Communities of Judgment

Towards a Teleosemantic Theory of Moral Thought and Discourse

Karl Bergman

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Geijersalen, Thunbergsvägen 3H, Uppsala, Friday, 11 October 2019 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Adjunct Professor Marc Artiga (Universitat de València).

Abstract
Bergman, K. 2019. Communities of Judgment. Towards a Teleosemantic Theory of Moral Thought and Discourse. 207 pp. Uppsala: Department of Philosophy.
ISBN 978-91-506-2786-2.

This thesis offers a teleosemantic account of moral discourse and judgment. It develops a number of views about the function and content of moral judgments and the nature of moral discourse based on Ruth Millikan's theory of intentional content and the functions of intentional attitudes.

Non-cognitivists in meta-ethics have argued that moral judgments are more akin to desires and other motivational attitudes than to descriptive beliefs. I argue that teleosemantics allows us to assign descriptive content to motivational attitudes and hence that even if the non-cognitivist is correct, moral judgments can be said to describe the world. Moreover, given further teleosemantic assumptions, this conclusion has consequences that are both surprising and interesting. First of all, while moral judgments have descriptive content, moral statements do not. The purpose of moral discourse is not to convey beliefs that are true *simpliciter*, but to convey attitudes that are descriptively correct when tokened by the addressee. Consequently, moral discourse requires speakers to adapt to hearers in order to secure their assent and bring them into "community of judgment" with themselves.

Secondly, the descriptive content of a motivational attitude is partly a matter of the subject's own preferences and circumstances. In particular, the descriptive correctness of a moral judgment is partly a function of the degree to which it is shared with others. Since a moral judgment also motivates the subject to spread it, it has the ability to, in a certain sense, make itself true. If regular descriptive beliefs are supposed to adapt the subject to the world, a moral judgment also has the capacity to adapt the world to the subject.

Keywords: Ruth Millikan, teleosemantics, biosemantics, content, descriptive content, meta-semantics, meta-ethics, cognitivism, non-cognitivism, moral objectivity, moral relativism, moral disagreement, moral psychology, evolution of morality

Karl Bergman, Department of Philosophy, Logic and Metaphysics, Box 627, Uppsala University, SE-75126 Uppsala, Sweden.

ISBN 978-91-506-2786-2
urn:nbn:se:uu:diva-391640 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-391640)

For my parents,
Britta Wännström and Dan Bergman
and my brother,
Erik Bergman

Acknowledgements

Philosophy can be a lonesome enterprise. You sit in your office, stewing in your own thoughts, and only intermittently do you benefit from the necessary correctives that are others' perceptions of your work. All the more reason, then, to extend gratitude to those who have supplied these correctives.

I want to thank my supervisors, Sharon Rider and Andrew Reisner, for all their help, encouragement, and patience. Lars-Göran Johansson was my secondary supervisor during an early phase of the work, and he is also due thanks.

Gunnar Björnsson served as opponent on my final seminar. Without his comments, this text would have looked very different. I am deeply grateful for his aid. Additional valuable feedback was provided by my departmental readers, Matti Eklund and Sebastian Lutz. Marcel Quarfood was the opponent at my half-time seminar, an ordeal I apologize for putting him through and thank him for enduring.

I have had the fortune of working with many intelligent, insightful, and pleasant colleagues. Nils Franzén is due special thanks for having read and commented on several late drafts. He and Henrik Rydéhn have provided insight into many difficult philosophical issues over the years, and without them, I imagine I would have been even less of a philosopher.

Other colleagues that deserve gratitude include, in no special order except alphabetical, Tobias Alexius, Per Algander, Johan Boberg, Björn Brunnander, Erik Carlson, Daniel Fogal, Anna Folland, Erik Hallstensson, Elinor Hållén, Erik Jansson Boström, Magnus Jedenheim Edling, Jens Johansson, Kasper Kristensen, Guilherme Marques Pedro, Carl Montan, Olle Risberg, Simon Rosenqvist, John Shaheen, Folke Tersman, Oda Tvedt, Rebecca Wallbank, Tobias Wilsch, and probably others. Rysiek Sliwinski and Anna Gustafsson deserve special mention for their assistance with all things practical and administrative.

Fabian Hundertmark graciously read and commented on a draft of a chapter that was subsequently chopped up and cannibalized for parts.

I presented chapter 3 at the eighth Philosophy of Biology and Cognitive Science research workshop (PBCS8) at Complutense University of Madrid in May 2018. I thank the organizers for giving me the opportunity to participate, and the participants for their comments and suggestions.

The last stretch of my PhD run was made possible by a generous grant from the Göransson Sandviken scholarship foundation. Without its manufacturing industry, Sweden would grind to a halt.

I want to thank my friends for the invaluable services of friendship they have rendered. Special mention is due to Nils Gasslander, who through innumerable conversations throughout the years has influenced my understanding of the human mind in ways I can't begin to mention; to Jonas Bååth, who has offered valuable insights into a sociologist's perspective on morality; and to Alexander Sohlman, who has provided philosophical as well as human companionship during my undergrad years.

Finally, I want to thank my parents, Britta Wännström and Dan Bergman, for making all this possible, and my brother, Erik Bergman, for his commendable discharge of the not always easy task of being my brother. I dedicate this work to them, with immense gratitude and love.

Contents

Introduction

The nature of moral judgments and of the language we use to convey those judgments to each other has long been a focus of interest and source of contention among meta-ethicists, those philosophers that study the nature of morality. *Cognitivists* about moral judgment believe that moral judgments are very much like regular beliefs, such as my belief that diamonds are hard, that Uppsala is north of Stockholm, and that the sun is currently shining. Regular beliefs purport to represent facts about the world, like the fact that diamonds are hard, and the same, claims the cognitivist, is true of moral beliefs. They, too, purport to represent certain facts about the world—moral facts!—and they are true or false depending on whether these facts obtain. The function of moral discourse, moreover, is to report on these moral facts and convey these moral beliefs. According to the cognitivist, moral judgment and discourse are *descriptive* in the sense that they purport to describe the world.

Non-cognitivists, on the other hand, reject the idea that moral judgments are like beliefs that purport to represent facts. To the non-cognitivist, beliefs are more akin to desires, intentions, or plans. Their function is to motivate action and guide conduct, not to describe the world. The job of moral discourse is to convey these action-guiding mental states in order to influence the behavior of others.[1] According to the non-cognitivist, moral judgment and discourse are not descriptive but *directive.* Their job is to direct behavior and action.

Underlying the dispute between cognitivists and non-cognitivists is a gaggle of other issues, psychological as well as epistemological and metaphysical, concerning the nature of morality. But the dispute also raises other questions of a kind more internal to the philosophical study of language and thought. What is it for a judgment to "describe the world," or for a statement to report a fact? What conditions does a piece of thought or an utterance need to meet in order to qualify as so doing? What, for that matter, does it mean for a piece of thought or an utterance to direct action?

To answer these latter questions is the job of *meta-semantics*, the branch of philosophy that studies the nature of *intentionality*. "Intentionality" denotes that characteristic feature of thought and language that allows a

[1] Some classical statements of non-cognitivism include (Stevenson 1944; Ayer 1952, chap. VI; Hare 1952; Gibbard 1990; Blackburn 1998b).

thought or a statement to be about the world, represent it, or describe it (as well as, often, failing to do so). Since the debate between the cognitivist and the non-cognitivist concerns what sort of intentionality moral judgments and statements possess, it seems as though an inquiry into the nature of moral judgment and discourse must be accompanied by an inquiry into the nature of intentionality. Meta-ethics and meta-semantics must be conducted jointly.

In this thesis, I address the question of the nature of moral judgment and discourse with the help of a particular meta-semantic theory: Ruth Millikan's teleosemantics. According to Millikanian teleosemantics, which I will refer to as just "teleosemantics" below, the nature of intentionality inheres in the *teleology* of thought and talk, i.e., their evolutionary functions. Thoughts continue to be thought and words continue to be spoken because they and the systems that produce them have accomplished certain things in the past—more precisely, their ancestors have accomplished certain things in the past—that allow thoughts to go on being thought and words to go on being spoken. It is by looking at how, in this manner, thoughts and talk "pull their evolutionary weight" that teleosemantics purports to clarify the nature of intentionality.

According to teleosemantics, the descriptive content of a thought or statement is, very roughly, the conditions that must obtain in order for the representation to perform its function successfully in a historically normal way. Thoughts, when successful, *adapt* their thinkers, and statements adapt their addressees, to certain circumstances in the world and increase the subject's chances of success in the face of those circumstances. This indirect connection between content and success makes the theory especially interesting to meta-ethical research, I posit, since it promises to supply an account of the success-conditions of moral judgment and discourse that are at once *descriptive* and *practical*.

A common non-cognitivist line is that a cognitivist theory of morality, one that takes moral thought and talk to represent facts, is incapable of accounting for the peculiar *practical* role that moral judgment and discourse have (e.g. Gibbard 2005, 113). It is observed that moral judgments and the discursive tools used to convey them serve to influence the actions of others and the outcomes of collective decision-making in order to coordinate action in social groups. According to the non-cognitivist, this is indeed *all* they do. And for the naturalistically inclined, it has seemed plausible that the practical role of moral thought and talk supplies *the* explanation for observable patterns of moral discourse, i.e., which moral positions people defend, which specific moral claims they make, and which convictions they form on the basis thereof. If that is so, what explanatory role remains for a view of moral thought and talk as *also* fact-stating? If moral thought and talk indeed purport to represent facts, but this fact-representing role makes no contribution to the explanation of *which* moral judgments and statements people make, then the conclusion would seem to be that people's moral thought and talk,

while purporting to represent facts, are epistemically unjustified. And since the consequent attribution of widespread epistemic irrationaliy to people is hard to accept, it is better to deny any fact-representing role whatsoever to moral judgment and discourse (cf. Street 2006).

But note that if, as teleosemantics maintains, the descriptive content of moral judgments and statements consists in the conditions under which these have historically been able to perform their functions successfully, one and the same theory can account for *both* their fact-representing *and* their practical role at the same time. Because if so, the content of moral representations simply consists in conditions necessary to explain their practical success. The epistemic reason for tokening a moral representation (the fact that makes it descriptively accurate) and the practical reason for tokening that same moral representation (the fact that makes it practically successful) are one and the same (cf. Harms 2000; Sinclair 2012; Artiga 2015).

This is, in outline, the idea that I will pursue in the pages to follow. I will ask how, if we suppose moral judgments to be motivational attitudes, we can go about assigning descriptive content to them, and what this entails for how we should understand moral discourse. In developing my views, I have relied heavily on my predecessors. I have mainly drawn from Allan Gibbard's theory of the function of moral judgments in *Wise Choices, Apt Feelings* (1990), and Neil Sinclair's development of this theory within a specifically teleosemantic framework (Sinclair 2012). Sinclair, in particular, exploits the same feature of teleosemantics that I do to assign descriptive content to moral judgments understood as motivational attitudes with practical import. In many respects, the ideas herein are an extrapolation of the program he sketches in his paper.

In one significant respect, however, I diverge from the precedent set by Sinclair, and that is in my discussion of moral discourse. Beginning in chapter 4, I develop a view on (a fragment of) normative discourse that I call "discursive non-descriptivism," since it entails that whereas normative *judgments* have descriptive content, normative *statements* do not. This view, though non-standard, follows more or less directly from the core tenets of teleosemantics together with auxiliary assumptions about normative judgments of the kind described above. I explain and defend this view in chapter 4, and apply it to specifically moral discourse in chapter 6.

One of the teleosemantic premises that ground the inference to discursive non-descriptivism has to do with the role this theory assigns to the *consumer* of a representation in assigning content to it. The consumer of a representation is whatever receives, interprets and makes use of it. In the case of discursive representations such as statements, the consumer is the statement's addressee or hearer. The analytic-philosophical tradition has, by and large, favored the *producer* of linguistic representations in its attempts to explain the nature of meaning and content. This preference is evidenced by Gricean meta-semantics, which tries to assign semantic properties to speech-act on

the basis of the intentions of the speakers who utter them (Grice 1957, 1989). It is also evidenced in earlier versions of naturalistic meta-semantics such as informational semantics, which tries to assign content on the basis of the information carried by the representation, this being a consequence of causal interactions at the representation's source. According to teleosemantics, however, the content of a representation has to do with its function, which is a matter not of how the representation is produced but of what it is supposed to do. A representation's function is an effect it is supposed to have on its consumer. A representation's job, again speaking very roughly, is to adapt the consumer to a certain state of affairs in the world.

In the case of discourse, the consequence is that the descriptive content of a speech-act derives from the content of the attitude it is supposed to produce, not (as has often been supposed) from the content of the attitude it is supposed to express. As long as we are talking about regular descriptive discourse, whose job is to convey belief, this amounts to a distinction without a difference, since the belief produced and the belief expressed have the same content. However, if normative and moral judgments are directive attitudes with descriptive content, it amounts to a real difference in the case of normative statements, whose jobs are to convey normative judgments. This is because, as I argue in chapter 4, tokens of the same directive attitude in the heads of different people will have different descriptive contents. The token in the speaker's head will therefore have a different content from the one produced in that of the hearer. Moreover, if there is more than one hearer, the token attitudes produced in each of their heads will have different contents. And since the content of a statement, on the teleosemantic view, depends on the content of the attitude it is supposed to produce, if the attitude it is supposed to produce has no *particular* content, the statement can have no particular content either.

If normative and moral statements have no particular content, why do we engage in normative discourse, debate, and attempts to convince each other and attain answers to normative questions? To explain why, I exploit what I would like to call the "ecological" picture of discourse implicit in teleosemantics. Since language is intentional, and since teleosemantics analyzes intentionality in terms of functions understood on analogy with biological functions, it is committed to a view of language as akin to a living thing. The title of Millikan's main work is *Language, Thought, and Other Biological Categories*, reflecting this commitment. That language is akin to a living thing doesn't mean that it is organic, only that it persists and proliferates due to mechanisms that bear close analogies to natural selection. But language devices "live" in the social interactions between people, and in order to persist and proliferate they must incentivize their own continued use. This means that they must encourage speakers to go on using them, and encourage hearers to go on responding to them, in predictable ways.

Language forms, like other living things, must adapt themselves to the social world in which they live their lives. For normative statements in particular, this means that they must adapt themselves to the preferences, desires, goals, and needs of people. A normative or moral statement, I will claim, has the function of directing people the same way, but in so doing it must ensure that people are predisposed to be directed that way. Otherwise the language form will be unsuccessful and tend to die out. The means whereby moral statements direct people the same way is by spreading moral judgments, but in order to be spread, others must be receptive to those judgments. In normal cases, this means that the judgments must actually adapt those others to the world, i.e., they must describe the world correctly. Through this dynamic, a semblance of objective truth-conditions for normative statements will emerge, discursive non-descriptivism notwithstanding, as certain normative statements prove better-adapted for proliferating their lineages in the human social world than others.

