanatory role in Millikan’s theory (Godfrey-Smith 1996: 184-187; Shea 2013: 63-80). Their main concern is about the *interplay* between the teleological and the allegedly pictorial
aspects of her theory. But in this appendix I am mainly concerned with the allegedly non-
teleological part of Millikan’s theory.

Millikan regards mapping relations or isomorphisms between representations and reality
as the core features that make representations true, and she intends her teleological account to
explain merely why representations are supposed but may fail to map or be isomorphic to reality:
the teleological part of her theory is only intended to explain misrepresentation (Godfrey-Smith
1996: 184-185; Shea 2013: 63). While Godfrey-Smith and Shea are skeptical about this, I will
here simply grant for the sake of argument that Millikan’s theory does contain a non-teleological
account of truth in terms of the relation that she calls “mapping” or “isomorphism”. My question
is what is precisely this relation. If it really is the Tractarian relation of picturing, as she
explicitly claims, then this part of her theory is undermined by the problems discussed in Chapter
2, Section 2.2. Millikan intends her theory to apply to all sorts of representations, from human
representations to the most basic cases of non-human representation. I will focus here on
sentences of natural human languages.

According to Millikan, sentences “correspond to states of affairs in accordance with
semantic-mapping functions” (Millikan 2005: 63). The notion of “function” here is the
mathematical one (Millikan 2005: 53), rather than the teleological notion that she deploys in
other parts of her theory. Semantic-mapping functions, she argues, are rules that determine the
truth-conditions of sentences:

The semantic-mapping function is given by rules according to which significant
transformations of the sentence that conserve its syntactic form yield different
Substitutions of component parts of a sentence of a specific syntactic form, she claims, change its truth-conditions in a systematic way. So, for example, ‘Fa’, can be transformed into ‘Fb’, ‘Ga’, ‘Gb’, etc., which have the same syntactic structure but represent different states of affairs. The difference regarding what portions of reality each of these sentences “maps” onto is determined exclusively by the different referential properties of their components. To the extent that mapping rules determine what states of affairs are represented by transformations of sentences of the same form—which is what Millikan suggests in the quoted passage—the “mapping” determined by these rules has nothing to do with the traditional notion of structural resemblance between sentences and facts. Of course the mapping rules that determine the truth conditions of ‘Rab’ and its transformations—‘Rba’, ‘Scd’, ‘Sdc’, etc.—will be different. It is clear that Millikan thinks that mapping functions relate each sentence to a state of affairs that would make it true, and that this relation depends on (1) the syntactic structure of the sentence and (2) the referential properties of its words. These two features are part of pictorial theories, but also of the contemporary non-pictorial account of correspondence that I discussed in Chapter 2 (although this theory does not explicitly posit fact or states of affairs, a variant of it may do so without relying on picturing relations). The questions are whether Millikan’s “mapping” rules are intended to relate sentences to states of affairs which have the very same structures as the sentences and whether this is actually required by her account of truth-conditions in terms of
Consider the following rule for the simplest form of subject-predicate sentence. \( SP: \) The sentence \( ^r \Phi x \downarrow \) is true if and only if the state of affairs in which the object referred to by \( ^r x \downarrow \) has the property referred to by \( ^r \Phi \downarrow \) obtains. The rule \( SP \) is compatible with Millikan’s claim that the relata of sentences are states of affairs and her professed Aristotelian realism about properties (Millikan 1984: Ch. 15 & 16). Also, \( SP \) covers all the substitution instances of \( ^r \Phi x \downarrow \) – i.e., ‘\( Fa \)’, ‘\( Fb \)’, ‘\( Ga \)’, etc.—assigning a specific state of affairs as the truth-maker for each instance. It yields in a systematic way the truth-conditions for of all the transformations of ‘\( Fa \)’. Assuming that these corresponding states of affairs have the same ontological structure—that of an object having a monadic property—\( SP \) pairs or correlates sentences that have a specific syntactic structure with states of affairs that have a specific ontological structure. \( SP \) delivers what Millikan calls “a correspondence by a given rule or function between form and some structure in the world” (Millikan 2005: 98). Everything that Millikan seems to require of a “semantic-mapping rule” is met by \( SP \). The crucial point is that \( SP \) does not require the combination of a name with a monadic predicate to structurally resemble the combination of an object with a monadic property. If \( SP \) is the kind of “semantic-mapping rule” that Millikan has in mind, then hers is a correlation-based rather than a resemblance-based form correspondence view. What is puzzling is Millikan’s talk of “mapping” as a form of “picturing” involving an “isomorphism” between representations and reality. A correlation account should only talk of “representing”.

Millikan’s favorite example of representation is the figure-eight waggle dance made by honey bees to communicate to each other the location of nectar—which she often compares to human sentences. The ethologist Karl von Frisch discovered that waggle dances communicate
the location of nectar and that they do this by representing its direction and distance from the hive: the angle of the waggle dance with respect to the vertical axis of the hive represents the angle from the line between the hive and the Sun where nectar is located, while the duration of the dance represents how far away from the hive the nectar is (von Frisch 1953: Ch. 11). According to Millikan, the bee dance “is a tiny map” that “maps the location of nectar by a certain rule of projection” (Millikan 2005: 83). “Variations in possible bee dances”, she points out, “correspond one-to-one to variations in possible locations of nectar” and “the principle involved”, she claims, “is mathematical isomorphism” (Millikan 2005: 97). The way she describes how the “bee language”—as von Frisch called it—represents the location of nectar has a distinctly Tractarian flavor. It is striking that bee dances do represent the location of nectar by means of mathematical rules of projection. A waggle dance $x$ represents the presence of nectar in direction $y$ at distance $z$ from the hive. There is a direct mathematical projection between the angle of $x$ and the direction $y$, and another one between the duration of $x$ and the distance $z$. Mathematical rules of projection determine the values of $y$ and $z$. Whether this makes the whole dance “isomorphic” to the location of nectar is not so clear and I will not discuss it. But what is this supposed to tell us about representations in general and about human sentences in particular?