A true belief adapts its subject to the world. In virtue thereof, it is also itself adapted to the world, since its fortunes are tied up with those its subject. A moral judgment, too, adapts its subject and itself to the world, but it also has the power to adapt *the world* to *itself* and produce conditions amenable for itself. It does so by spreading itself, placing tokens of its own type in the heads of other people and thereby producing what I call a *community of judgment*. Since, as I will claim, moral judgments secure their teleological success by producing coordination among people, by getting itself spread the moral judgment actively produces the conditions required for its own success. Since, according to the teleosemantic, picture, the conditions required for the (historically normal) success of a judgment determines its descriptive content, a moral judgment can in a way contribute to making itself true.

This has been a preview of the views I will be defending in the coming pages. Most of what I will say relies on the assumption that teleosemantics is true, so before I give a rundown of the structure of the thesis I would like to motivate my decision to rely on this particular meta-semantic theory.

The Role of Teleosemantics in the Argument

When I speak of teleosemantics, I primarily have in mind a theory developed by Ruth Millikan in a number of books and articles (Millikan 1984, 1989a, 2004, 2017, etc.) Versions of teleosemantics have been developed by several other authors,[2] and the differences between the various views are often subtle. I have chosen to focus on Millikan's version for a number of reasons. It is the one I know best and feel the most sympathy for. It is developed to a

[2] For example, David Papineau (1984, 1993), Carolyn Price (2001), Karen Neander (2017), and Nicholas Shea (2018).

degree of detail that none of the other versions are. Most importantly, for present purposes, Millikan has paid more attention to how the teleosemantic ideas apply to language and discourse than any other teleosemanticist, and we are going to need an understanding of language if we are to understand such a language-laden phenomenon as morality.

I personally believe that teleosemantics is largely correct and that, for this reason alone, it is worth engaging in teleosemantic meta-ethics. But why should somebody who is less convinced than I am about the value and explanatory power of Millikanian teleosemantics take interest in an exercise of this kind? Consider two approaches one could take to the task of adjudicating claims made in the debate between cognitivists and non-cognitivists. Such claims include, for instance, that the function of moral judgments is to represent facts or, alternatively, to direct action. The first, more common approach is simply to start with the claims themselves, to attempt then to draw out their consequences, and finally to evaluate those consequences for plausibility. Some application of this method is of course indispensable in any kind of philosophy. But as long as the claims are made in general terms, like "the function of a moral judgment is to represent the facts," it will be an open question exactly what their consequences are. This need not be a vice, especially not in the early stages of inquiry, when views have yet to take determinate shape and freedom of exploration and the generation of new ideas is at a premium. But in due time, a defender of a view like that must commit themselves to a determinate interpretation of their claim, i.e., a determinate view about what follows from it, what its alternatives are and what follows from them, and how the plausibility of these consequences are to be evaluated. That amounts to a theory. It will be a theory not just about moral judgment but about judgment generally, because it has to say something about what it means for a judgment to have this or that function, what different functions there are, what a judgment is, and so on.

Another approach is to come to the question with a theory of thought and talk already in hand, one that has been developed not out of any particular concern for *moral* thought and talk but simply out of a general theoretical interest in thought and talk as such. If this theory is sufficiently general and powerful, it should be applicable to the case of moral thought and talk as well. It will not, perhaps, allow us to adjudicate between rival views, but it will allow us to get new perspectives on them. It will permit us to draw out their consequences in detail and reveal options and implications that have been overlooked. Teleosemantics is a theory of this kind. It encompasses a number of quite specific principles for answering questions such as, "under what conditions does a judgment or a piece of discourse possess a given function?", "what conditions must a judgment meet in order to count as representing facts?", and so on. It has been developed without any particular view to solving problems in meta-ethics. So, granted that teleosemantics has *some* independent plausibility, the exercise of applying it to meta-ethical

questions serves to advance our understanding of these questions and gives us new options for answering them, some of which I have indicated above.

In order to reveal these options, I have systematically presupposed the truth of the main tenets of Millikanian teleosemantics. The current thesis is therefore an exercise in theory *application*. This is not to say that there won't be cause, in the course of the discussion, to submit the theory to criticism and propose modification. I do that in section 2.5, and again in chapter 3. But these critiques and modifications are peripheral. I am primarily interested, not in discussing teleosemantics, but in using it. I could not, at any rate, hope to defend the theory better than others have already done.

Importantly, however, the application of teleosemantics to a domain where it has seen scant application thus far can itself be seen as a test of the theory. If teleosemantics can advance our understanding of moral judgment and discourse, this will in itself lend credence to the theory. Conversely, if the attempt to apply the theory to this domain involves us in confusions and riddles, this will weigh against it. The thesis can therefore be read as serving two purposes simultaneously. It tries to apply rarely-used tools to some old problems in meta-ethics and thereby advance our understanding of those problems, and it serves as an indirect test of teleosemantics itself. In the first capacity, it directs itself to meta-ethicists interested in acquiring new perspectives on their subject-area. In the second, it directs itself to teleosemanticists interested in exploring the power and scope of their favored theory.

One feature of teleosemantics that should make it interesting for meta-ethical purposes is its naturalism. The naturalistic commitments of teleosemantics consist in its aspiration to provide a constitutive explanation (or, as I will prefer to say, a metaphysical analysis) of intentionality that appeals only to natural phenomena, which we can gloss, roughly, as phenomena of the same kind as those studied by the natural sciences. The naturalist about a subject-matter is somebody who aspires to provide an account of that subject-matter that allows us to understand its *place* in relation to the rest of the natural world. She believes that the subject-matter can be *situated* in the world that science describes. The naturalist *simpliciter* is somebody who believes that everything can be situated in this manner: that everything that exists is natural or, perhaps, that nature—in its alternative sense of "reality"—constitutes a coherent whole. These descriptions obviously fall short of precise definitions. However, the definitional issues seldom become relevant in debates among naturalists. Everyone involve seems to have a fairly firm, if intuitive, grasp of what the relevant desiderata are. I hope I can rely on a similar unspoken understanding with my readers.[3]

[3] At the end of the day, we don't judge a theory on whether it is "naturalist," but on whether it explains the phenomena. I suspect that what distinguishes naturalists from non-naturalists, more than anything else, is their different conceptions of what constitutes a satisfying explanation. The naturalist wants her explanations to show her how the target phenomena hang together with the rest of nature and science in the broadest possible picture.

Teleosemantics is (or aspires to be) a consistently naturalist program. Naturalism is also an important desideratum for many meta-ethicists, and non-cognitivism in particular is often motivated—historically as well as in contemporary arguments—by naturalist commitments. If these commitments are naturalist in some sense sufficiently similar to the one in which teleosemantics is naturalist, this alone should suffice to make teleosemantics worth meta-ethical consideration.[4]

When it comes to naturalism in meta-semantics, Millikan's teleosemantics is one of few existing theories that attempt to give a completely general account of intentional phenomena and does so to a high degree of detail. When I call it "naturalist," I do not only mean to contrast it with avowedly *non*-naturalist theories of intentionality (of which there are some), but also with such theorizing about semantic phenomena that simply remain agnostic about the relationship of intentional phenomena to the natural world. Theories of the latter kind may or may not be consistent with naturalism, but they do not submit that consistency to a systematic *test* in the way Millikan's theory does. Thus, we have no concrete grounds for confidence in their ability to ground a naturalist theory of moral language, and any theorizing of the latter kind that is done with their help will have a provisional status.

The careful detail in which Millikan's teleosemantics is worked out, the range of phenomena it tries to account for, and its often oblique and innovative approaches to old problems almost guarantee that even if it should be wrong, it will be interestingly wrong. In other words, it will be important for future efforts to work out systematic naturalist theories of intentionality to study and understand the approaches of Millikanian teleosemantics and how they failed. That goes for anyone attempting to give a naturalist theory of *moral* thought and talk as well as anyone simply studying thought and talk in general. Thus, even if teleosemantics is wrong, the meta-ethicist ought to study its consequences for theories of moral judgment and discourse.

Structure of the Thesis

The first two chapters are entirely devoted to an introduction to the field of naturalist meta-semantics (chapter 1) and to teleosemantics specifically (chapter 2). The purpose of these chapters is both to introduce the main concepts that I will be relying on in the rest of the thesis and to orient the reader

[4] I do not deny that one can be a naturalist in meta-semantics and a non-naturalist in meta-ethics, or vice versa. Many who defend naturalist views in specific domains, however, are motivated to do so by a commitment to naturalism *simpliciter*—the view that everything is natural. Even those who are agnostic about this global view may accept the methodological principle that we ought to aspire in our theory-building to give naturalist accounts of as wide a range of phenomena as possible. For meta-ethicists who hold either of these views, a naturalist meta-semantics ought to be highly interesting.

in the background of and motivation for teleosemantics. Readers who are already acquainted with teleosemantics and the underlying discussion can skip or skim these chapters.

In chapter 3, I discuss an argument due to Marc Artiga (2014) that purports to show that if teleosemantics is true, all representations, including propositional attitudes like beliefs and desires, are *hybrid* in the sense that they possess both descriptive content and a directive function. I assess Artiga's argument, conclude that it is sound, and argue that teleosemantics is no worse off for it.

In chapter 4, I evaluate the possibility of accounting for some of the intuitions speaking in favor of cognitivism on the non-cognitivist premise that normative judgments are directive attitudes by exploiting the conclusion of chapter 3 that even directive attitudes have descriptive content. I try to assign descriptive content to a range of directive attitudes, culminating in judgments conventionally expressed by sentences of the form "*A* ought to φ," where *A* denotes an agent and φ an action. I observe that my proposal has the implication that different tokens of such judgments, in the heads of different people, have different descriptive content, and that therefore the sentences conventionally used to express and convey those judgments have no descriptive content at all. I try to explain how ought-statements can nevertheless be truth-apt and governed by more-or-less intersubjectively valid standards by developing the view I call *discursive non-descriptivism*. According to this view, your assessment of the truth of an ought-statement amounts to an assessment of whether the token judgment that the statement is supposed to produce in *your own* head would be descriptively correct. I also sketch a program for a compositional semantics compatible with discursive non-descriptivism.

In chapter 5, I propose an account of the evolutionary function of *moral* judgments that I call the *coordination of responses hypothesis*. This view, which builds on the works of Neil Sinclair, Allan Gibbard, and others, entails that the function of a moral judgment is to produce widespread conformity to some pattern of action within a population, and that the mechanism whereby it normally does this is by coordinating responses of approval and disapproval, including overt sanctions. I make a distinction between the *adapted* and the *invariant* function of a moral judgment, where the former consists in the aforementioned production of widespread conformity whereas the latter consists in bringing about the ultimate, evolutionarily beneficial outcomes that such widespread conformity has historically contributed to. I discuss various views on the invariant function of moral judgments, including the popular *cooperation view*: that the function of morality is to help bring about cooperation.

In chapter 6, finally, I discuss consequences of the coordination of responses hypothesis for long-standing debates about the objectivity and absoluteness of ethics. I propose a definition of moral objectivism using the con-

cepts I have developed and try to assess whether the hypothesis can vindicate objectivism as thus defined. I also offer a tentative proposal for how the objectivist features of moral thought and discourse can be accounted for, should objectivism turn out to be false, by again using the resources of discursive non-descriptivism developed in chapter 4.

The following investigations will of necessity be exploratory and tentative. The teleosemantic literature is generally focused on fundamental questions in the philosophy of intentionality, questions that often concern subpersonal representational states, the mental states of lower animals, and how and whether these accounts can be generalized to apply to human propositional attitudes. In contrast, relatively little attention has been paid to how the teleosemantic ideas interact with more conventional problems in the philosophy of language: problems concerning truth, inference, and the semantics of linguistic expressions. These problems, however, often come to the fore when discussing the nature of moral judgment and discourse. When problems of this nature have arisen in the course of my discussions, I have sometimes had to rely on somewhat speculative solutions.

I have also had to make many assumptions regarding normative and moral psychology and the evolution of morality. There is a vast literature on these topics, and I have only scratched the surface. Much of this literature, moreover, brings to the fore deep conceptual questions about, for instance, the nature of morality itself, or what qualifies as a moral judgment. I broach some of these issues in the text, but claim no conclusiveness for my attempts to address them.

I hope, therefore, that the ideas presented here will be received in the spirit in which they were intended, as proposals rather than as definitive claims. Hopefully, the general ideas have some merit independently of the details.

1. Naturalism in Meta-Semantics

Teleosemantics, I said in the introduction, is a theory in meta-semantics, and meta-semantics is the philosophical study of the nature of intentionality. Teleosemantics is also naturalistic, meaning that it aspires to understand intentionality as a natural phenomenon. Naturalistic meta-semantics, as a field of research, is defined by its attempt to discharge this task, which I will describe as the attempt to supply a *naturalistic analysis of intentionality*. To understand teleosemantics, we must first understand this task itself, what it entails, and the problems encountered along the way. To provide that understanding is the purpose of the present chapter.

To give a naturalistic analysis of intentionality is not, regrettably, a simple matter of applying analytical tools to an already clearly delineated subject-matter. The notion of intentionality is semi-technical and intuitions about it theory-dependent. Our approach to it must therefore of necessity be somewhat indirect and circumspect. Intentionality is best understood as a generalization over a number of more familiar, everyday ideas that seem to have interesting things in common, ideas like that of the *meaning* of a word or phrase, what some claim or thought or idea is *about*, what a word *refers* to, what *is said* in a statement, what somebody *has in mind*, and so on. The notion of intentionality is meant to capture the fact that all these everyday ideas seem to imply a way for language or mind to be related to the world, to *stretch out* towards it, as it were.[5] Often enough, the ideas also imply that a standard of success is involved: in stretching out towards the world, our language and thought can manage to reach it, or fall short. What we say, or think, can be false. What we desire to happen can fail to come about.

In section 1.2, I will offer a characterization of the phenomenon of intentionality in terms of the interdefinable notions of *representation* and *content*. I will discuss the relation between content and truth-conditions, and discuss the idea that content is normative. Before that, however, I must explain what I mean by a naturalistic analysis. That will be the task of section 1.1.

In section 1.3 I offer some preliminary remarks on what it means to give a naturalistic analysis of intentionality specifically. What constraints bear on this task? What problems does it involve us in? In section 1.4, I discuss *indicator semantics*, a broad family of approaches in naturalistic meta-semantics

[5] The word "intentionality" itself comes from the Latin "*intendere*", which means "to stretch."

from which teleosemantics is an outgrowth. This discussion will prepare us for understanding teleosemantics when I present it in chapter 2, by acquainting us with its fundamental commitments as well as with the problems arising from those commitments that it constitutes an attempt to solve.

1.1. Naturalistic Analysis

The subject-matter of naturalistic meta-semantics is intentionality. What naturalistic meta-semantics attempts to accomplish with respect to this subject matter is a *naturalistic analysis* of it. As I will use this term, it means explaining *in virtue of what* an entity has the intentional properties that it does have. Let me clarify.

As commonly understood, philosophical analysis is in the business of producing necessary biconditionals on the following form:

(BICONDITIONAL) Necessarily, for every *x*, *x* is *F* if and only if *x* is *R*

Different varieties of analysis entail different ways of understanding what underpins a biconditional of this form when it constitutes a successful analysis. For instance, if we are engaged in *conceptual* analysis, the *analysandum* is the concept expressed by *F*, and we want *R* to *mean the same thing* as *F* or to express, explicitly, the conceptual components already implicit in the concept expressed by *F*. When we have accomplished this, the BICONDITIONAL should count as an analytic truth.

Naturalistic meta-semantics is not, however, primarily engaged in conceptual analysis. The project is metaphysical. A *metaphysical* analysis is an analysis, not of the concept expressed by *F* but of the entity or property that *F* refers to. This kind of analysis will also produce BICONDITIONALS. Here, the *analysandum* is *F*'s referent, and *R* constitutes the description or set of conditions whereby *F* is analyzed. Different metaphysical relations can underpin and explain a biconditional of the above kind. One possibility is property identity: the properties denoted by *F* and *R* respectively are one and the same property. Another is grounding, commonly understood as a metaphysical explanatory relation that can hold between *distinct* facts (cf. Rosen 2010).

I will be assuming that when we engage in naturalistic analysis, we are ultimately interested in finding out what things *are*. Much like the physicist who identifies heat with average molecular kinetic energy, or the chemist who identifies water with a certain chemical with the molecular composition H_2O, we are trying to produce descriptions that refer to familiar phenomena in more perspicuous ways, allowing us to subsume them under general laws

and understand their relations to other phenomena. We are searching, in other words, for property *identities*.[6]

Obviously, not every necessary, universally quantified biconditional counts as a good analysis, even provided that it is true. The *goodness* of the analysis depends on the description substituting for *R*, and whether that description gives us a deeper understanding of the target phenomenon than we had before. The naturalist, in particular, wants to supply a *naturalistic* analysis, which means, to a first approximation, that the terms making up *R* should all refer to entities and properties that are natural, i.e., of the same general kind as the entities studied by natural science (cf. p. 15).

One thing that I will put particular stress on is that a good analysis should allow us to see why the target phenomenon is important enough that we have bothered to keep track of it using dedicated vocabulary. An analysis that renders this fact mysterious, I claim, is *prima facie* a bad analysis. This assumes a certain picture of human language use: only things that are salient enough, recurrent enough, and significant enough are awarded conventional linguistic labels. I will discuss different ideas about what makes intentional notions important beginning on p. 33 below.

We turn now from the *form* of analysis to its *method*. Traditionally, philosophical analysis has employed the method of intuitions and counterexamples: constructing a candidate analysis and then attempting to come up with examples of (possible) entities that satisfy the description on the right-hand side of the biconditional but that, intuitively, fail to count as instances of the *analysandum*, or *vice versa*. In essence, this is also the method used by naturalist meta-semantics, and we will see many instances of it below. However, it is worth saying a few words about its limitations and pitfalls.

As already mentioned, I take the task of naturalist meta-semantics to be metaphysical rather than conceptual analysis. According to an influential tradition in philosophy, the inferences that are constitutive of possession of a concept, and therefore count as analytic, are accessible through *a priori* conceptual intuition. In contrast, many naturalistic meta-semantic theories, including Millikanian teleosemantics, entail strong versions of semantic *externalism*, according to which competence with a term does not necessitate any particular knowledge about the term's referent, *a priori* or otherwise.

This doesn't necessarily mean that intuition is worthless as a tool for analysis, only that "intuition" must be understood as a more general capacity, one that engages high-level, abstract understanding of the world, much of

[6] Teleosemantics, at least of the Millikanian variety, is generally hostile to conceptual analysis as traditionally understood. The ideology of conceptual analysis as a method relies on the possibility of *a priori* knowledge of sameness and difference of meaning, which is something Millikan's *meaning empiricism* entails is impossible (Millikan 1984, 325–33). As for the idea that meta-semantics should be in the business of searching for *a posteriori* property identities like those pursued in natural science, it is the understanding of the endeavor favored by Millikan (1989b, 16–17) as well as other teleosemanticists like David Papineau (1993, 93).

which has an empirical origin. It also means that, though we can be confident that our intuitions *collectively* paint a fairly accurate picture of the world,[7] any individual intuition can in principle be inaccurate. As a consequence, an analysis that revises some of our intuitions about the target phenomenon cannot for that reason automatically be ruled out (cf. Millikan 2010, 36–37). Candidate analyses must be assessed holistically, by their capacity to give an account of the analysandum that is maximally respectful of our intuitions while at the same time being maximally consistent with pre-existing commitments in other domains.

For this and related reasons, Millikan approaches the task of naturalistic meta-semantics in a *constructive* fashion. Rather than directly approaching the target *analysanda* and attempting to determine necessary and sufficient conditions for them, she proceeds in a top-down fashion, constructing an overarching theory using stipulated concepts and attempting to show how this theory can account for the target phenomena in an overarching and systematic way (Millikan 1989b, 14–15). I sympathize with this approach, and I will proceed in a similar way in the present work.

I have now outlined the nature of the task that teleosemantics has set itself: to provide a naturalist analysis of intentionality. In the rest of this chapter, I will explain in more detail what the task entails, the various desiderata that bear on it, and how the shortcomings of earlier approaches have motivated the development of teleosemantics, all in the hope that this background will allow the reader to better understand the latter theory when it is introduced in the next chapter.

The first step is to get clearer about what intentionality is. That will be the task of the next section.

1.2. Intentionality: Representation and Content

I have mentioned the difficulties involved in characterizing the notion of intentionality. An encyclopedia entry defines intentionality as "the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs" (Jacob 2019). But this definition, despite its generic character, is already tendentious, because it excludes many phenomena that others would be happy to treat as intentional: sentences and speech-acts, animal warning calls, even (to some extent, at least) systems like the magnetosomes that indicate the direction of oxygen-poor water, possessed by certain species of ocean-living bacteria to whom few would be tempted to attribute minds (Dretske 1986, 26–27; Millikan 1989a, 290).

[7] We must, at any rate, allow ourselves this assumption, or we would lack a method altogether.

It is clear that sentences, assertions, thoughts, and even magnetosomes can all be described as being in some sense "about" the world, or as "representing" it. It is less clear that this common way of talking reveals an underlying similarity of ontological structure. The existence of an underlying structure is something that must be demonstrated through theoretical analysis, by supplying a theory that assigns common properties to the diverse phenomena mentioned above and thereby explains why we use the same locutions to talk about them. Naturalistic theories of intentionality have, as a general rule,[8] fallen on the side of attempting to treat a maximally wide range of phenomena under a unified account.[9] This is a strategy that can only be justified by its fruits—by its capacity to actually yield a unified explanation of the phenomena it considers to be intentional. For the moment, I will have to treat it as a mere assumption, but I hope the reader will see that the assumption is rewarding.[10]

If thoughts, sentences, animal warning calls, and magnetosomes are all instances of the same general phenomenon, we need a term for the kind that they all instantiate. I will use "*representation*." Representations, I will say, are entities that bear *content*. Content is what makes a representation into a representation, and so everything with content is a representation and all representations have content.

So what is content? The term "content" will be familiar from the philosophy of language and mind, where it typically denotes *propositional* content, the feature of beliefs and sentences that determine their truth-conditions and their inferential relations to one another. The notion of content used in this thesis is essentially the same, with qualifications to be mentioned shortly. The content of a representation, I will say, is something that determines how the world must be in order for the representation to qualify as representing it correctly. We can therefore think of content as a possible state of affairs—one that may or may not be actualized—to which the representation bears a

[8] Though not one without exceptions. See (Fodor 1990, 47; Shea 2018).

[9] However, it is worth mentioning in passing that not all theories of intentionality vindicate this unification hypothesis. Some, which we can call "two-step theories," make a distinction between *original* and *derivative* intentionality (the terminology is from (Brandom 1994, 58 ff.)). According to such theories, there are at least two types of representations, and some of them derive their intentionality from the intentionality of the other(s). Typically, such theories distinguish specifically between linguistic and mental representations (not necessarily using this terminology). They can then be further subdivided into two categories, depending on which of the two types they take to have original, and which derivative, intentionality. One can take the intentionality of thought to be original, and the intentionality of language to be derived from it. A view like that has been defended by Paul Grice, among others, who held that the meaning of a speech-act is derived, in complex ways, from the communicative intentions of the speaker (Grice 1989). Conversely, one can take the intentionality of language to be original and the intentionality of thought to be derived from it. A view like that has been defended, for instance, by Hartry Field (1978).

[10] This is not meant to imply that the cases don't differ in important ways. They clearly do. For instance, one difference between linguistic representations and simpler representations like animal warning calls is the recursive structure that the former, but not the latter, possess.

relation. Alternatively, we can think of it as a property of the representation that *picks out* one of these possible states of affairs. Not much hinges on the distinction, and I will freely move between these two ways of talking in what follows.[11]

A representation's content determines how the world must be for it to represent it correctly, but there are several ways to represent the world. The world can be represented *descriptively*. Descriptive representations purport to say what the world is like. They include such things as beliefs and factual assertions which, when the conditions specified by their content obtain, are *true* or *correctly describe* the world. We will call their content *descriptive content*. Another way to represent the world is *directively*. The content of a directive representation picks out conditions that must obtain in order for it to be *satisfied*. Directive representations comprise such things as commands and desires, which say, not how the world is, but how it is to be made. We will call their content *directive content*.

There may be yet other ways to represent the world. In this and the following chapter, for simplicity's sake, I will restrict my discussion to descriptive representations and descriptive content, which have been the main concerns of the meta-semantic literature. In chapter 3, I will complicate the picture by discussing other kinds of content such as directive content, as well as introducing the idea that a representation can have several different contents.

In paradigmatic cases, then, the notion of content employed here subsumes the familiar notion of propositional content, at least if the latter is understood in a particular way. Later, we will see that the notion of content relevant for understanding teleosemantics is "unstructured." Contents do not, as propositions are sometimes taken to do, bear their inferential relations on their sleeves. They have no compositional structure of their own.[12] Relatedly, and again in contrast to one common understanding of propositions, they are *intensionally* rather than *hyperintensionally* individuated.[13] These niceties will only become important later (section 2.4), but for this reason, and because I will find other uses for the notion later in chapter 4, I will be talking about contents and truth-conditions but avoid propositions-talk in what follows.

[11] It is natural to think that the content of a belief, for example, is simply *identical* to its correctness-conditions, and that is essentially the line I'm taking here, but on some views, especially if contents are identified with structured propositions (see below), contents "cut finer" than correctness-conditions so that one and the same set of correctness-conditions can be picked out by several different contents.

[12] The view that propositions are structured entities is associated with Frege (1948, 1956) and Russell (1903, 47), who defended different versions of the idea. As the notion of content is used here, it is more closely associated with propositions *qua* sets of possible worlds.

[13] This means that roughly, two contents are the same just in case they pick out conditions that obtain in exactly the same possible worlds. Cf p. 69.

Our notion of descriptive content, furthermore, subsumes the notion of truth-conditions. However, many of the entities that can be profitably treated as representations with descriptive content do not, intuitively, possess truth-conditions. Many sub-personal cognitive representations, for instance, are likely to be more akin to maps than to sentences. Maps, intuitively, are not truth-apt, although they do describe the world. They have correctness-conditions rather than truth-conditions (cf. Rowlands 2010, 116). I will continue to assume that *when* a representation has truth-conditions, they coincide with its descriptive content (though I will qualify this assumption in chapter 4). I will go on using "true" and "truth" indiscriminately about representations that correctly describe the world, but when precision is required I will use "descriptively correct" instead.

Representations comprise what is in an important sense the most fundamental and analytically basic class of intentional entities. But not everything that is intentional is a representation, because not everything intentional has content in the above sense, i.e., truth- or satisfaction-conditions. A sentence can have truth-conditions, but the words that compose it do not. The words instead make systematic *contributions* to the content of the sentences they help compose. We began this section with the question what intentionality is, and if we take into account that there are two kinds of intentional entities—representations and sub-representational entities that help determine the content of representations—we can define intentionality as follows:

> (INTENTIONALITY) Intentionality is that property which an entity or a feature of an entity has in virtue of possessing content *or* in virtue of contributing to the content of the entity whose feature it is.

We can now give a more precise statement of the task of meta-semantics: to determine, for any given entity, whether and in virtue of what it possesses content (is a representation); and for any feature of a representation, in virtue of what it makes the contribution it does make to the content of the representation, if any.

Some words on features of representations. A linguistic representation, such as a speech-act, has linguistic features: it is formed by uttering a certain sentence, with certain syntactical structure and so on. But it also has non-linguistic features, like the time and place of utterance and the person who utters it. These can all contribute to the content of the speech-act (the assertion that *I am here* has different content depending on who utters it and where it is uttered). Other features of the speech-act, like the exact pitch at which it is uttered, are irrelevant to its content. The features of a representation that informs its content can be called its *significant features*.

Non-linguistic representations also have significant features. Take, for instance, the waggle dance that honeybees make to guide their nest mates to nectar (a favored example of Millikan's). The waggle dance has two signifi-

cant features: the duration of the waggle, and its angle to the perpendicular. Its content, *there is nectar at location* L*,* is determined as a function of the value of these two parameters: the duration of the waggle corresponds to the distance to *L*, and its angle corresponds to the angle between *L* and the sun, with the hive at the vertex. We can thus give a "syntax" of sorts for the waggle dance and, on the basis thereof, a "semantics" (though it will not, of course, be the kind of recursive syntax and semantics we are familiar with from the study of language).

It is an important fact about representations that one and the same meaning-bearing feature can be shared among several representations belonging to the same general class and that when this is the case, the contents of the representations sharing the feature typically relate to each other in predictable ways. This phenomenon is well-known in the case of language. Two linguistic representations containing the same word typically have contents that involve the same thing, namely, the word's referent or denotation. Similarly, two linguistic representations with the same syntactic structure have contents where the referents of their component terms are related in analogous ways. These two features of language are closely related to the principle of compositionality, a fundamental principle of semantic theory:

> (THE PRINCIPLE OF COMPOSITIONALITY) The meaning of a complex expression is a function of the meanings of its parts and of the way they are syntactically combined.[14]

But the phenomenon generalizes: two bee-dances made at the same angle to the hive represent nectar at the same angle to the sun.

A naturalistic meta-semantics should therefore allow us to explain how the content of a representation is constrained by its significant features, and also how the intentional properties of the significant features relate to the content of the representations they are features of. We will see how teleosemantics addresses these issues in sections 2.1 and 2.4.

1.2.1. Content and Normativity

Using the notions of representation and content, defined in generic terms, I have tried to give a characterization of intentionality that is maximally neutral between various possible theoretical commitments, while still compatible with the commitments of teleosemantics. I have not been able to remain *entirely* neutral, and some of the assumptions I have made will pay off only later, as we learn how teleosemantics understands the notions of content,

[14] For an introduction to the notion of compositionality, and a discussion of different formulations of the principle, see (Pagin and Westerståhl 2010).

representation and intentionality generally. Hopefully, I have at least been able to present a recognizable picture.

The notion of content, in particular, while it may be familiar as long as it is thought of in terms of beliefs and assertions, starts looking strange once it is abstracted away from those paradigm cases. I have tried to capture it in terms of conditions that need to obtain for a representation to correctly represent the world, but while this characterization hopefully elucidates the relations between the intentional notions, it is obviously of no help unless we already have some independent grasp on the notions involved. It doesn't allow us to understand intentionality in terms of anything *non*-intentional.