Bee dances are Millikan’s favorite example because she thinks that they illustrate basic features of representations in general. Here is how she explains what “mapping functions” do in any system of representation: “Semantic mapping functions define isomorphisms between the set of possible signs in a certain sign domain and the set of their possible signifieds.” (Millikan 2004a: 49). This is meant to apply also to the linguistic and mental representations of human beings, although she warns that we should “expect the... mapping functions to be much more
abstract” in the case of human representations (Millikan 2005: 102). Furthermore, she insists that she uses the term ‘isomorphism’ in “the abstract mathematical sense” (Millikan 2004a: 84). In their pictorial period, Wittgenstein and Russell also intended true sentences to be isomorphic to facts only in an abstract sense. They intended only the underlying logical form of a true sentence to resemble the structure of what it represents, independently of its superficial structure which may be different. If this were the kind of abstract isomorphism that Millikan expects to obtain between true sentences and what they “map” onto, her theory would be undermined by the problem discovered by Russell: that a sentence ‘Rab’ would represent a fact containing two objects united by a dyadic relation, while the sentence contains three words united by a triadic logico-syntactic relation (see Chapter 2, Section 2.2). However, despite her talk of “picturing”, “mapping”, “rules of projection” and “isomorphisms”, she never claims explicitly that a true representation has the same structure as what it represents.

Recall Millikan’s claim that, in human languages, the mapping rules systematically determine for different sentences different states of affairs as truth-makers (Millikan 2005: 63-64). This is compatible with a correlation-based view of correspondence. Recall her claim that mapping rules determine a correspondence between the form of a true sentence and some structure in the world (Millikan 2005: 98). This claim also falls short of being genuinely pictorial. As I argued above, the rule $SP$ would determine a correspondence between a sentence of the form $^r \phi_x^\gamma$ and a state of affairs with a specific structure—that of an object having a monadic property—regardless of whether the sentence’s form matches that structure or not (which is good because it does not, as Russell showed). Now consider the following passage:

We can think of beliefs as... mental sentences... and think of... their truth makers...
as determined by some sort of Tarskian mapping... Certain substitutions of sentence parts correspond to substitutions in aspects of world affairs mapped...

That is, there is a functional isomorphism between the whole set of possible mental sentences... and the set of world affairs to which these sentences would correspond... (Millikan 2004b: 231-232)

This passage is very revealing of what kind of “mapping” Millikan expects to obtain between true sentences and reality. The rule $SP$ I proposed is precisely the “sort of Tarskian mapping” for simple predicative sentences that a realist about properties and structured states of affairs (or facts) like Millikan would need. Analogous “sort of Tarskian” rules would be required to cover other syntactic structures. What those rules would determine is a systematic one-to-one pairing of all possible sentences with all of their possible truth-makers, which is what Millikan in the passage calls—rather confusingly, in my opinion—a “functional isomorphism”. This pairing is clearly what a fact-based view of correspondence as correlation requires, which is less than what a genuinely pictorial theory requires. Incidentally, notice that the set of all instances of the equivalence schema also seems to provide a systematic pairing between sentences and what they represent (‘$p$’ is true $\leftrightarrow p$, ‘$q$’ is true $\leftrightarrow q$, ‘$r$’ is true $\leftrightarrow r$, etc.). But this “pairing” is not a relation. A correlation-based correspondence theory requires the systematic pairing to be determined by substantial relations between sentences and truth-makers.

All of this, however, leaves open the possibility that Millikan’s talk of relations of “picturing” between sentences and their truth-makers may additionally require actual structural resemblance. The fact that she does not explicitly require structural resemblance is not enough to sort this out, since her pictorial talk and her comparison between sentences and bee dances
suggests otherwise. However, Millikan has recently made a distinction between “projected correspondence” and “substitutional correspondence” (Millikan 2013b: 84-86) that helps to clarify this issue. This distinction is intended as a response to an objection by Shea. So I will provide some context before explaining the distinction and discussing how it settles the issue of whether Millikan requires true sentences to structurally resemble states of affairs.

In his objection to Millikan’s isomorphism requirement, Shea suggests an interesting hypothetical scenario. Suppose that the bee rules for representing distance were as follows: 1 waggle = 75 m, 2 waggles = 300 m, 3 waggles = 60 m, 4 waggles = 150 m, etc. (Shea 2013: 74). Of course, these are very different from the actual rules, discovered by von Frisch, relating bee dances to the distance of nectar. But in principle—while very unlikely—some species or subspecies of bee could have evolved to have dances which represent by Shea’s rules, provided their foraging range was limited, so that a finite list of rules could be used to represent all the required distances. I will call these creatures “bees*”. One of the features of Shea’s bee* rules is that they are list-like: there is a different rule for each distance represented. Another feature is that the list is arbitrary: there is no general principle involved, such as more waggles representing more distance. The only principle is that each number of waggles represents a particular distance. Bee* dances are not “isomorphic” in any minimally strict sense of the word to the distance of nectar they represent. Yet, as Shea points out, his arbitrary rules seem to determine what qualifies as a “functional isomorphism” in Millikan’s sense (Shea 2013: 75). As I indicated above, a systematic one-to-one pairing of representations with states of affairs seems indeed to be what she calls a “functional isomorphism”. Shea’s hypothetical bee* rules deliver such a pairing. The example echoes and supports Godfrey-Smith’s objection that Millikan’s notion of isomorphism
is too liberal—that “nearly any relation can constitute a «mapping» in Millikan’s theory” (Godfrey-Smith 1996: 185). Shea and Godfrey-Smith make these points in the context of discussing the relation between Millikan’s isomorphism requirement and her teleological theory of content. As far as Shea is concerned, Millikan’s teleosemantic theory has the resources to explain his hypothetical bee* rules in terms of evolutionary grounded substantive relations. He is not objecting to Millikan’s teleosemantics. What Shea intends the example to show is that the isomorphism requirement does not play a genuine role in the theory.