One way in which philosophers have attempted to acquire an independent grasp of intentionality is via the idea that content provides *norms* for when to token a representation. This idea, or something very much like it, finds expression in the common dictum that "meaning is normative" (cf. Kripke 1984, 23–37). It is also associated with the common view that a theory of content has to explain the possibility of *mis*representation. A norm is essentially something with which one can fail to conform, just like a representation is essentially something that can fail to accurately represent the world.

In what sense is content normative, if indeed it is? Perhaps it is normative in the sense that it gives a norm for the *person* who proposes to be the subject of a representation, an "ought" that forbids the subject from tokening the representation unless its correctness-conditions obtain. Applied to assertions, this principle entails the normative claim that one ought to ensure, before one speaks, that what one says will be true. If the "ought" in question is understood as a *pro tanto* rather than an *all-things-considered* ought, this claim is plausible enough. Already when we apply the principle to beliefs, however, things start to look stranger. It then entails the claim that before one forms a belief, one ought to ensure that the belief one forms is true. The strangeness here is due to the plausible principle that *"ought" implies "can,"* i.e., that for one to be under an obligation to do something, that something must be under one's voluntary control. But it is far from obvious that the forming of a belief is under a person's voluntary control.[15]

As we go from human, person-level propositional attitudes to representations employed by simpler organisms or as part of the unconscious processing of the cognitive system, the idea that contents specify oughts for the representation's subject becomes even more implausible. Simple organisms, surely, cannot be subject to norms, and sub-personal cognitive representations either have no subject at all, or if they do—if their subject is identified

[15] For more detailed discussion and criticism of the view that the contents of beliefs specify norms for their subjects, see (Glüer and Wikforss 2009).

with the person in whose mind they occur—the *"ought" implies "can"* problem returns with a vengeance.[16]

Despite all this, it is difficult to shake the impression that there is a quite plain sense in which contents are normative and generate oughts. It seems like a perfectly acceptable thing to say that beliefs ought to be true. But if these oughts are not in the first instance oughts for *persons*, then what kind of oughts are they?

The teleosemantic tradition has proposed that the norms associated with content are *teleological* norms, norms that specify the conditions under which a system can be said to function properly. Just like a heart *ought* to beat rhythmically or a knife *ought* to be sharp, a representation *ought* to be tokened only under certain circumstances. Millikan calls the conditions required for an entity to fulfill its function in a proper way its *Normal* conditions, where "Normal" is intended to evoke the notion of a norm, and has been capitalized in order to mark it out as a semi-technical term. The notion of Normal conditions will play an important role in this thesis. We will return to it in section 2.1.

Are teleological norms really "norms" in the same sense as agent-level norms? This is a question that, ideally, a teleosemantically informed meta-ethics (or meta-normative theory) should suggest an answer to. I think we can say one thing outright, albeit a bit impressionistically, and that is that both teleological norms and person-level norms create certain kinds of expectations, something like *justified expectations of good outcomes*. Given that circumstances are otherwise normal, if a person is subject to a norm, we have some reason to expect that she will conform to it, and given that she does conform to it, we have some reason to expect that the outcome will be good. Similarly, given that circumstances are otherwise normal, we have some reason to expect that an entity subject to a teleological norm will function in accordance with it, and we have some reason to expect that if it does, the outcome will be good. I will defend these claims with respect to teleological norms on p. 52, when we have the teleosemantic theory of teleological norms in hand. To defend the corresponding claim for person-level norms will have to wait, however, until the very end.

Even if I'm right that there is this underlying unity between teleological norms and person-level norms, that in itself does not mean that the teleologi-

[16] If one is wedded to the idea that contents specify norms for representational subjects, one may be tempted to conclude that simple organisms do *not* possess representations, or do not do so in the same sense as human subjects do. Robert Brandom defends a view according to which intentionality is constitutively normative, and draws the appropriate conclusion that only humans possess *original* intentionality, other beings having intentionality only in a *derivative* sense (Brandom 1994, 58 ff.) Brandom's theory is not naturalistic, and so I will not contend with it here, but it strikes me as a less convoluted view overall to treat intentionality as a unified phenomenon that bears no essential relation to norms for representational subjects, and view whatever oughts there are as specific to how humans relate to their representations.

cal norms for a representation *themselves* are, or in any way directly entail, norms for the person who tokens them. Rather, if contents are teleological norms, the connection between content and person-level norms for representational subjects becomes *less* direct than Kripke, and those following him, have supposed. I will discuss the relation between speech-act content and the norms governing speech on p. 64, in connection with my discussion of Millikan's theory of discourse.

1.3. Naturalizing Intentionality

In the last section I gave a definition of intentionality in terms of the notions of *representation* and *content*. In this section, I will begin to introduce the problem of giving a *naturalistic analysis* of intentionality—to determine those natural properties in virtue of which an entity has the intentional properties that it does have—and suggest the shape of a solution.

There are a number of different approaches to intentional phenomena in naturalistic philosophy of mind. Teleosemantics is a *realist* theory of intentionality, contrasting with, e.g. instrumentalist (Dennett 1971), fictionalist (Egan 2014), and eliminativist (Churchland 1981; Stich 1983)[17] approaches. In this section, I will discuss a number of broad metaphysical commitments that teleosemantics has in common with other approaches within naturalistic realism about intentionality and try to explain what motivates them. By looking at these commitments, we will get an understanding of some of the central problems of naturalistic meta-semantics, which will help guide the discussion to come. Keeping in mind that teleosemantics attempts to be a *general* theory of intentionality that accounts for linguistic as well as mental intentionality, I will show how these commitments generalize to the linguistic case, where relevant.

1) *Mental representations are particular physical events or states*. Representations are states or events rather than things. It is not a symbol as such, but *that* the symbol is instantiated in so-and-so circumstances in this-or-that system, or *that* the system undergoes these-and-these processes involving the symbol, that represents the world a certain way. To say that these states and events are *physical*—i.e., consist entirely of physical entities and properties subjected to physical laws and capable of causal interaction with other physical states and events—is simply to commit oneself to naturalism.[18]

[17] Churchland 1981 is the *locus classicus* for eliminativism, but although Churchland is an eliminativist about the folk-psychological propositional attitudes, he is not an eliminativist about intentionality *per se*.

[18] Mental representations are often assumed to be states of, or events in, the brain, but it makes little difference to the view, at this level of abstraction, whether these states/events are exclusively defined with respect to the brain, or whether they can involve extra-cranial entities and features, as per (Clark and Chalmers 1998).

The reason for construing representations as events or states rather than things, and as *particular* events or states rather than event or state types, is that they should be able to stand in causal relationships to other particular states and events, like states of the world and behavioral events. They should be able to stand in these relationships because they should be able to *explain* behavior and be *explained by* states of the world (see below). Specifically, they should be able to explain particular behavioral episodes and be explained by particular states of the world. Presumably, states and events rather than things are the primary relata of causal relations. And presumably, only *particular* states and events can stand in the relevant sort of causal relations to other particular states and events (Millikan 2017, 125).

When this principle is generalized to the case of language, it requires us to be careful about describing *sentences* as linguistic representations. When philosophers speak of sentences they often mean sentence *types*, which are *not* concrete particulars but rather abstract entities that can best be understood as *significant features* (p. 25) of concrete particulars such as speech-acts. Sentence *tokens*—these concrete marks on the page, or those concrete uttered sounds—are concrete particulars, and can therefore be representations. Speech-acts are events, and they can also be representations. When I talk about linguistic representations, I have these two categories in mind.

2) *The intentional properties of mental representations supervene on their physical features*. Supervenience is the relation that holds between two classes of properties A and B just in case, roughly, there cannot be a difference with respect to the A-properties without a difference with respect to the B-properties (Hare 1952, 145). The A-properties are then said to supervene on the B-properties. Supervenience on physical properties is normally understood to be a minimal requirement for naturalism or physicalism in the philosophy of mind, and so this second commitment is simply a further way of expressing commitment to naturalism. Supervenience, being a "weak" metaphysical relation, is compatible with a number of "stronger" metaphysical relations, up to and including identity (everything trivially supervenes on itself). By framing the naturalist commitment in terms of supervenience, we can therefore remain neutral with respect to controversies about the exact metaphysical relation involved (cf. section 1.1).

The supervenience principle entails that two representations cannot differ with respect to their intentional properties without differing with respect to their physical properties. In evaluating this claim, it is important that we are sufficiently liberal in what we count as a representation's physical properties. We cannot only count the "intrinsic" physical properties of states (such as their shape), but must also count some of their relational properties. This is for the following simple reason: it is easy to imagine that the same shape can have different intentional properties in different systems, just as the same sound can mean different things in different languages. The intentional

properties of a representation must depend somehow on its context, on the system it is part of.

As some influential arguments from the second half of the 20th century show, the effects of this context-dependence on the required supervenience base of intentionality are very pervasive. Hilary Putnam (1975) discusses a scenario involving a planet, Twin Earth, that is identical to our world with respect to all its physical properties except for one: instead of water, its oceans and rivers are filled with a different chemical, indistinguishable from water in all observable respects but with a different chemical composition: XYZ instead of H_2O. Putnam points out that, although the denizens of twin earth are identical to us in all physical respects (except that they have XYZ in their bodies instead of H_2O), their brains are wired up the same way, and they speak languages that sound just like ours, it seems undeniable that their word "water" refers, not to water, but to XYZ, and instead of thoughts about water, they think thoughts about XYZ.

This and related arguments (Kripke 1980; Burge 1979) have convinced many that the determinants of linguistic meaning and mental content must include factors outside of the mind of the thinker/speaker. As we will see in the next chapter, teleosemantics too belongs to this *externalist* tradition in the philosophy of intentionality. Specifically, it maintains that the content of a representation depends in a particular way on its *history* and on the entities and circumstances that have played explanatory roles in that history.

3) *The causal properties of mental representations reflect their intentional properties*. One of the most powerful ideas of 20th century philosophy of mind is that the intentional idiom—including, centrally, the idiom of propositional attitudes like beliefs and desires—is a sort of *theory* for explaining and predicting human behavior (e.g. Sellars 1997, 102-07). According to this idea, when we talk about beliefs, desires etc., we are talking about posited theoretical entities with specific causal powers and relations, capable of explaining how perception is mapped onto behavior.

What, then, are the causal powers of intentional states that allow attributions of them to predict and explain behavior? The idea here is that a well-functioning mind is set up in such a way that the causal relations between representations, to some degree short of perfection, reflect modal relations between their contents (e.g. inconsistency and entailment), so that, for instance, if the content *P* of one belief entails the content *Q* of another, the first belief will tend to cause a tokening of the second; if the contents of two beliefs are inconsistent, the cognitive system will tend to ensure that they are not held at the same time; and so on. In addition, the causal relations between representations and *extra-mental* reality will also tend to reflect *rational* (reason-giving) relations defined on the former's content, so that the belief that *P* will tend to cause behavior that is appropriate given *P*; evidence that *P* will tend to cause the belief that *P*; and so on.

The view that causal relations of mental representations *reflect* their modal and rational relations must be distinguished from a stronger view: that the causal properties of a mental state *determine* its intentional properties in such a way as to make this reflection a matter of metaphysical necessity. The latter view is a form of *functionalism*, which, applied to intentional states, constitutes a distinct meta-semantic view. For instance, according to the "analytic functionalism" defended by David Lewis (1974, 1994), it is constitutive of intentional mental states that they interact causally in such a way as to render the subject maximally rational and consistent.

Functionalism, in its various guises, constitutes the main contender within meta-semantics to the tradition to which teleosemantics belongs (which tries to analyze intentionality in terms of causal and historical relations to real world features, rather than in terms of mind-internal causal roles). It would take us too far afield to subject functionalism to an exhaustive critique (for review, see Ryder 2009, 264–70). Here, I will mention one reason to be skeptical of functionalism that bears on the current discussion.

The functionalist has to explain how the content of a thought is *determined by* its causal role. She must therefore find a way to associate each possible content with a unique causal profile. One way is to assign causal roles to contents in such a way that the causal interactions between representations reflect, in the above sense, the modal and rational relations of their contents. This approach, however, runs the risk of generating *too much* rationality. Sometimes people are irrational. But if the content-determining causal roles of representations were those that reflected the modal and rational relations of the contents thus determined, a representation could never stand in a causal relation that *didn't* respect its modal and rational relations. In other words, inconsistency and irrationality would be impossible.

Functionalism thus needs some other way of singling out the content-determining causal relations. One possibility is to identify them with the causal relations that a representation stands in under *normal* conditions. Pondering the nature of normal conditions may bring us away from functionalism and towards the sort of externalist and historical theory of which teleosemantics is an example.

If the intentional properties of mental states are not *determined* by their causal properties, then the claim that mental causal processes *reflect* these intentional properties must amount to the claim that mental systems are, as a matter of fact, "wired up" so as to meet this criterion. In other words, even if content is not determined by causal role, it can still be a robust empirical generalization that the causal effects a representation has on the cognitive system it is part of tends to reflect the modal and rational relations of its content. This accommodates the possibility of irrationality, because "tends to" admits of exceptions.

Still, it would be strange if the metaphysical basis of intentionality had *nothing* to do with the disposition of mental states to enter into causal rela-

tions reflecting their intentional relations, i.e., if it were a wholly contingent matter that cognitive systems sometimes respected those intentional relations in operating on the mental states. It would be, as it were, as if Nature had found mental states already lying around, complete with contents, and decided, on a whim, to put them in causal relations that reflected the modal and rational relations between those contents. Couldn't Nature then just as well have arranged their causal relations some other way, perhaps substituting every other belief in the causal network for a contradictory belief? What would have been lost, expect for a certain aesthetically pleasing symmetry?

Another, less fanciful way of putting the same point is this: if the intentional properties of mental representations have no *essential* relation to their causal properties, then we seem to get no explanatory mileage out of mentioning those intentional properties. They cannot figure in scientific generalizations about the causation of human behavior. If representations are supposed to supply us with causal explanations of human behavior, we would do just as well to simply point to their causal properties directly.

These types of considerations have led some to outright deny the scientific value of intentional notions (e.g. Stich 1983; Chomsky 1995). It lies very close at hand to deny that intentional notions have any interest at all: if science has no use for intentional notions, then maybe, ultimately, *we* don't either. That is a very radical conclusion. We use intentional notions all the time, and we sure *seem* to be using them to explain and predict various aspects of human behavior. If this ubiquitous practice turned out to rest on some kind of confusion, it would deeply affect not only our psychology, but all those of our intellectual and cognitive practices that are concerned with our mutual understanding of each other. In Fodor's words, "if we're that wrong about the mind, then that's the wrongest we've ever been about anything" (Fodor 1987, xii).

We probably don't want to accept this conclusion, then. But the challenge is real. If intentional categories do not individuate mental states according to their present causal powers, then what is their cognitive value? Why do we spend so much time and effort keeping track of them? I said on p. 21 above that an analysis of intentionality that renders this fact mysterious is, *prima facie*, a bad analysis. We should be able to explain the contribution intentional notions make to our understanding of the world and of ourselves.

Teleosemantics has what I believe to be a very good answer to this question. Teleosemantics rejects the functionalist claim that the content of a representation is determined by its present causal role. However, it doesn't make the relation between present causal role and intentional content entirely contingent, because it ties the intentional content of a representation to the causal interactions of *past* representations related to the present token by a certain kind of ancestor-descendant relation, insofar as those causal interactions have contributed to the past persistence and proliferation of the systems producing the representations. And it ties the present causal properties of a

representation to the causal properties of its ancestors, which have helped determine the properties of the present token via a process of *reproduction* or *replication* that preserves certain physical, hence causal, features.

All of this will be explained in greater detail in the next chapter. Here, let me just briefly indicate what it means for the present conundrum. It means that attributing intentional properties to a mental state is tantamount to saying something about a) the conditions that have historically constituted *success* for a state of that kind, and/or b) the conditions that have been conducive to that type of success. Here, "success" is to be understood teleologically, i.e. in terms of outcomes that have contributed to past persistence and proliferation. Now, this type of historical reference does not directly entail anything about the causal powers of a present mental state. But by inductive inference, it allows us to draw fallible but useful inferences about the state's likely future causal interactions (cf. Millikan 1986, 2007).

In particular, teleosemantics identifies the descriptive content (the truth- or correctness-conditions) of a representation with the conditions that figures in explanations of the persistence and proliferation of the representation's ancestors. This means that, knowing the content of a belief, we can draw some fallible conclusions about the conditions under which it is likely to be successful and the conditions under which it is not (cf. Godfrey-Smith 1994, 1996; Shea 2007). This falls far short of an exhaustive mechanistic understanding of the cognitive system and its interactions with the environment. But that kind of exhaustive mechanistic understanding is typically unavailable to us anyway, whereas the rough indication of likely outcomes supplied by knowledge of teleological properties is not.

Moreover, the types of likely outcomes fallibly predicted by teleological notions are of a kind that is of special interest to us. This is because, although the "success" involved is defined in terms of contributions to past persistence, it is a type of success that most of us are disposed to have a great deal of practical and emotional investment in, for the simple reason that we are its causal products and have evolved to pursue it.

In this section, I have discussed some of the general considerations that bear on the project of naturalizing intentionality. In the next section, I will discuss indicator semantics, a family of views that comprises the closest historical predecessors to teleosemantics. I hope this discussion will allow the reader to better appreciate the motivations behind the central claims of teleosemantics.

1.4. Indicator Semantics

A representation, we learned in the preceding sections, is a state or event that bears a particular relationship to a set of conditions on the world, picked out by the representation's content. The family of views known as "indicator

semantics" takes that relationship to be that of *carrying information about*. Carrying information, on these views, is taken to be a non-intentional, nomological relationship, one that is instantiated between two states when (roughly) both states obtain and there is some lawlike relationship between states of the first type and states of the second that explains why they both obtain.[19]

Consider a mercury thermometer. It has a number of possible states, each consisting in a different height of its mercury column.[20] There are laws or lawlike principles (pertaining to the relationship between temperature and the density of mercury) that correlates the height of the mercury column with the ambient temperature. So, provided the thermometer functions normally, the state of the thermometer carries information about the ambient temperature. It also seems natural to say that the state of the thermometer *represents* the ambient temperature. The basic idea of indicator semantics is to analyze the latter state of affairs in terms of the former: the state of the thermometer represents the ambient temperature *in virtue of* carrying information about it.

Representations often *do* seem to carry information about what they represent, and "carries information about" paraphrases one common sense of the word "means" (compare: "those dark clouds mean that it will rain soon") (Grice 1957), so many authors have found the basic idea of indicator semantics compelling (e.g. Stampe 1977; Dretske 1981, 1986, 1988; Fodor 1987) Here, we will not be concerned with the details of this significant literature,

[19] I have kept this formulation intentionally vague, not to exclude any of the notions of information employed by indicator theories. We could also say, what I take to be equivalent, that the obtaining of a state of one type should nomologically necessitate, or else raise the probability of, the obtaining of the state of the other. Exactly how strong the nomological connection must be has been the source of dispute. Dretske (1981, 65) requires necessitation, whereas Millikan's notion of information is significantly more liberal (e.g. Millikan 2004, chap. 3). See (Ryder 2009, 254) for a review.

Typically, the requisite lawlike relationship is identified with *causation*, so that the representation must be caused by the *representandum*. But the account of information relevant to the needs of meta-semantics could easily be extended to allow states to carry information about each other even if neither is the cause of the other, as long as they are connected by some other lawlike relationship. For instance, they could have a common cause or be connected by non-causal nomological necessity (the way the mass of an object correlates with its acceleration under a given constant force) (this is a feature of the account of information favored by Millikan 2017, 139–40).

One could protest against this more liberal conception that the representation-relation is an *asymmetric* relation: my belief that *P* represents that *P*, but the fact that *P* doesn't represent my belief that *P*. Hence, we need the information-relation to be similarly asymmetric if we are to have any hope of analyzing the representation-relation in terms of it. However, the indicator theory cannot hope to capture all features of representation in terms of information-carrying anyway. If it tried, it would wildly overgenerate representations: a cloud causes rain, but the rain does not represent the cloud. So, liberal conception of information or not, indicator semantics needs to put additional constraints on representationhood. More on this below.

[20] To be more precise, the semantically significant state-space of the thermometer has *two* parameters: the height of the mercury column, and the location of the thermometer. The height of the mercury column in a thermometer located at a place at a time represents that the ambient temperature *at that place and time* is so-and-so.

only discuss the (well-known) limitations of the basic idea and how they motivate further developments.

The information-relation is factive. If A carries information about B, then B *ipso facto* obtains. However, the fact that A has the intentional content that B obtains in no way guarantees that B obtains. Representations, characteristically, can misrepresent the world, and indicator semantics has trouble accounting for misrepresentation. Suppose a 5 cm high mercury column in thermometer *T* represents the ambient temperature as being 10°C. On one occasion, this height may indeed carry the information that the ambient temperature is 10°C. But suppose that the thermometer gets a small hole so that some of the mercury leaks out. On a later occasion, the same height no longer carries the information that the temperature is 10°C. But intuitively, the height still *represents* the ambient temperature as 10°C. This just happens to be a *mis*representation of the actual temperature.

One initially promising way to deal with this problem is to allow a state A to carry the content *that state-type B is instantiated*, even if there is no corresponding B for A to carry information about, as long as A bears a suitable relation to states that *do* carry information about Bs. Since a state carries information about another only in virtue of a lawlike relation that holds between some types that the states belong to, we would seem to get the suitable relation for free: A has the content *that state-type B is instantiated* because A belongs to the As and there is a lawlike relation between the As and the Bs in virtue of which some As sometimes carry information about some Bs. If this is to work, the lawlike relation in question cannot be one that *necessitates* that an A will be accompanied by a B, or an A could, again, never misrepresent (cf. n. 19, p. 35). It has to be one that merely underwrites a non-accidental correlation between As and Bs, in virtue of which an instantiation of type A raises the probability that type B will also be instantiated.

By itself, however, the idea that representation reduces to correlation faces a number of difficulties, most infamous of which is the *disjunction problem* (Fodor 1984). If there is an imperfect correlation between the As and the Bs, then there will always be some disjunctive state-type C such that there is a stronger correlation between the As and the Cs. For instance, if a mercury column height of 5 cm correlates with a temperature of 10°C, it will correlate more strongly with a temperature of 10°C *or* a temperature of 20°C and the presence of a hole in the thermometer, etc. The strongest correlation will be between the As and a vast disjunction of circumstances that jointly necessitate that an A is tokened. The problem, then, is to determine *which* of these correlations is the one that picks out the representational content of the As.

Fodor notes that

> [A]ll the standard attempts to solve the disjunction problem exhibit a certain family resemblance. The basic idea is to distinguish between two types of situations,

> such that lawful covariation determines meaning in one type of situation but not in the other. (Fodor 1990, 60)

In the first type of situation, which we can call the "normal conditions," the tokening of an A guarantees the tokening of a B. Misrepresentation—an A without a B—occurs when conditions are abnormal.

The problem then becomes that of specifying the normal conditions in a non-circular way. This, as Fodor points out (*ibid.*), is far from trivial. As we shall see in the next chapter, teleosemantics solves this problem by identifying "normal" conditions with (uppercase) Normal conditions, defined as conditions that have contributed to explaining past episodes of successful operation among ancestors of the current representational system.[21]

A second, related problem for indicator semantics concerns the fact that, even with a solution to the disjunction problem, a representation under normal conditions is likely to bear information about a number of different things: about the cause of the representation but also about the cause of the cause and about everything that correlates, for lawlike reasons, with that cause. A normally functioning thermometer carries information about the ambient temperature, but also about the mercury levels in other, nearby thermometers. Indicator semantics needs a way to pick out the content-determining *relatum*. We can call this "the specification problem."

Teleosemantics purports to solve the specification problem by, again, appealing to the technical notion of Normal conditions. Normal conditions are not just any conditions that have normally obtained in the past, but conditions that help *explain* the success of past episodes of representation. Consideration of our thermometer example can illustrate this idea. If we define "success" for the thermometer as permitting human users to learn about the ambient temperature and take appropriate steps in response (this, after all, is why we have thermometers), then it seems clear that past successes are explained by the fact that thermometers have carried information about the *temperature*, while the fact that it has also carried information about the state of other thermometers is irrelevant to their success.

I would like to pause here to reflect briefly on the nature of the specification problem. The problem arises because a prospective content determination principle fails to pick out a *unique* content for each representation. A number of different candidate contents equally meet the condition given by the principle. This is the general form of an *indeterminacy problem*, about which we will hear more in section 2.5, and again in chapters 3 and 4. Indeterminacy problems are recurring problems for naturalistic theories of intentionality, and teleosemantics is no exception from this pattern.

[21] For discussion of some other strategies for solving the disjunction problem, see (Ryder 2009, 255–64).

To solve an indeterminacy problem, we can add further conditions to the content-determination principle and hope to narrow the eligible content candidates down to one. Most likely, however, there will be several different narrowed-down content determination principles to choose from. Which one do we pick? We should of course pick the one that gives the *right* content for each representation, and for the philosopher, it is natural to suppose that our guide to finding the right one must be pre-theoretic intuition.

At this point, a type of skeptical worry easily insinuates itself into the discussion. Suppose we have a number of prospective content-determination principles D_1, D_2, D_3, and so on, each of which picks out a unique content candidate for each representation. Suppose we find that D_1 is the one that best accords with pre-theoretic intuition, and we decide that the *content*$_1$ of a representation should be identified with its content *simpliciter*. Yet, the skeptic says, even if this is correct, what is it about the *content*$_2$, the *content*$_3$, etc. of a representation that make them less interesting, less worthy of consideration and theoretical attention, than its *content*$_1$? Even if, when we talk about the content of a representation, the thing we talk about is the representation's *content*$_1$, is it anything other than parochialism and force of habit that makes us attend to this particular feature of a representation? Perhaps what we *should* talk about is the *content*$_2$ of representations, and we should revise our intentional language accordingly, so that the word "content" referred (referred$_2$?) to *content*$_2$.[22]

Meeting this skeptical challenge is another reason why we want an account of intentionality that doesn't just capture our intuitions but also accounts for the cognitive value of intentional notions. Indicator semantics accomplishes this to a degree. It doesn't just capture (some of) our intuitions: it also provides a (partial) explanation of *why*, if the theory is a correct account of content, content-notions have cognitive value, i.e., why content is an interesting phenomenon worth keeping track of. It is interesting because the content of a representation says something about the states of affairs the representation can be expected to carry information about, and information is interesting because it allows us to learn about the world and navigate it successfully. We would hope that a solution to the specification problem would

[22] A point very much like this, although regarding the causal theory of *reference* (which is something other than indicator semantics, though it gives rise to similar problems), has been made by Stephen Stich (1993, 110 ff.) Stich makes the argument that for every causal relation the causal theorist proposes as being our usual reference-relation, there will be a number of other nearby relations REFERENCE*, REFERENCE** etc. that differ from the proposed relation only in minute respects. While the truth of a sentence depends on what its component terms refer to, the TRUTH* of a sentence depends on what its component terms REFER* to, and there is no obvious reason why we should value truth over TRUTH*, so even if the causal theorist has correctly identified our reference-relation, she has given us no reason to care about it, and in fact, her theory illustrates that we might *not* have any such reason. Stich makes his argument in the course of a skeptical attack on the value of truth, but we can also read it as indicating the need, in giving a theory of reference, to ensure that this theory also explains *why* we should value truth over TRUTH*.

further explain why content (as opposed to *content*$_2$ or *content*$_3$) is worth keeping track of. And as we have already seen, the teleosemantic solution to the specification problem meets this desideratum, by telling us that the content of a representation picks out conditions that will tend to make the causal effects of the representation successful.

A third, less widely discussed problem with the basic indicator theory is that, while it gives us the content of representations, it doesn't tell us which states are representations. Any state-type that is part of the causal order will correlate with other state-types, but not every state-type is a representation-type. The clouds presaging rain are not representations of rain, and neither is the rain a representation of clouds. An informational analysis of content must therefore be supplemented with conditions on representationhood, or it will wildly overgenerate representations.

An appeal to teleology seems like a promising approach here as well. But it doesn't suffice to say that representations are entities that *are supposed to* carry information. As Millikan points out (1989a, 282), that too would overgenerate representations. Many biological systems are designed to vary reliably with conditions in the environment, i.e. to carry information, but are not plausibly thought of as representations. For instance, the size of our pupils are designed to vary with the intensity of incoming light, but they do not, presumably, represent the intensity of incoming light.

Millikan purports to solve this added problem by introducing the idea of a *consumer*, a system that relies on another system, the *producer*, in order to perform its (teleological) function. A representation, then, is a functional state that mediates causally between a producer and a consumer. We will return to this notion, too, in the next chapter.

This has been a short introduction to the basic ideas of indicator semantics and its various pitfalls. Teleosemantics can be seen as an attempt to solve these problems. In the course of this attempt, a theory emerges that bears very little resemblance to the simple thermometer model described above. Nevertheless, teleosemantics retains the fundamental idea that representations are in the business of carrying information about their represented. In the next chapter, we will see how teleosemantics uses the teleological notions of function and Normal conditions to overcome the disjunction and specification problems of indicator semantics.

It is worth acknowledging the fact that though indicator semantics has some initial plausibility as a theory of simple, hard-wired representations, it simply seems like a non-starter for explaining the sort of complex, productive representational systems that we find, for instance, in human thought and language. It is hard to see what the requisite correlations and type-identities would even be when we are dealing with the complex compositionality of human language, where a representation can be the first of its kind that has ever been tokened in the history of the world, yet is immediately comprehensible to any competent interpreter. Much of the interest that

Millikanian teleosemantics has for me personally derives from its ambitious attempt to solve this problem. We will return to it in the next chapter.

1.5. Summary and Conclusion

In this chapter, I have introduced the task of providing a naturalistic analysis of intentionality. I defined intentionality in terms of the notions of representation and content, discussed some of the commitments uniting attempts to naturalize intentionality and some of the problems facing that project, and gave an overview of indicator semantics.

The chapter leaves us with some questions and problems:

1) How does teleosemantics account for the possibility of false representations and of misrepresentation?
2) How does teleosemantics solve the specification problems of indicator semantics?
3) What is the explanatory role or cognitive value of intentional notions, which justifies their role in everyday as well as scientific thinking?
4) How should we explicate the intuitive sense in which the content of a representation provides a *norm* for its tokening?
5) How does teleosemantics account for complex, productive systems of representation, such as human thought and language?