Millikan draws the distinction between projected and substitutional correspondence in her response to Shea. The language of actual bees, she suggests, is a clear example of mapping by means of projected correspondence. There are mathematical functions or projection rules relating features of different bee dances to different locations of nectar. She offers the following illustration of what the rule for representing distance may look like: “For example: add one waggle to the dance = add 1,000 yards to the distance of nectar. If this equation applies to every possible bee dance, it describes a projected correspondence between dances and distances” (Millikan 2013b: 84). While also imaginary, this rule does capture the basic principle involved in actual bee dances: a longer dance represents a longer distance of nectar from the hive. About Shea’s bee* rules Millikan says: “Shea suggests for bee dances a list giving arbitrary correspondences between numbers of waggles and distances. This would be a list of correlated substitution transformations...” (Millikan 2013b: 84). Her point is that there is no mathematical equation or projection rule relating the number of waggles to the distance of nectar. Instead, she

87 The list-like nature of bee* rules may raise concerns about whether they determine substantive correspondences, since deflationists offer precisely list-like characterizations which, they claim, involve no substantive relation. Presumably the reply to this concern is that the bee* dances are substantially related to locations of nectar by evolutionary history.
claims, there are substitutional rules that arbitrarily relate changes in the number of waggles to changes in the distance of nectar from the hive represented—e.g., replacing a 2 waggles dance with a 3 waggles dance transforms the distance represented from 300 m to 60 m (Millikan 2013b: 84-85). Even though they are arbitrary, she acknowledges, bee* rules do determine a systematic one-to-one pairing of representations and what they represent (Millikan 2013b: 85). Setting aside the representation of direction, her claim is that a bee dance is true when it has a projected correspondence to the distance of nectar, while a bee* dance is true when it has a substitutional correspondence to the distance of nectar.\textsuperscript{88}

Millikan further illustrates the distinction with another example where, she says, both kinds of correspondence are at play: a human-made map of a county where lines represent “the shape and placement of roads” and “different colors... represent different road surfaces, such as dirt, gravel, paved, and so forth” (Millikan 2013b: 86). This is basically a hybrid case where, if the map is accurate, the shape and placement of the lines has a projected correspondence to the actual shape and placement of the roads, while the colors of the lines have a substitutional correspondence to the actual surfaces of the roads.

A way of clarifying Millikan’s point is that what she calls a “projected correspondence” involves a relation of isomorphism between representation and reality, while what she calls a “substitutional correspondence” involves merely a correlation determined by some substantial relation other than isomorphism. Millikan does not make the point explicitly in these terms. Perhaps she still wants to regard substitutional correspondences as involving some form of “abstract isomorphism”. While this is a terminological issue, I see no point in applying the term

\textsuperscript{88} Millikan now claims that the teleological part of her theory is what grounds any such correspondences on real relations (Millikan 2013b: 82-83).
‘isomorphism’ to non-projective correspondences. This would be misleading, since it would obscure the standard distinction made in the literature on the nature of truth between correspondence as congruence, structural resemblance or isomorphism, on one hand, and correspondence as mere correlation, on the other. Once a theory admits both non-projective and projective correspondences, it is better characterized as a correlation theory, since what it claims to be the common feature of all true representations is that they are paired to their truth makers by one or another substantive relation—projective relations being merely one such kind of relation that some representational systems (e.g., bee dances) may rely on.

Now we can address Millikan’s position regarding sentences in human languages. Drawing on her distinction between projective and substitutional correspondence, she says that “almost all of the correspondences found in natural language are substitutional” (Millikan 2013b: 84) and claims that “this is the main difference between simpler representations and the representations of human language” (Millikan 2013b: 86). I am skeptical about the second claim. But let me start by commenting on the first one. Millikan has always argued that what sets apart sentences from other representations is their syntactic structure: they have subject-predicate form and can be negated (Millikan 1984: 308-309). She has pointed out, for example, that “bee dances are not sentences, for they have no subject terms, and a rabbit’s danger thumps are not sentences, for they never predicate of a time and place anything other than danger... [while] the alarm calls of vervet monkeys are not sentences because... [they] are never negated” (Millikan 1993: 118). Also, she has always characterized the “mapping” rules of sentences as mainly substitutional, as shown by some passages I have already quoted. Here is an earlier example: “In the case of sentences that map, the significant transformations... are mostly substitution transformations...”
(Millikan 1984: 108). However, only now she has made the explicit claim that substitutional rules are not projective rules. This makes it clear that she does not regard now the kind of “substitutional” correspondence of true sentences to be like the “projective” correspondence of accurate bee dances. In the clearer terms I prefer to use, her conception of sentential truth is one of correspondence as correlation, rather than one of correspondence as congruence, structural resemblance or isomorphism. I argued before that Millikan’s theory does not really require true sentences to structurally resemble states of affairs. Whatever she intended in her previous work—which is not entirely clear—she now clearly does not think that true sentences structurally resemble what they represent.

I close this appendix with couple of critical remarks. The first one is substantive and the second one terminological. My first remark is against Millikan’s claim that the distinction between substitutional and projective truth-conditions is what sets apart human sentences from simpler representations. (I think that her syntax-based distinction is more appropriate, although we may find syntax also in non-human languages of thought: whether syntax is uniquely human is an empirical question.) Notice that some very sophisticated representations (e.g., computer generated models) may rely at least partly on projective rules. More importantly, many simple non-human representations seem to have “substitutional” truth-conditions. Vervet monkey alarm calls are an obvious example: they are merely correlated to the presence of the appropriate predators when accurate. Indeed, plenty of extremely simple “detection” systems represent reality without relying on any projective rules or mathematical isomorphisms. The hair detectors used by the Venus flytrap to detect that an insect has landed are a clear example. (By the way, the triggers only represent—when successful—that an insect has landed, so there are neither
projective nor substitutional “transformations” involved. The distinction does not apply across the board.) It is likely that the perceptual representations of more sophisticated creatures do rely on some forms of isomorphism (Millikan 2005: 102). But correspondence as correlation seems to be ubiquitous and not merely a sophisticated human phenomenon.

My second point is that once correspondence by means of mathematical isomorphism has been discarded as a necessary condition for representational truth, there is no motivation for regarding true representations in general as either picturing or mapping what they represent. It is obvious that any further talk of picturing would be merely metaphorical. Perhaps bee dances and many other representations literally are maps of what they represent. But consider simple human sentences like ‘Socrates is wise’ or ‘Snow is white’. Saying that these sentences map something seems clearly metaphoric. What could that possibly mean other than that they represent something? Talk of mapping does not convey anything that cannot be as easily conveyed by explicit talk of representing.

In Chapter 2, Section 2.2, I argued that the pictorial account of sentential truth as structural resemblance to facts is hopeless. In this Appendix I have argued that the “pictorial” account promoted by Millikan, despite her claims, does not require genuinely pictorial relations. If she originally intended correspondence relations in general to involve structural resemblances, that feature of her theory should be replaced by a correlation-based account. It seems that Millikan herself has now reached this conclusion, at least for the case of sentential truth. The only viable alternative for a correspondence theory of sentential truth is to offer a correlation account based on substantial relations between sentences and reality.
Appendix 2:

More on Attitude Ascriptions

In Chapter 1, I argued against the view that attitude ascriptions of the form ‘S believes that p’ or ‘S said that p’ assert relations between people and abstract objects called “propositions”. I defended a non-propositionalist alternative, proposed by Devitt (1996), according to which the ‘that’-clauses in attitude ascriptions function as *indefinite* singular terms. I showed how this account has crucial advantages over the propositionalist analysis. In this appendix, I will discuss yet another analysis of attitude ascriptions that constitutes an important non-propositionalist alternative to the one I defended in Chapter 1: the multiple-relation analysis offered by Friederike Moltmann (2003 & 2013). Moltmann’s analysis is based on Russell’s multiple-relation theory of beliefs. Russell himself abandoned this view due to a serious problem he encountered while trying to develop it—and the view was not taken seriously afterwards. Moltmann, however, has resurrected the view an attempted to articulate it in a way that overcomes the difficulties faced by Russell’s version. In what follows, I will first present Russell’s original proposal and the problems it faces. Afterwards, I will present Moltmann’s contemporary version and argue that it also faces serious difficulties. Finally, I will argue that the analysis I proposed following Devitt has crucial advantages that favor it over any multiple-relation analysis of belief ascriptions.

To understand Russell’s multiple-relations theory, it will be convenient to briefly review the view he was trying to avoid: his earlier view that the primary truth-bearers are what today are called “Russellian propositions”. (I discussed this in Chapter 1. Here I provide a brief summary.) Let me use a simple example: the proposition that $Rab$—i.e., that the object $a$ has the relation $R$
with the object $b$. The early Russell (1903) held that the constituents of propositions were entities in the world. The constituents of the proposition that $\text{Rab}$ would be the objects $a$ and $b$, as well as the relation $R$. But what glues these parts into a unity? Russell suggested that in the proposition that $\text{Rab}$, the fact that $R$ actually relates $a$ and $b$ is what provides the unity of the proposition (Russell 1903: 48-53). As I pointed out in Chapter 1, on this view true propositions collapse into the facts that make them true. Russell embraced this, proposing that a proposition is true if and only if it is identical to a fact (Russell 1904). But a serious problem is that on Russell’s view there cannot be false propositions (King 2007: 23). If it is not a fact that $\text{Rab}$—if $R$ does not relate $a$ and $b$—there is no unity that we can call the “false proposition” that $\text{Rab}$.

Eventually Russell attempted to avoid the problem of falsehood by rejecting propositions and replacing them with beliefs as the primary truth-bearers. Treating a belief as a relation to a “Russellian” proposition would defeat the purpose. Russell proposed instead that a belief is complex unity involving multiple relations between entities in the world and the believer: that $S$’s act of believing that $\text{Rab}$ contains $R$, $a$ and $b$ bound to $S$ by the believing relation (Russell 1912: 124-29). Notice that there was an important continuity with Russell’s earlier account: the constituents of his truth-bearers still were mind- and language-independent entities, rather than representations standing for them. But on Russell’s new account the unity of the truth-bearer was provided by the believing relation: “The relation $[R... \text{when } S \text{ believes that } \text{Rab}]...$ occurs in the act of believing... [as] one of the objects—it is a brick in the structure, not the cement. The cement is the relation ‘believing’.” (Russell 1912: 128). Russell’s analysis of the structure of $S$’s act of believing that $\text{Rab}$ can be logically symbolized by ‘$B_sRab$’—where ‘$B$’ is a polyadic predicate standing for the believing relation but ‘$r$’ is a name for a relation rather than a dyadic
predicate. Similarly, the analysis of $S$’s act of believing that $Fa$ can be symbolized by ‘$BsFa$’—where ‘$F$’ is a name for a property instead of a predicate. The constituents united by the believing relation are all treated as objects.

The multiple relation account allowed Russell to make some room for falsehood and led him to define truth as correspondence. On this account, whether $BSRab$ is independent of whether it is a fact that $Rab$—the believing relation can bind $R$, $a$ and $b$ one by one with $S$ even if $R$ does not bind $a$ and $b$: “it is easier to account for falsehood if we take judgment to be a relation in which the mind and the various objects concerned all occur severally” (Russell 1912: 125). This independence of truth-bearers from facts is what originally led Russell to regard truth as a correspondence rather than an identity relation. He proposed that $S$’s belief that $Rab$ is true if and only if there is a corresponding fact where $R$ binds $a$ and $b$ in the “same order” as $R$, $a$ and $b$ are bound to $S$ by the believing relation (Russell 1912: 128-29). As Kirkham points out, this is “the (original modern) correspondence-as-congruence theory of truth” (Kirkham 1992: 132).89

Russell’s multiple relation theory was an ingenious attempt to avoid propositions, but it is undermined by two insurmountable problems which show that the theory ascribes the wrong constituents and the wrong structures to truth-bearers. The solution to these problems requires a shift to structured representations as truth-bearers.

One problem faced by Russell’s theory is that it does not truly make room for all kinds of falsehood: it fails to achieve its main goal. Suppose that $S$ falsely believes that $Rab$, not merely