Let us keep these questions in mind as we turn, in the next chapter, to acquaint ourselves with teleosemantics.

2. Teleosemantics

In this chapter, I will introduce teleosemantics. *Teleosemantics* is so-called because it purports to analyze the intentional properties of representations—their "semantics"—in terms of their *teleological* properties, i.e., their functions or purposes and the way those functions are supposed to be performed. Millikan often uses the alternative term "biosemantics" for her version of the view, reflecting the fact that the understanding of teleology operative in teleosemantics is biological: the teleology in question is the kind possessed by biological traits. Biological traits have been shaped by natural selection. According to the *etiological theory of function* it is *in virtue of* its selection history that a biological trait possesses the teleological properties that it does possess. Teleosemantics is the result of combining the etiological theory of function with an analysis of intentional properties in terms of functions.[23]

Versions of teleosemantics have been defended by, among others, David Papineau (1984, 1993), Carolyn Price (2001), Karen Neander (2017), and Nicholas Shea (2018). As advertised in the introduction (p. 13), I will rely almost exclusively on Millikan's version of the theory (towards the end of the chapter we will hear some dissenting voices). The term "teleosemantics" will be used to refer to Millikanian teleosemantics unless otherwise noted.

This chapter is an attempt to introduce (Millikanian) teleosemantics in sufficient detail that the reader will be able to appreciate and, to some extent, assess its main points. Limitations of space force me to gloss over many interesting details and points of controversy. I have attempted to indicate problematic points in footnotes and provide references for the interested reader, but I have largely abstained from addressing common objections except insofar as these help illuminate some important claim or argument.

Rather than being a thorough critical evaluation of teleosemantics, then, the chapter is mainly intended to prepare the ground for the discussion to follow. Doing so will require two things. The first is to show how teleose-

[23] These two views—the etiological theory of function and the analysis of intentionality in terms of function—are logically independent. Seizing on this, and in the light of some purported problems for the etiological theory (see n. 26, p. 44; n. 28, p. 47), Bence Nanay (2014) has proposed a revision of teleosemantics that replaces the etiological theory with an alternative account of function. However, as I hope the coming discussion will show, the etiological theory is such an integral part of teleosemantics, informing its implications and conclusions on all levels of analysis, that any attempt to replace it yields a theory of a very different kind, one that it strikes me as misleading to market under the same rubric.

mantics answers the five questions we were left with at the end of the last chapter (p. 40). The second is to explain how teleosemantics accounts for *discourse*. If we want a teleosemantically informed theory of moral thought and discourse, we clearly cannot remain content with a version of the theory that restricts its scope to mental representations. But discourse, as can be expected, presents its own range of problems when one tries to subsume it under a general theory of intentionality. One such problem is the fact of productivity referred to in the fifth of the aforementioned questions. Another, specific to teleosemantics, is the plain fact that discourse is a cultural phenomenon which may at first sight appear difficult to subsume under a biological model of explanation. How teleosemantics purports to solve these problems will be the topic of sections 2.3 and 2.4.

The chapter is structured as follows. In section 2.1 I introduce the basic concepts of Millikanian teleosemantics and show how these combine into a content-determination principle that allows us to answer questions 1-3 above. In Section 2.2 I discuss Millikan's theory of concepts and show how teleosemantics can be applicable to person-level conceptual thought, i.e. propositional attitudes. In section 2.3 I discuss Millikan's theory of discourse and explain how discursive representations can possess the sort of ancestry necessary for assigning content to them in accordance with the content-determination principle given in 2.1. The discussion of discourse continues in section 2.4, where I explain how Millikan proposes to account for the productivity and compositionality of human language. Finally, in section 2.5, I address some remaining worries about the ability of Millikanian teleosemantics to avoid indeterminacy problems.

In presenting Millikan's views, I will refer freely across her collected *oeuvre* from *Language, Thought and Other Biological Categories* (1984) to *Beyond Concepts* (2017), on the understanding that the theory remains essentially the same across that period. There have been some revisions, of course, and when those become relevant I will point them out in the text.

2.1. Basics of Teleosemantics

Teleosemantics tries to account for intentionality in terms of the functions of representations and of the systems that produce them. The relevant sense of "function" is that of *teleological* functions, i.e., the sort of functions that tools and biological traits possess.[24] Millikan calls these "proper functions." We will follow her terminology, interchangeably with the simple "function." We will also use other teleological terms, like "purpose," "goal," "job" etc., to talk about proper functions. Teleosemantics understands these notions

[24] Hence, this use contrasts with that of the "functionalist" theory in philosophy of mind, where functions are identified with causal roles. Cf. p. 32.

according to the *etiological theory of function*, so-called because it accounts for the function of an entity in terms of its etiology, i.e., the causal history that explains its existence.

Biological traits have been naturally selected for because they have helped accomplish certain tasks that contributed to the past persistence and proliferation of the organisms whose traits they were, and hence to their own past persistence and proliferation. According to the etiological theory, when we speak of the function of a biological trait—when we say, for instance, that the heart's function is to pump blood—it is usually these selected effects we have in mind. By generalizing the notion of selection to apply not only to biological traits but also to artefacts, linguistic devices and so on, the etiological theory purports to provide a completely general account of teleology.

Chapters 1 and 2 of Millikan's main work *Language, Thought and Other Biological Categories* (Millikan 1984; henceforth, *LTOBC*) lay out her version of the etiological theory. The details are involved, so what follows is only a brief overview. In order to possess a function (or, as I will also say, in order to *be function-bearing*), an entity must have an *ancestry* or, equivalently, belong to a *family*. Millikan in *LTOBC* calls such families "reproductively established families" (LTOBC, 23–25). I will write "RE-families" for short. Within an RE-family, ancestors are related to descendants via particular types of causal relations. In the most fundamental cases, this is simply *reproduction*, defined as any process whereby token entities are produced that resemble other, pre-existing token entities in determinate respects (their "reproductively established properties"), and resemble them *because* those pre-existing tokens have those properties (LTOBC, 19–23). The replication of a DNA molecule is thus an example of reproduction, and so is the child's repetition of a word she hears her parents say. An RE-family whose members are produced by reproduction is what Millikan calls a *first-order* RE-family (LTOBC, 23).

There are also *higher-order* reproductively established families, whose members are not reproductions of one another. Instead, they bear a more indirect ancestor-descendant relationship in virtue of the fact that they are each the product of the proper function of members of some *other* RE-family or families and resemble each other for that reason (LTOBC, 24). It is worth noting that one and the same entity can belong to several different RE-families, having inherited different properties from different ancestors (LTOBC, 20–21). This will become important when we discuss language, since a linguistic representation is a prime example of an entity with many different, crisscrossing ancestries.

When something has an ancestry, it can also have a function. Very roughly, a function of an entity (a *direct* proper function, in the terminology of *LTOBC*) is something ancestral entities have done that helps explain, via a process of selection, the fact that the current entity exists. For instance, a function of the heart is to pump blood. Pumping blood is something earlier

members of the heart-family have done that helps explain why present hearts exist. Those earlier hearts have also done other things, like producing a thumping sound. But the thumping sound presumably does not help explain why present hearts exist, and thus, it is not a proper function of the heart.[25] Since an entity's ancestors may have contributed to its past success in several different ways, an entity can have several different functions. For details, see (LTOBC, 25–31).[26]

In addition, Millikan requires that a function be something the ancestral entities have done that is explained by their reproductively established properties. This is to ensure that idiosyncratic features of a trait token, which may have contributed to the fitness of the lineage but are not reproduced, will not be functions of descendant tokens.

Function is not the only teleological notion important to understanding the teleosemantic analysis of intentionality, however. Just as important is the notion of *Normalcy* or of something's being Normal (cf. p. 37). Like proper function, Normalcy is defined etiologically. If a proper function is an activity that ancestors have performed that helps explain the existence of their descendants, Normalcy (always capitalized) describes the shape this explanation has actually, historically, taken: *how* the ancestors have managed to contribute to bringing their descendants into being, and what other factors have contributed to this process. Millikan talks in particular of Normal explanations and Normal conditions:

> A Normal explanation is a preponderant explanation for those historical cases where a proper function was performed. Similarly, Normal conditions to which a Normal explanation makes reference are preponderant explanatory conditions under which that function has historically been performed (LTOBC, 34).

---

[25] Note that I am shifting between talking about *token* hearts and the heart *kind*. Most of the time, this vacillation is completely innocuous, but it should be kept in mind that strictly speaking, a proper function is a property that belongs to *tokens* rather than types. Since the notion of proper function is historical, ancestral members of a family can lack a function that descendants possess.

[26] One type of criticism against the etiological theory is based on the fact that it precludes the possibility that *novel* traits without ancestry (e.g. mutations) can have proper functions. For a version of this objection, see (Brunnander 2011). This type of criticism is closely related to the so-called Swampman objection (see n. 28, p. 47).

Another type of criticism has been advanced by Paul Sheldon Davies (2000; see also Nanay 2014), who argues that the etiological theory fails to account for the possibility of a *malfunctioning* trait, i.e. one that is incapable of performing a function that it nevertheless shares with other, functioning traits. Davies derives this conclusion from two assumptions: 1) that the etiological theory assigns functions in the first instance to trait *types*, whereas trait tokens have functions derivatively in virtue of belonging to types; and 2) that types are individuated by the ability of their members to perform a function. It follows from (2) that a malfunctioning trait cannot belong to the same type as a functioning trait, and so, by (1), that the two cannot share a function. As far as I can tell, this objection fails to find purchase on Millikan's version of the etiological theory, which assigns functions directly to tokens, on the basis of their ancestry (cf. n. 25 above).

To see the relation between proper function and Normal conditions, consider my desktop lamp. One proper function of the lamp is to emit light. Its Normal way of performing that function is by way of an electric current passing through the gas in the fluorescent lightbulb. An explanation of the fact that the lamp emits light (performs its proper function) which makes reference to the electric current, the physical reaction in the gas, and so on, is a Normal explanation of the proper function's performance. If the lamp catches on fire it will emit light, and so, in a way, fulfill its function, but it will not do so in a Normal way and therefore not according to a Normal explanation.

The notion of Normalcy makes precise an intuitive distinction between two ways in which a system can perform its function, viz., more or less "by accident" or "by a fluke." That a lamp can emit light by catching on fire is an accident of its design. The same goes for evolved systems: that the human circulatory system can operate with an Aqua-lung is an "accident" of natural selection (though not an accident of the design of the Aqua-lung, of course) (cf. LTOBC, 33). If it turned out, *per mirabile*, that there were some chemical substance on another planet which could substitute for oxygen in our circulatory system, then that would constitute an even greater accident.[27]

Note that the distinction between functions and Normal conditions is not absolute. Systems often perform their functions *by* first performing other, more immediate functions. The heart oxygenizes the tissues *by* pumping blood. Here, performance of the latter function is part of the Normal explanation for performance of the former.

How, then, does teleosemantics account for the intentional properties of a representation in terms of its teleology, i.e., its functions and the Normal conditions for their successful performance? Not everything that has a function is a representation, of course. The heart has a function—to pump blood—but is presumably no representation. One could be forgiven for thinking that what sets representations apart from other functional systems is the *kind* of function they have. It would be uninformative to learn that a representation's function is to "represent"—representation is what we are trying to analyze—but perhaps a representation's function is to carry information about the world (in something like the sense discussed in section 1.4)? Milli-

[27] Normal conditions, it must be emphasized, are not statistically normal or common. They can be very rare. The point is that they are conditions under which a trait has historically managed to perform its function (Millikan 1984, 34).

In a review of Millikan's *Beyond Concepts* (2017), Paul Griffiths suggests that the notion of Normalcy "points to a significant and largely unexplored problem for Millikan's approach. The idea of the Normal assumes that if a trait has evolved through natural selection, then there is a reasonably unified causal explanation in which properties of the trait interact with the environment to explain its success. But this cannot be assumed" (Griffiths 2019, 2). It is doubtlessly true that Millikan does assume "a reasonably unified causal explanation" for historical success, at least on more proximate levels of explanation (Millikan 1984, 33–34), although to what extent this is essential to the account, I cannot tell. It may be worthwhile to further explore Griffiths' worry.

kan, however, is adamant that a function is always something that an entity *does*, an intervention it makes in the causal order (e.g. Millikan 2004, 67), and only occasionally do things *effect* the things they carry information about. Many times, things carry information about their causes or about things that correlate with their causes.

In fact, the *function* of a representation, according to Millikan, can be anything (LTOBC, 71). What makes something a representation is not the type of function it has, but the relations it bears to certain other function-bearing systems and the functions of *those* systems. A representation is something that is produced by a *producer* and interacts with a *consumer* (or interpreter), and whose function involves helping the consumer to perform *its* function, by adapting it to external conditions (Millikan 1989a, 285–86).

A consumer has what Millikan calls a *relational* function. What it is supposed to do varies depending on external circumstances (LTOBC, 39). The notion of a relational function is very broad, and encompasses everything from the human pupil (it has the function to expand or contract depending on the intensity of incoming light) and the skin of a chameleon (it has the function to change color with the background) to the human cognitive system (its function is, in very abstract terms, to produce behavior that is adaptive given the circumstances). Such systems have persisted and proliferated because their activities have born a certain relation to features of the environment. In the case of the chameleon's skin, the relation is that of bearing the same color as its background. In the case of the human cognitive system, it is a much more complex and abstract relation.

In addition to this relational function, a system also has what we may call an *invariant* function (LTOBC, 41): to permit the eye to see while protecting it from strong illumination, to make the chameleon harder to detect, to produce adaptive behavior. It is *by* performing its relational function that it has, at least Normally, managed to perform its invariant function. This means that, for any given activity that the system can engage in, the way the world has to be for that activity to count as performance of the system's *relational* function is a Normal condition for performance of its *invariant* function.

The consumer, then, is a system with a relational function. So how does it Normally manage to perform it? That is where the producer and the representations come in. A producer is a system that is supposed to *help* the consumer perform *its* relational function, by producing representations that interact causally with the consumer. By interacting with the consumer, the representation "adapts" the consumer (LTOBC, 43–45) and makes it behave in a specific way. To cause the consumer to act in such-and-such a way is the representation's own function, what Millikan calls its *adapted* function (it is, as it were, adapted to the state of affairs it represents; cf. *ibid*.), and if the producer works Normally, it has produced a representation that "fits" external circumstances in the sense that the consumer response it causes is

the right one for the circumstances, the one that constitutes performance of the consumer's relational function.

In essence, this is what it means for a representation to *correctly represent* the world, according to teleosemantics: that circumstances are as they're *supposed to be* when the consumer responds to the representation as it's supposed to do. In Millikanese, the representation is correct when Normal conditions obtain for the consumer's proper performance, as adapted to the representation (LTOBC, 100). A common paraphrase of this idea says that a representation represents the *success-conditions* for the consumer response, and this is a fair approximation, but it is important to keep in mind that, as I stressed above, not every success is a Normal success. Success can be obtained through a fluke or by accident.

Now I ca state, to a first approximation (there will be some complications below), the teleosemantic content-determination principle: The content of a representation is the Normal conditions for successful performance of the consumer's proper function, when adapted to the representation.