---

89 The account relies on a supposed resemblance or congruence between the structures facts and true beliefs. But it also relies on an alleged identity of the constituents of facts and true beliefs. While this is the first contemporary attempt to explain truth as resemblance-to-facts, the truly influential one is the second attempt by Wittgenstein and Russell himself, which treats representations as truth-bearers and relies on referential rather than identity relations between their constituents and reality.
because $R$ does not relate $a$ and $b$, but because $b$ does not exist. As Kirkham points out, on Russell’s analysis $S$ cannot have such a false belief: one of the constituents is missing, since the theory requires $R$, $a$ and $b$—rather than ideas or concepts purporting to stand for them—to be the constituents of the belief (Kirkham 1992: 122-123). I will call this the “missing constituents” problem. The upshot of this problem is that the constituents of a truth-bearer must be *symbols* or *representations* that purport—but may fail—to refer to entities in the world. In fact, the early Russell’s account of propositions is also undermined by the problem: if $a$ does not exist, there cannot be a false proposition that $Fa$, not because it lacks unity, but because it lacks a constituent. The early Russell had a reply: ‘$a$’ is not a genuine name, but an abbreviation for a uniquely identifying description (Russell: 1905), so the proposition *expressed* by ‘$Fa$’ does not contain $a$ to begin with: there is no missing constituent. An analogous attempt may be made to rescue his multiple relation theory: the belief *expressed* by ‘$Fa$’ does not contain $a$, so it is not missing a constituent (see Kirkham 1992: 123). But Saul Kripke has shown that Russell’s treatment of *ordinary* names as disguised descriptions utterly fails (Kripke 1980). Additionally, Russell acknowledged that there are *genuine* names. The missing constituents problem would anyway arise for Russellian propositions or beliefs expressed by sentences containing empty genuine names. While Russell assumed that genuine names cannot be empty, Kripke has also debunked this assumption (Kripke 2011).

The missing constituents problem may be avoided by a modification of the multiple relation theory that replaces the relevant entities with concepts purporting to stand for them. The truth of a belief would then depend on whether there is a corresponding fact where the *referents* of the concepts are bound together in the same order as the *concepts* are bound to the believer by
the believing relation—which would make room for beliefs that are false due to reference failure. If we use underlining to symbolize concepts, so that the concepts purporting to stand for $R$, $a$ and $b$ are $R$, $a$ and $b$, we can symbolize the resulting analysis of $S$’s act of believing that $Rab$ by $‘B_{s}Rab’$—where ‘$R$’ is a name for a predicational concept, rather than a dyadic predication. This is still a multiple relation account: the concepts are related one by one to the believer by the believing relation. If the missing constituents problem were the only problem faced by Russell’s multiple relation theory, this variant may work. In any case, Russell did not realize that his theory was undermined by this problem (perhaps he thought that he could deploy his theory of disguised descriptions to avoid it). What led him to abandon the multiple relation theory was another problem, which happens to affect any variant.

The other problem faced by the multiple relation theory was discovered by Wittgenstein. In a letter from June 1913, Wittgenstein pointed out to Russell that ‘$Rab \lor \sim Rab$’ must “follow directly” from the correct analysis of ‘$S$ believes that $Rab$’ and that this “condition is not fulfilled” by his theory (Wittgenstein 2008: 40). Notice that ‘$Rab \lor \sim Rab$’ is a tautology and that any arbitrary tautology can be validly deduced from any statement. So what is the problem? Wittgenstein was hinting that there must be a non-arbitrary connection between $S$’s belief that $Rab$ and specifically ‘$Rab \lor \sim Rab$’ which—unlike other tautologies—expresses that this particular belief is either true or false. On his Notes on Logic from September 1913, Wittgenstein elaborated on why the multiple relation theory fails to establish this connection:

When we say A judges that, etc., then we have to mention a whole proposition which A judges. It will not do... to mention only its constituents... [A] right theory of judgment must make it impossible for me to judge that “this table penholders
the book”... (Wittgenstein 1969: 96)

In the ordinary ascription ‘S believes that Rab’, S’s belief is specified by ‘Rab’, which by itself constitutes meaningful unit that can be true or false—‘Rab ∨ ~Rab’—because the syntactic arrangement of ‘R’, ‘a’ and ‘b’ forms a sentence. Such a sentence is the “whole proposition” that, according to Wittgenstein, must be mentioned to specify S’s belief. (I discuss below Wittgenstein’s treatment of sentences as propositions). The belief ascriptions made by Russell’s theory were very different, since they replaced ‘S believes that Rab’ with ‘BsRab’, which does not contain ‘Rab’. This was an intended feature of Russell’s analysis. But Wittgenstein objected that in ‘BsRab’, the string ‘rab’ is not by itself a meaningful unit that can be true of false, so it cannot specify S’s belief. The concatenation of ‘r’, ‘a’ and ‘b’ does not form a sentence because ‘r’ does not play the role of a dyadic predicate: ‘rab’ is merely a string of names analogous to the nonsensical string of ‘table’, ‘penholder’ and ‘book’ mentioned by Wittgenstein. As Russell explained the problem some years later, he was “putting the subordinate verb on a level with its terms as an object term in the belief” (Russell 1918: 59). Russell’s account of S’s act of believing that Rab did not have the right form: it did not relate S to a structured unity that can be true or false. I will call this the “structure problem” of the multiple relation theory.90

Notice that the structure problem also affects the concept-based variant of the multiple relation theory that may overcome the missing constituents problem. In ‘BsRab’ the string ‘rab’ is merely a concatenation of names for concepts, rather than a meaningful unit that can be true or false. Nicholas Griffin (1985) argues that Wittgenstein’s main concern was that the multiple relation theory did not provide the necessary type restrictions on the constituents of the belief: it did not treat R as a relation in S’s belief that Rab. Peter Hanks (2007) argues that Russell could have provided the required type restrictions, but Wittgenstein’s main concern was the unity of the truth-bearer: even if R were specified to be a relation, it would not relate a and b, since R, a and b would still be severally related to the believer. Of course, if R is not treated as a relation to begin with, its unity with a and b is also broken. On either interpretation of Wittgenstein’s objection, Russell’s truth-bearer does not have the right structure.
false such as the mental sentence symbolized by ‘Rab’. Regarding beliefs as \textit{structured mental representations} containing concepts or ideas purporting to refer to entities in the world avoids both the structure and the missing constituents problems.

Russell unsuccessfully tried to amend his theory by adding \textit{a form} as yet another \textit{object} of the multiple relation—instead of ‘\textit{BsRab}’, the ascription would be ‘\textit{BsRab\delta}’, were ‘\textit{\delta}’ names the form \(\Phi_{xy}\)—but it was unclear how this separate form could structure the other constituents (Griffin 1985: 242; Hanks 2007: 127). Russell eventually abandoned the theory and accepted Wittgenstein’s proposal that truth-bearers are linguistic and mental representations. The shift to representations was exclusively motivated by the structure problem. The missing constituents problem played no role. Even in Russell (1918) and Wittgenstein (1922) empty genuine names are wrongly assumed to be an impossibility (Kripke 2011: 60-61).

After Russell himself abandoned it, the multiple relation theory was relegated during the remainder of the twentieth century to a “historical curiosity, rather than a serious philosophical position” (Griffin 1985: 213). But Michael Jubien (2001) and Friederike Moltmann (2003) have recently resurrected it in an attempt to avoid treating beliefs and other attitudes as relations to mind- and language-independent propositions. Peter Hanks argues that these attempts face the very same problem that led Russell to abandon the theory (Hanks 2009: 471-72 & 484n). Moltmann (2013), however, maintains that her analysis of attitude ascriptions overcomes the problem. Her version of the multiple relation theory is the most developed one and provides an alternative to the non-propositionalist analysis of attitude ascriptions I proposed in Chapter 1. But I will argue that it is unclear whether her ingenious proposal succeeds and that my proposal avoids altogether the problems faced by any multiple relational analysis of attitude ascriptions.
Moltmann (2013) treats attitude verbs as multigrade predicates, following a proposal by Alex Oliver and Timothy Smiley (2004). A multigrade predicate is alleged to be one that can take a variable number of arguments. Consider (1) ‘Tom cooked dinner’ and (2) ‘Tom, Dick and Harry cooked dinner’. The logical form of (1) is ‘Ct’, but Oliver and Smiley contend that ‘cooked dinner’ is multigrade, so the form (2) is ‘Ctdh’ rather than ‘Ct & Cd & Ch’ (Oliver and Smiley 2004: 609-612). Oliver and Smiley also draw a distinction between argument places and positions. Consider (3) ‘Adam fought with Yuri’ and (4) ‘Adam and Bill fought with Yuri and Zero’. Each of the two places of ‘x fought with y’, they propose, can take a variable number of internal positions; so the form of (3) is ‘Fa;y’ and the form of (4) is ‘Fab;yz’—where the non-standard ‘;’ is used to separate places (Oliver and Smiley 2004: 615-618).

Moltmann suggests that ‘believes’ and other attitude verbs are multigrade predicates with a first place reserved for the agent and a second place with a variable number of positions reserved for the constituents of the “attitudinal object” (which she claims is not a proposition); so that the logical form of ‘John thinks that Mary likes Bill’ is: ‘THINK(John; LIKE, Mary, Bill)’ (Moltmann 2013: 132-149). Like Russell, she regards the constituents as severally related arguments of the attitude verb. Unlike Russell, she treats the attitude verb as dyadic: ‘LIKE, Mary, Bill’ occupies a single place, within which ‘LIKE’, ‘Mary’ and ‘Bill’ occupy different positions. The attitude verb, she proposes, establishes relations of “intentional predication”: the attitudes are “ways of predicating a property of its arguments” (Moltmann 2013: 147). Her suggestion is that the attitude verb assigns “different roles” to the “different positions” within its second place: “one distinguished argument position for a property, meant to be predicated of the other arguments, as well as further argument positions matching the argument positions of the
property” (Moltmann 2013: 149). What ‘S believes that Rab’ states, she argues, is that S predicates the property R of a and b (Moltmann 2013: 146). This analysis, she claims, overcomes the problem that undermined Russell’s account: “the fact that the constituents are arguments of a multigrade attitudinal relation will ensure that the sequence forms the content of an attitudinal object that... has the right truth conditions.” (Moltmann 2013: 153).

Does Moltmann’s theory overcome the problems of Russell’s theory? Let us start with the missing constituents problem—which she does not discuss. Moltmann says: “An agent is successful in predicating an n-place property of n objects just in case the property holds of the objects” (Moltmann 2013: 146). If it is not the case that Rab, then S is not successful in predicating R of a and b. This makes room for some falsehoods. But Moltmann’s analysis, like Russell’s, relates S to R, a and b themselves. If b does not exist, then it seems that S is not only unsuccessful in predicating R of a and b, but the predication cannot take place. In fact, if ‘b’ is an empty term, ‘S believes that Rab’ seems to be a false ascription under her analysis instead of the ascription of a false belief. Moltmann makes a tentative suggestion that may provide a solution to this problem: “[the] constituents may be concepts, occupying a position in the multigrade place of the attitude verb specifically marked for such concepts” (Moltmann 2013: 159). Moltmann here is not committing herself to this suggestion, and she is focusing only on the predicate embedded in the ‘that’-clause (which is why she mentions only a single position that may be occupied by concepts). But by adopting this suggestion and extending it to cover the arguments of the embedded predicate, her theory may overcome the missing constituents problem. This would be a concept-based modification of her analysis.

Moltmann’s analysis also faces a dilemma regarding the structure of the ascription. The
quasi-unity provided by treating the attitude verb as dyadic is not enough to overcome the structure problem: what unifies the different positions in its second place into a structure that can be true or false? Moltmann claims that the attitude verb assigns “different roles” to the different positions. This could either mean that in ‘S believes that Rab’ the ascription contains a name ‘R’ referring to a property R—which S is “meant” to predicate of a and b—or that the ascription itself contains a relational predicate ‘R’. The dilemma is that both options are problematic: the first one is undermined by the structure problem, while the second one seems to collapse into propositionalism. According to the first option, the form of the ascription is ‘Bs;Rab’. The string ‘Rab’, just like in Russell’s version, is a concatenation of names. Even if one of these names has the specially assigned role of referring to a relational property, this does not turn the string into a unity that can be true or false. Clearly this is not what Moltmann has in mind: recall her analysis of ‘John thinks that Mary likes Bill’ as ‘THINK(John; LIKE, Mary, Bill)’. What she has in mind is the second option, according to which the form of the ascription is ‘Bs;Rab’. In the string ‘Rab’, the predicate ‘R’ does appear to take ‘a’ and ‘b’ as its arguments: we do seem to have a structure with “the right truth conditions” (setting aside the missing constituents problem: can S have this belief if b does not exist?). But this analysis seems to collapse into the very kind of propositionalism she wants to avoid. In the traditional propositionalist view, the ascription ‘S

91 Oliver and Smiley suggest this option. They argue that this solves the problem of the multiple relation theory using “predicates with variably many arguments”, but they do not claim that it solves the problem—which they do notice—of “the conversion of verbs into nouns (‘loves’ into ‘loving’) which the multiple relation theory demands” (Oliver and Smiley 2004: 628). Moltmann departs from Oliver and Smiley on this matter.