For an example, let us return to the waggle dance from p. 25. The producer and consumer of waggle dances are devices in the nervous systems of honey bees—the producer in the bee that performs the dance, the consumer in the observing bee—and the representation is, of course, the dance itself. The producer's relational function is to produce a dance that bears a specific geometric relationship to a nearby location with nectar. But dance-producers have ultimately persisted only because they have sometimes led other bees to bring home nectar. So the producer must "cooperate" with its consumer. The consumer's relational function is to make the bee fly to a location where there is nectar. In order to do so, it relies on the dance produced by the producer, and so, for the whole system to perform what is ultimately its invariant function, to bring home nectar to the hive, the place where the dance Normally causes the consumer to bring the bee must be a place with nectar. If there is nectar at the place, the dance has described the world correctly.

Note that what matters for determining the correctness-conditions of a representation is not what it *actually* causes the consumer to do, but what it is *supposed to* cause it to do, according to its own adapted function. So a representation cannot be just anything that causally impacts the consumer, but must itself be a function-bearing entity belonging to an RE-family, with an ancestry consisting of other representations that have impacted consumers in similar way, thus giving rise to fitness-enhancing outcomes.[28]

[28] The fact that a representation must have an ancestry forms the basis of the most common and persistent objection to teleosemantics, the so-called *Swampman* objection. First introduced in a paper by Donald Davidson (1987), Swampman is a perfect duplicate of a human being who forms through random chance in a steamy swamp. Since Swampman has no evolutionary history, teleosemantics entails that he lacks intentional states. This strikes many as an unacceptable consequence.

However, not every representation will have ancestors of *exactly the same kind* as itself. It is possible that tomorrow, a particular waggle dance will be danced whose duration-angle combination has no precedent in the history of waggle dances. How, then, are functions and correctness-conditions assigned to novel representations? How, in other words, can representational systems be productive?

It helps to consider that in general, what explains the past persistence of some function-bearing entity's ancestors is not just that the ancestors have performed certain activities, but that they have performed those activities *under certain circumstances*. What explains past persistence is *these-activities-under-these-circumstances* or, put differently, that a certain *relation* has obtained between activities and circumstances. For a given activity, the circumstances that must obtain for the two to bear the right relation to one another, the past-persistence-explaining relation, are Normal conditions for that activity. Producing representations is an activity of the producer, and the producer's job is to ensure that a certain relation obtains between the representation it produces and the circumstances that obtain, the same relation that explains past persistence. We can represent the relation in question as a mathematical function or mapping, defined over the significant features of the representation, that maps every possible representation that the producer can produce to a possible state of the world or set of possible worlds. For the representation to be Normally successful, this state must obtain. The actual world must be in the set picked out.

This relation can hold even for completely novel representations. A waggle dance can relate, according to the semantic rules for waggle dances, to the location of nearby nectar even if no similar waggle dance has ever been produced.[29] What explains the past persistence of the waggle-producer is its ability to produce waggle dances that relate according to *this* rule to the environment. It is when waggle dances have related according to this rule that they have impacted the consumer in such a way (explained, ultimately, by the engineering details of how the producer and the consumer are wired together) that the system's invariant function has been performed, and nectar

---

Teleosemanticists have reacted in different ways to the Swampman objection. Millikan's response has been to bite the bullet and to argue that denying intentionality to Swampman is, after all, acceptable (2010, 2017, 95–96). Other responses include attempts to foreswear the status of metaphysical necessity to the teleosemantic analysis of intentionality (Papineau 2001) and revisions of teleosemantics to allow that states lacking evolutionary history can possess content (Shea 2018, 170–73).

[29] Indeed, though I have abstracted from it in my presentation so far, Millikan often insists (e.g. 1989a, 296, 2017, 119–20) that the *time* at which a representation is *tokened* is often a significant feature of the representation. A dance *now* means nectar at so-and-so-location *now* (and of course, since the sun moves across the sky, the dance needs a time index). But then it becomes trivial that a dance cannot have ancestors with the exact same significant features as itself.

has been brought home to the hive. And it is in virtue of that fact that the rule *is* the semantic rule for waggle dances.

This has consequences for the relation between *correct representation* and *information*. As I mentioned above, Millikan denies that the *function* of a representation is to carry information about what it represents. But even so, when a representation is produced Normally, it *will* carry information about what it represents. To use Millikan's own equivalent terminology, representations will Normally be *natural signs* of what they represent. This is precisely because the producer's function is to ensure that a certain relation obtains between representation and world. For this to be its function, Millikan insists, there must be a way for it to perform that function, a Normal explanation for how it is accomplished, that involves a correlation or "lawlike relation" (p. 35) between representations and the things they represent that obtains in virtue of the producer's Normal mode of operation. If there were no such lawlike relation, if the producer had just *happened* to sometimes produce some items under certain conditions, then even if those items and those conditions actually had yielded fitness, we couldn't say that *producing the relation* between item and condition was a function of the producer. This is because it would not be something that had happened in virtue of any reproducible feature of the *producer*. It would just be something that had happened to it, by a fluke. Consequently, the items it produced would not be representations (Millikan 1993b, 127–28, 2004, 85, 2007, 444–45).

What kinds of mechanisms allow producers to Normally produce representations that are natural signs of what they represent? One possibility is that the producer is itself causally sensitive to the represented, the way the human pupil is sensitive to the intensity of incoming light. Another possibility is that the producer is sensitive to one or more *other* natural signs of the represented, the way the human visual system registers information about the distal arrangement of physical objects by being causally sensitive to the proximal pattern of incoming light.[30]

There is an ongoing debate among teleosemanticists about the proper status that should be assigned to information in a theory of content (cf. Shea 2007; Neander 2013). As we have seen, it is a consequence of Millikan's views that a Normally produced representation *will* carry information about

[30] The theory of information and natural signs, and how it contributes to making knowledge possible, occupies a large portion of the Millikanian oeuvre (e.g. Millikan 2004, 31–61, 2017, 109–54). The sense of "information" in which a Normally produced representation carries information about its *representandum* is that of *locally recurrent natural information*, which Millikan defines as obtaining between two states when states of one type correlate with states of the other for non-accidental reasons that obtain within a domain that the organism can either keep track of or maintain itself within (Millikan 2004, 40). What is important about this definition, for present purposes, is that information in this sense need not be anchored in absolute natural laws, but can be underpinned by local non-accidental connections including those that are maintained by the inner workings of the organism itself.

its content. But her theory also leaves room for representations, both correct and incorrect, that fail to carry information about the thing they represent.

In the case of incorrect representations this is trivial, as false representations strictly speaking do not represent anything. There is nothing for them to represent. But here we come to the first of our five questions above, namely, how teleosemantics explains the possibility of false representations in the first place.

The answer looks straightforward at first. If a representation represents the Normal conditions for the performance of the consumer's function as adapted to it (p. 47), then a false representation should simply be a representation produced under abNormal conditions. The representation is *supposed to* represent but fails to do so, because Normal conditions fail to obtain.

There is a complication, however. As we saw above, in order to count as a representation that falls under a certain semantic rule—in order for it to be supposed to represent something—the representation must belong to an RE-family of representations that have persisted because past tokens *have* conformed to that semantic rule. However, a *false* representation has to be one that is produced abNormally. It is therefore unclear whether it *can* belong to the same RE-family as those earlier, successful tokens, since members of the *same* RE-family must be produced by the *same* mechanisms (cf. p. 43), but an abNormally produced representation is, by that very token, not produced by the same mechanism as Normally produced ones.

To allow for false representations, Millikan must therefore countenance some looseness in what counts as "the same mechanism." Something can count as belonging to an RE-family if it is "produced in accordance with an explanation that approximates in some (undefined) degree to a Normal explanation for production of members" of the family (LTOBC, 25, 42). Similarly, a deformed eye can still count as an eye, since the process that produces it approximates the process that produces healthy eyes. False representations, on Millikan's view, are akin to *deformed* or *incomplete* traits: "to understand what a [representation] is we need first to understand what a true or satisfied [representation] is […] This is like saying that to understand what a scythe handle is (scythe handles are very puzzlingly shaped) one must first understand what a scythe is" (Millikan 2017, 156). Millikan acknowledges that this admits for some vagueness about what belongs to a family and hence what counts as a representation's falling under a given semantic rule. This, I believe, is not a weakness of the theory. Sentences can fail by being false, but also by containing non-referring terms, by being asyntactic, garbled, etc. There is no clear line demarcating representations from non-representation.

Often enough, we can use the semantic rule picked out by the Normal explanation to say how the world *would* have to have been in order for the representation to have been correct. Those merely possible conditions constitute the representation's content. This is of course perfectly analogous to

how I defined content in chapter 1: it specifies conditions that would have needed to obtain for the representation to have been correct. For sufficiently malformed representations, on the other hand, there may be *no* possible state of affairs that could have rendered it correct.

Failed representations will also have functions, though they will not be able to fulfill all their functions at once, at least not in a Normal way. In general, all the representations belonging to a given family will have a shared *invariant* function that it shares with its producer and consumer. In the case of the waggle dance, the invariant function is to bring the observing bee to nectar. Each representation, at least the "syntactically well-formed" ones, will also have an adapted function: to make the bee fly in this-or-that direction. The problem for false representations is that they cannot Normally perform both their adapted and their invariant functions at the same time.

This also allows us to see how a representation can be true despite the fact that it doesn't carry information about what it represents. It can be true *by accident*, by being formed by an abNormal process—one that has not been causally sensitive in the right way to the represented—but nevertheless happen to relate to the world in the Normal way (LTOBC, 45, 2004, 76).

We have now answered the first of our questions. Let us address the second: How does teleosemantics solve the specification problem of indicator semantics? The specification problem, recall, arises because a state typically carries information about many more things than it represents (p. 37). At first sight, this problem doesn't arise for teleosemantics. Although a representation Normally carries information about what it represents, not everything it Normally carries information about is something it represents. What it represents are the Normal conditions for successful consumer response.

Yet an analog of the specification problem arises for teleosemantics as well. If what the representation represents, the Normal conditions for successful consumer response, has Normal causes and Normal ways of being detected, then these, too, will be Normal conditions for the consumer response. Under Normal conditions, a place with nectar will be a place with flowers, and a place with flowers will be a place that has recently seen sufficient precipitation. Yet intuitively, the waggle dance represents the location of nectar, not the location of sufficient recent precipitation. It is also a Normal condition for a successful consumer response that it is caused by the right representation, but a representation hardly represents its own causal influence over the consumer.

It is to address these issues, I take it, that Millikan introduces the notion of a *most proximate* Normal condition:

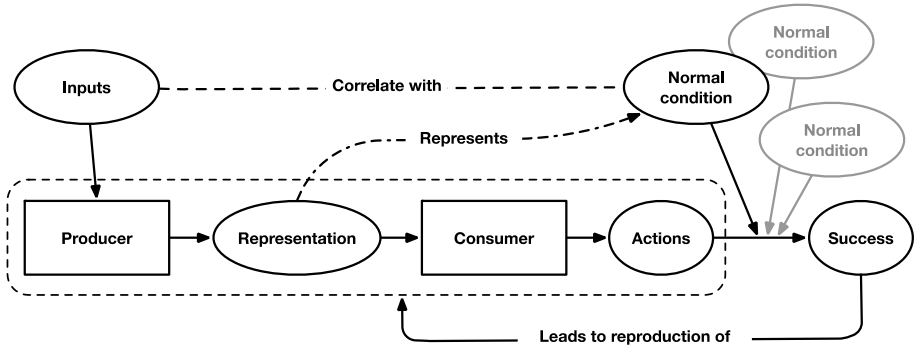
> [A representation] P [represents] whatever it maps onto that must be mentioned in giving the *most proximate* Normal explanation for full proper performance of its [consumer] as adapted to [P]. (LTOBC, 100)

In other words, a representation represents whatever condition must be mentioned in the "most proximate" Normal explanation of the consumer response. A Normal explanation becomes less proximate as it cites more explanatorily relevant background conditions. For instance, a proximate Normal explanation of the human circulatory system's function of oxygenizing the tissues need mention only the supply of oxygen from the lungs. A less proximate Normal explanation also mentions the source of that oxygen, Earth's atmosphere (LTOBC, 33).

Why the most proximate Normal conditions? The underlying idea, I take it, is that the most proximate Normal conditions are all we need to keep track of in order to estimate whether the consumer response is likely to be successful. As long as there is nectar in the flowers, the hive is likely to get food. It doesn't matter how the nectar got there, or why the consumer responds as it does. Consequently, the most proximate Normal conditions are what best serves the explanatory and predictive role teleosemantics assigns to content, i.e., to pick out conditions for likely success (p. 33; see also below).

One may worry that this move is insufficient to assign determinate content to representations. And indeed, some worries have been raised in the literature, and we will acquaint ourselves with them in section 2.5 below. For the moment, let us leave it at that and consider other aspects of the theory.

To sum up what has been said so far, the following diagram, courtesy of Gunnar Björnsson (2018), may help the reader understand the various components of the content-conferring producer-consumer system when it operates Normally and how representational content is assigned on its basis.



Let us now consider question 3, "What is the explanatory role or cognitive value of intentional notions, which justifies their importance in everyday as well as scientific thinking?" In the last chapter (p. 33) I suggested that according to teleosemantics, a representation's content does not directly entail anything about its present causal powers. We can now see why: although a representation will *Normally* interact in determinate ways with its consumer and, ultimately, its environment, there are many ways for conditions to be

abNormal. But content notions still have cognitive value for us. Since they pick out Normal conditions for successful functioning, they will allow us to draw conclusions about the likelihood that a representation will succeed, given further information about the world.

In general, the functions of biological systems involve the production and maintenance of states far from thermodynamic equilibrium. There is no reason to expect these states to come about under arbitrary environmental conditions. A biological trait has evolved because it has managed to bring about states of this kind under *certain* environmental conditions, the Normal conditions, but there is no particular reason to expect that it will bring about those states under different conditions. If it does so nevertheless, it will be a fluke, an unlikely stroke of luck.

Normal conditions provide a fallible epistemic guide for predicting future success, where "success" means *teleological* success, i.e., successful performance of function. If we know that (some features of) the Normal conditions for a function obtain, this raises the probability that the function will be successfully performed. Conversely, if we know that (some features of) the Normal conditions do not obtain, this lowers the probability that the function will be successfully performed. For the same reason, if we know that a function was successfully performed, this raises the probability that Normal conditions for that function obtained. And if we know that a function was not successfully performed, this raises the probability that Normal conditions for that function did not obtain.

Can we, just knowing the content of a representation that has been tokened, use it to infer that the content obtains, i.e. that the world is the way the representation says it is? Clearly, we do in fact do this all the time, and the reliability of language as a means for conveying information and disseminating knowledge depends on the sometimes-validity of this type of inference. As we saw above, a Normally produced representation will carry information (which, recall, is a factive notion) about what it represents. If conditions are otherwise Normal, a well-functioning producer will produce a representation only if the represented state of affairs obtains. "The producer's job […] is to make it that when it produces an R that raises the probability that a corresponding C obtains" (Millikan 2007, 444–45). How reliable is this statistical inference?