92 This is not a well-formed formula of first-order logic, even when extended to include multigrade predicates. But suppose that ‘B’ is treated instead as a multigrade higher-order predicate that takes both names and predicates as arguments. In this case, the ontological commitments of the ascription may seem as problematic as the commitment to propositions that Moltmann wants to avoid. However, Moltmann (2013) interprets ordinary language expressions—including second-order predications—as having less ontological commitments than usually supposed. Roughly, she interprets expressions usually understood as referring to a property qua universal or abstract object as merely plurally referring to the various particular instances of the property (Moltmann 2013: Ch. 1-3). I will grant that Moltmann can account for ‘Bs:Rab’ being well-formed and not having problematic ontological commitments related to the seemingly higher-order predication it contains.
believes that *Rab*‘ is analyzed as having the form as ‘*Bs<~Rab>*’—where ‘~Rab>’ functions as a *name* for the proposition expressed by the sentence ‘*Rab*’. How different is ‘*Bs;Rab>* from ‘*Bs<~Rab>>’? Instead of angled brackets, Moltmann uses multigrade positions. Nonetheless, her analysis seems to plug the same unity—by other means—into the second argument place of the attitude verb. She actually acknowledges that in her analysis “the structure of the multigrade position matches a structured proposition”, but she insists that there is a difference: that there is no propositional object in her analysis, but merely “an ordered plurality of propositional constituents” (Moltmann 2013: 150). Moltmann wants the constituents to be separated enough so that there is no unity left that may be regarded as a “propositional object”, but she also wants them to be united enough so that the “attitudinal object” they compose can be true or false. It is unclear whether she succeeds balancing these two conflicting *desiderata*.

Things get even less clear when we consider ascriptions containing logically complex embedded sentences. Moltmann tentatively suggests that ‘John believes that either Mary or Bill will win’ may have the form ‘believe(John; OR, f[WIN, Mary], f[WIN, Bill])’—where the connective ‘OR’ works as a multigrade *predicate* “taking attitudinal objects of entertaining as arguments in any of their places” (Moltmann 2013: 152). This suggestion seems outright implausible. Are *quantifiers* and *negations* embedded within ‘that’-clauses also multigrade *predicates*? Consider ‘John believes that God does not exist’. Its multiple relation analysis will have to be some variant of ‘BELIEVE(John; NOT, EXIST, God)’. Placing logical operators and quantifiers in predicate argument positions is not only logically problematic, but ontologically opens Pandora’s box. Of course, the Russellian view of propositions has difficulties handling complex truth-bearers as well: what are the constituents of, and what provides the structure for,
propositions like <Mary or Bill will win> and <God does not exist>? Moltmann acknowledges that her theory “shares a number of issues concerning complex sentences with structured propositions approaches” (Moltmann 2013: 152). What she has in mind is mainly Russellian approaches. But these shared issues are serious enough to undermine both kinds of approach.

The analysis I propose in Chapter 1, drawing on Devitt’s (1996) suggestion that the ‘that’-clauses in attitude ascriptions function as indefinite singular terms, completely avoids commitment to mind- and language-independent propositions, while also avoiding the problems faced by multiple relation analyses. According to this analysis, ‘S believes that p’ states that S has a belief and specifies that the content of this belief is the same as that of the embedded sentence ‘p’. The form of the attribution in the analysis is ‘∃x(Bxs & Sx<p>’): some belief of S has the same content as ‘p’. In Chapter 1, I showed that this analysis has crucial advantages over the propositionalist analysis. First, it avoids the Substitution Problem that, as Moltmann (2003) shows, undermines the propositionalist analysis. Second, it does not treat the attributed content as a mind- and language-independent entity, which could not possibly play a causal role in the explanation of people’s behavior. Additionally, I showed that the analysis is able to account for the validity of inferences like ‘S believes that p, so S believes something’ without quantifying over propositional objects.

93 The issue is that the clause ‘that p’ cannot be replaced with ‘the proposition that p’ for all attitude verbs without changing the meaning of the ascription. ‘Mary believes that it will rain’ can be rephrased as ‘Mary believes the proposition that it will rain’, but ‘Mary fears that it will rain’ cannot be rephrased as ‘Mary fears the proposition that it will rain’. Strictly speaking, the substitution problem only undermines the traditional version of the propositional analysis, where ‘S believes that p’ is analyzed as having the form ‘Bs<→p>’. As I pointed out in Chapter 1, a propositionalist can go along with Devitt’s proposal to treat the ‘that’-clause as an indefinite singular term, but give a propositionalist twist to the analysis. On this sophisticated propositionalism, ‘S believes that p’ is analyzed as having the form ‘∃x(Bxs & Cx<p>’): Some belief of S has the proposition that p as its content. As I showed in Chapter 1, this analysis avoids the substitution problem. However, the resulting view is undermined by the main problem with propositionalism: the ascribed contents cannot play causal roles in the explanation of behavior.
Now I want to point out that this analysis also does better than multiple relation analyses. Consider (1) ‘S believes that \(Fa\)’, (2) ‘S believes that \(Rab\)’ and (3) ‘S believes that \(Fa\ or \ Gb\)’. On the analysis I propose, the forms of these attributions are simply (1’) ‘\(\exists x(Bxs \& Sx\ ‘Fa’\)’, (2’) ‘\(\exists x(Bxs \& Sx\ ‘Rab’\)’ and (3’) ‘\(\exists x(Bxs \& Sx\ ‘Fa\ or \ Gb’\)’—where names for the sentences embedded in the ‘that’-clauses (formed by enclosing them within quotation marks) are placed in the second argument place of the relational predicate ‘\(Sxy\)’ (‘\(x\) has the same content as \(y\)’). Suppose that ‘\(Gb\)’ is true—so ‘\(Fa\ or \ Gb\)’ is true—but ‘\(Fa\)’ and ‘\(Rab\)’ are false because \(a\) does not exist. Then the beliefs attributed to \(S\) in (1’) and (2’) are false, while the belief attributed in (3’) is true. The analysis avoids the **missing constituents problem**: if the embedded sentence is false due to reference failure, the ascribed belief will be also false. Since any two representations (whether mental or linguistic) have the same truth-value if they have the same content, the analysis entails that the belief has the same truth-value as the sentence embedded in the ascription’s ‘that’-clause. Consequently, the analysis also avoids the **structure problem**: by mentioning the complete embedded sentences, it specifies the right truth-conditions for the ascribed beliefs. (Notice that this is precisely what Wittgenstein recommended in his criticism of Russell’s theory, although he did not articulate an analysis of ascriptions.) For example, the analysis of (1) as (1’) entails that \(S\)’s belief is true if and only if \(Fa\). The full inference is the following:

\[
\begin{align*}
(P1) & \text{ Some belief of } S \text{ has the same content as ‘} Fa \text{’.} \\
(P2) & \text{ ‘} Fa \text{’ is true if and only if } Fa. \\
(P3) & \text{ Any two representations have the same truth-value if they have the same content.} \\
(C) & \text{ Therefore, some belief of } S \text{ is true if and only if } Fa.
\end{align*}
\]

Here (P1) is the analysis’ paraphrase of ‘\(S\) believes that \(Fa\)’, (P2) is an uncontroversial instance of the equivalence schema and (P3) is an equally uncontroversial principle. The conclusion (C)
shows that the analysis ascribes a belief with the right truth-conditions. The logical form of this inference is as follows:

\[
\exists x(Bxs \land Sx'Fa')
\]

\[
T 'Fa' \leftrightarrow Fa
\]

\[
\forall x \forall y[Sxy \rightarrow (Tx \leftrightarrow Ty)]
\]

\[
\therefore \exists x[Bxs \land (Tx \leftrightarrow Fa)]
\]

Where ‘\(Tx\) = ‘x is true’. (Recall that ‘\(Bxs\)’ = ‘x is a belief of S’ and ‘\(Sxy\)’ = ‘x has the same content as y’). This is a valid argument form and it remains valid when replacing the all mentioned and used instances of ‘\(Fa\)’ with instances of any other sentence. This shows that any belief ascription under the proposed analysis has the right truth-conditions.

Finally, the analysis I propose can handle attitude ascriptions with logically complex embedded sentences as easily as it can handle simpler ones, just with the resources of standard first-order logic. This is already shown by the analysis of (3) as (3’). But consider (4) ‘S believes that \(\neg \exists y(Fy)\)’. The analysis of (4) as (4’) ‘\(\exists x(Bxs \land Sx'\neg \exists y(Fy))'\)’ does not use ‘\(\neg\)’ or ‘\(\exists y\)’ in a predicate argument position: they are merely parts of the name for the sentence that specifies the content of S’s belief. (3’) and (4’) are well-formed and ontologically unproblematic. They are as plausible and harmless as the statement that the Spanish sentence ‘Dios no existe’ has the same meaning as the English sentence ‘God does not exist’.

Even if a sophisticated multiple relation analysis like Moltmann’s were able to avoid the problems of Russell’s version without collapsing into propositionalism—which is unclear—its treatment of logically complex ‘that’-clauses would nevertheless be implausible. Consequently, I
doubt that the multiple relation theory can be successfully resurrected. Russell did well to abandon it following Wittgenstein’s advice.

Neither the early Russell’s account of propositions nor his subsequent account of beliefs solved the problem of the *unity or structure* of the truth-bearer. On his *Notes on Logic* from 1913, Wittgenstein proposed a simple solution: “Propositions... are symbols” (Wittgenstein 1969: 98). The *sentential symbol* ‘\(Rab\)’, for example, has the symbols ‘\(Rxy\)’, ‘\(a\)’ and ‘\(b\)’ as constituents (Wittgenstein 1969: 98). There is no problem accounting for the unity of the *sentence*, which is provided by its *syntactic structure*. As Wittgenstein wrote in his *Notebooks* on 1914: “Does the subject-predicate form exist? Does the relational form exist? ...every*thing* that needs to be shewn is shewn by the existence of subject-predicate SENTENCES” (Wittgenstein 1969: 2-3). Of course, the structure of sentential symbols was never a problem for Russell. In fact, he used symbolic logic to uncover the hidden structures of sentences that traditional logic mistakenly analyzed as having simple subject-predicate forms—such as ‘The present King of France is bald’ which, Russell (1905) showed, has the complex form ‘\(\exists x ([Kx \& \forall y (Ky \rightarrow x=y)] \& Bx)\)’: i.e., ‘there is unique thing which is King of France and this thing is bald’). But Russell was convinced that the primary truth-bearers themselves contained *no* symbols—whether in his early account of propositions or in his multiple relation theory of beliefs—and regarded the symbols used in logic as a “theoretically irrelevant convenience”. This conviction made syntactic structures unavailable as candidates for the structure of truth-bearers. Wittgenstein’s suggestion that “propositions” *are* symbols places not only symbols, but also syntactic structures, in the truth-bearers themselves.

By 1918, Russell finally accepted sentences as truth-bearers: a “proposition is just a... complex symbol... that... has parts which are also symbols: ...a sentence containing several
words... is therefore a complex symbol” (Russell 1918: 10). By 1919, he extends the view to cover mental representations, whether they are made out of mental words and mental images (Russell 1919: 29). Russell’s truth-bearers are finally concrete linguistic and mental representations. Similarly, Wittgenstein (1922) uses the word ‘proposition’, as Horwich points out, to refer to “a sentence with its meaning, and not, as is more common these days, to refer merely to the meaning itself that a sentence might have” (Horwich 2012: 76n). Wittgenstein’s “propositions” are structured combinations of meaningful symbols or representations. Like Russell, he intends the view to cover mental representations: “The applied, thought, propositional sign is the thought.” (Wittgenstein 1922: 3.5). While Russell (1919) and Wittgenstein (1922) do not offer an account of attitude ascriptions, notice that neither the analysis of ascriptions as stating single relations between agents and mind- and language-independent propositions, nor Russell’s multiple relation analysis, is compatible with their view that beliefs are concrete structured representations that are truth-bearers in their own right—i.e., primary truth-bearers. But the analysis of ascriptions as proposed by Devitt, which I have defended in Chapter 1 and in this appendix, does offer a compatible alternative.
References


Devitt, Michael (1997). On Determining Reference. *Sprache und Denken / Language and


Harman, Gilbert (1987). (Nonsolipsistic) conceptual role semantics. *New Directions in*


Cambridge, MA.