In the case of biologically hard-wired representational systems, where the consumer has no choice but to "trust" the producer, representations have to have been sufficiently reliable in the past to explain the persistence of the system. As Millikan likes to point out (e.g. 1989a, 283), that need not be particularly reliable at all. If the fitness cost of failing to detect an environmental feature is sufficiently high, and the cost of acting on a false positive is sufficiently low (as is often the case for detection of predators or mates), representations may be false most of the time. The benefits still outweigh the costs. Humans, of course, have some ability to determine whether a repre-

sentation has been Normally produced, by comparing it with other representations concerning the same matters, or by relying on experience of producer behavior under various circumstances. At bottom, this is just yet more representations, and somewhere deep down in the cognitive machinery there must be mechanisms that blindly trust their producers in the "hope" that conditions will remain sufficiently similar to the historical norm for this trust to be vindicated.

We still have two more questions to address. Question 4 concerned in which sense intentional properties provide norms for their tokening, and we have seen how teleosemantics proposes to answer this question: the norms in question are teleological norms, defined according to the etiological theory. Something remains to be said about in what sense teleological norms are indeed "norms" and how they may or may not relate to agent-level norms. I will return to these questions in section 2.4.

Question 5 concerned how teleosemantics can account for complex productive systems of representation like human thought and language. We have seen how the limited productivity of simple representational systems like waggle dances can be accounted for within the teleosemantics framework, but the productivity of language and thought is of a different order entirely. How teleosemantics seeks to accommodate this fact will be the topic of the next three sections.

2.2. Conceptual Thought

With the fundamentals of Millikanian teleosemantics in place, we will now look at how the theory accounts for the intentionality of human thought and discourse. These two things are closely linked. Human thought is itself discursive, whether or not we have a "language of thought." It is adapted to the needs of a language-using animal and reliant on discourse to function well. Human discourse, for its part, is principally a means of conveying thought and for influencing the thoughts of others.

We will consider thought first. Thought paradigmatically takes the form of propositional attitudes like beliefs. Propositional attitudes possess content. Hence, if teleosemantics is to be a viable general theory of content, it must be possible to apply its principles of content determination to propositional attitudes and not just to bee dances. That presupposes that beliefs (and the other propositional attitudes) mediate between producers and consumers in the way described in the previous section. How should we understand these notions, when applied to something as nebulous as belief?

If there are consumer systems that rely on beliefs to perform their functions, they must be quite different from the consumers described in the examples above, those simple systems that translate representations directly into specific kinds of actions like flying away in a certain direction. Beliefs

are utilized in complex processes of inference, learning, and decision-making, where they interact with other beliefs and other propositional attitudes to generate further attitudes and ultimately behavior. Can we, on the assumption that the cognitive mechanisms that enable these processes are the consumers of beliefs, construct a theory about the functions of beliefs that allows us to assign appropriate contents to them?

Underlying the aforementioned cognitive processes is the *conceptual* structure of beliefs. Concepts and conceptual thought are the topic of two of Millikan's monographs (2000, 2017). In *On Clear and Confused Ideas* (2000; henceforth *OCCI*), Millikan characterizes a concept as essentially involving an *ability to reidentify* its object (OCCI, 50).[31] The ability to reidentify or recognize objects is central to conceptual thought. Take inference, for instance. In an inference, the premises must typically have some co-referring terms in common: all *men* are mortal, Socrates is a *man*. For us to be able to *draw* the inference, we must be able to tell that the two instances of the term refer to the same thing, i.e., we must be able to *identify* their referents with each other (OCCI, 141-42). Similarly, in learning, the subject must be able to recognize or reidentify an object across several encounters in order accumulate information about it and treat that information as being about one and the same thing.[32] We do this by having concepts for the things we reidentify. We can have concepts for individuals, kinds, properties and stuffs, i.e., any sort of entity about which we could profitably accumulate information and learn in order to deal with this entity better when we encounter it again.

How do concepts let us accomplish acts of reidentification? A concept, to Millikan, is a feature that can be shared among attitudes. An image can be used to clarify this idea. We can think of a concept as a *mental word*, which helps compose a *mental sentence* that specifies the content of a propositional attitude. In the same way that you can't talk about a thing (except via description) unless you have a name for it, you can't think about something—can't form an attitude about it—without a concept.

[31] In *Beyond Concepts* (2017), Millikan eschews the common term "concept" in favor of the neologism "unicept." She also introduces the notion of a "unitracker," and distributes the theoretical roles indiscriminately attributed to concepts in *OCCI* across unicepts and unitrackers. While a unicept plays the role of component of propositional attitudes, the unitracker associated with that unicept corresponds to the ability to reidentify the unicept's object (2017, 7–9). Not to complicate matters unnecessarily, I have largely employed the apparatus of the earlier *OCCI* here, but I will indicate how *Beyond Concept* diverges from that framework in footnotes like this one.

[32] We can imagine a creature such that each time it encountered a particular thing *x*, it stored information about it, but which had no way of telling that these different pieces of information consisted in information about the *same* object. In a sense, this creature would "accumulate information" about *x*, but it could not draw inferences on the basis of this information and could not use it to predict *x*'s behavior the next time it encountered it. In short, it would have no way of *profiting off* the fact that the information it had accumulated concerned the same thing. According to Millikan, such a creature would not possess a concept of *x*.

But as Millikan reminds us, the very fact that two attitudes share a feature doesn't entail that the cognitive system is capable of treating them as being about the same thing (OCCI, 133). This shared feature must also have the *function* to signal, to the cognitive system, that "these two attitudes are about the same thing." The feature must be a *sameness-marker*. This is one of the fundamental roles of concepts in cognition: to function as sameness-markers for purposes of inference, learning, etc. (OCCI, 144)

Concepts can also be treated as repositories of information, tying together all the attitudes that are supposed to be treated, by the cognitive system, as being about the same thing. This idea is akin to that of a *mental file* (cf. Recanati 2012). The connection between it and the image of concepts as mental words is direct: all beliefs (mental sentences) that contain the same concept (mental word) can be thought of as jointly constituting a repository of information about the object of that concept. Think of searching for a word in a database and getting a list of hits.

As the "mental word" image suggests, the content of an attitude and the objects of the concepts that it employs must mutually constrain each other. This is to be expected, if the function of an attitude is to aid processes of conceptual thought (inference, learning) and if its conceptual components constrain *which* processes it is supposed to aid and, hence, the conditions that need to obtain for those processes to be Normally successful. Consider, for example, my belief that Donald Trump has a weird haircut. One function of this belief is to allow me to learn more about Donald Trump by helping me recognize him (in pictures etc.) and so add new information to my repository of information about him. Why is *this* a function of the belief? Well, because it is a belief *about Donald Trump*. And why is it a belief about Donald Trump? Because it contains my concept of Donald Trump. And in order for it to fulfill its function—to allow me to recognize Donald Trump—in a Normal way, it seems it must also be true, i.e., its descriptive content *Donald Trump has a weird haircut* must obtain. If Donald Trump doesn't have a weird haircut, the belief will not help me recognize him.

Hence, teleosemantics vindicates the intuitive principle that the content of a belief is partly determined by the objects of the concepts that compose it. What, then, determines the object of a concept? It should not come as a surprise that according to Millikan, a concept's object is determined by its *function*—by what it's supposed to help track, recognize, and learn about (OCCI, 193).[33]

However, the mind doesn't come pre-equipped with concepts of all the different things that an adult person can think about. It must be able to form new concepts as it encounters new entities. The functions of those newly-formed concepts, which is what determines what they are concepts of, must

[33] In *Beyond Concepts,* the object of a unicept is determined by the function of its associated unitracker, by what that unitracker is supposed to track (p. 72).

somehow derive from the mechanisms Normally governing the formation of new concepts in encounters with new entities.

Detailed discussion of Millikan's theory of the mechanism of concept-formation, and the many ways in which the process can go wrong, would take us too far afield. To get a sense for some of the difficulties involved, consider that in order to assign a unique object to a newly-formed concept, it is not sufficient to identify that object with the thing whose causal influence occasioned the concept's formation. To see why, consider that we may have concepts for both *individuals* and *kinds*, among other things. Suppose I encounter a marten for the first time and form a concept MARTEN. Of course, the particular marten I encounter is both an individual marten and an instance of the kind *martens*. Is my concept, then, a concept of the individual or the kind? Both entities figure in the cause of my concept formation.

If the function of a concept is to track a *specific* entity (whether individual or kind), then, according to the etiological theory of functions, it must have been formed by a process that has, in the ancestral past, been able to produce concepts that have actually been capable of tracking *specific* entities. For MARTEN to be a concept of a kind rather than an individual, it must have been formed by a process that has historically produced concepts capable of tracking kinds rather than individuals upon encounters with members of those kinds. To do this, the concept-formation machinery must "seed" the concept with tools dedicated to re-identifying kinds rather than individuals. In the apparatus of *OCCI*, these tools derive from a "template," an implicit understanding of how to keep track of things of the general category to which the target entity belongs,[34] such as the categories *individual* and *kind*.[35] (OCCI, 28-32). For instance, kinds but not individuals can be instantiated at several different places at once. The difference can be exploited to ensure that it is the kind rather than the individual that is tracked, or vice versa.

The crucial role played by templates in determining the objects of our concepts entails that our minds are particularly adapted, not to say hard-wired, for thinking about certain kinds of things but not others. Which kinds of things we possess templates for will depend on what kinds of things the concept-formation machinery has, in the ancestral past, been able to produce successful trackers of. Now, something is easier to track if its properties stay roughly the same over time or, insofar as they vary, vary in predictable ways. It is also more useful to have a concept for something like this, since the information stored under the concept will then be applicable upon new encounters. Hence, the kinds of things we have templates for, and therefore the kinds of things we are able to form concepts of, will tend to be things

[34] In the apparatus of *Beyond Concepts*, they come "pre-packaged" in the unitracker associated with the fledgling concept.

[35] They could also correspond to the categories *animal individual* and *animal kind*. Templates need not correspond to the highest levels of ontological categorization

that stay constant over time or vary in predictable ways. The mind is therefore particularly well adapted to forming concepts of what Millikan calls *substances*. Substances are a large category of entities that includes individuals, kinds, and stuffs that are "natural" in the sense that for each of them, there exists some natural explanation for their similarity in determinate respects over time or across instances. It includes individuals like Karl Bergman, kinds like *cars*, and properties like *blue*, but not, for instance, kinds like *incars* (cars that are indoors; cf. Hirsch 1976) or the property *grue* (the property of being green and observed before some future time *t*, or blue and not so observed; cf. Goodman 1983, chap. 3). (for details on the underlying metaphysics, see LTOBC, chaps. 16–17, 1998, 2000, chaps. 1–3, 2017, chap. 1). We can, evidently, acquire the concepts of incars and grue, but doing so presupposes discursive abilities that are grounded in our prior possession of concepts of ordinary things like cars and the color blue.

We should emphasize that a concept, for Millikan, is not equivalent to a *description* or *theory* of, or set of necessary and sufficient conditions for, an entity. Each one of a person's concepts may be *associated* with such a theory—what Millikan calls the person's *conception* of the thing—but these conceptions need not be shared among people who possess concepts of the same thing. This feature of the theory is correlative with its thoroughgoing externalism. To have a concept for something, it suffices that the concept has been formed by the right processes under the right sort of causal influence from that thing. By the same token, the theory is also *minimalist*. It entails that we can form concepts of novel entities on the basis of very little acquaintance with them. In principle, it is sufficient to see or read about something once in order to form a concept of it. As long as the concept comes equipped with the right tracking equipment, a minimal amount of seeding information can be sufficient to allow the concept to do its work of allowing us to recognize and accumulate information about its objects. Seeing a term for the first time in print often suffices to give me the ability to think about its referent, because using this minimal seeding information—that the thing is called by that name, plus whatever the text says about it—I can then Normally go on to accumulate more information about it (OCCI, 88-91).[36]

But this process can easily go wrong. Upon meeting a new person for the first time and then immediately meeting his identical twin, I may erroneously store those two encounters under the same newly formed concept. Unless

[36] When you know the word for a thing, discourse can become a means for you to learn more about that thing, to *recognize* that thing when it is spoken of and add the conveyed information to your store of information about it.

On several occasions, Millikan emphasizes the parallels between language comprehension and perception. Discourse is a medium through which we can learn about objects, just like the media (light, sound, etc.) through which we learn about objects perceptually. Learning in this way is a fallible process, since the medium can operate abNormally—when the speaker is insincere or ignorant, etc.—but the same goes for perceptual media. See OCCI, 89.

I discover my mistake, the concept will continue to be filled with information from two different sources. The result is a *Confused Idea* (as in the title of *On Clear and Confused Ideas*) or, in the more technical terminology of *Beyond Concepts*, an "equivocept" (2017, 91). We can also happen to form concepts that do not in fact have any object, because the causal influence under which they were formed did not originate in an entity matching its associated template. We mistake some moving shadows in the woods for a person, though nobody is there. Then we have an empty concept or a "vacucept" (2017, 93). And it may happen that we form two distinct concepts for one and the same thing—the morning star and the evening star, or the distinguished gray-haired man and Bernard J. Ortcutt[37]—incorrectly treating them as distinct.[38]

In the next section, we will look at Millikan's theory of discourse. The remarks in this section will prove relevant for understanding that theory. Discourse is dependent in several ways on the capacity for conceptual thought. For one thing, learning a language requires us to have concepts of individual words and other linguistic devices, to be able to recognize them and accumulate information about them. Correlatively, we must have concepts of *things* to learn the words for them. Most relevant, for our purposes, is the necessity of conceptual capacities to understand and participate in linguistic communication. Linguistic interpretation, according to Millikan, crucially involves the ability to token concepts whose objects are the referents of the terms used, hence to identify those referents (LTOBC, 71).

The compositional structure of sentences, composed of referring terms (and some logical connecting tissue), and the conceptual structure of thoughts mirror each other, and if Millikan is right they are designed to do so. Here, we touch upon a topic I have already briefly broached (p. 24): the contrast between intentionality generally, defined as the possession of content, and what we may call the inferential articulation of representations, their functional capacity to enter into inferences that recapitulate modal relations between their contents. Contents are not, as such, inferentially articulated, and consequently, a representation cannot enter into inference simply in virtue of having content. This is as it should be, on a view of intentionality that treats waggle-dances and magnetosomes as intentional, because these clearly cannot enter into inferences. The conceptual structure of beliefs gives them inferential articulation and can be seen as a way for the cognitive sys-

[37] The example, of course, is from Quine: "There is a certain man in a brown hat whom Ralph has glimpsed several times under questionable circumstances on which we need not enter here; suffice it to say that Ralph suspects he is a spy. Also there is a gray-haired man, vaguely known to Ralph as rather a pillar of the community, whom Ralph is not aware of having seen except once at the beach. Now Ralph does not know it, but the men are one and the same" (Quine 1956, 179).

[38] The methods whereby the mind continuously attempts to discover and correct mistakes of these kinds are discussed in OCCI, chapter 7.

tem to approximate the ideal of causal relations between beliefs reflecting the modal relations between their contents (p. 31). But the conceptual structure of beliefs is not uniquely determined by their contents. In Millikan's words, "intentionality and rationality are *not* two sides of the same coin" (LTOBC, 140).

An analogous point holds for linguistic representations: the modal relations between their contents are not transparently revealed by their overt inferential or deductive relations. We will return to this point after having discussed the teleosemantic theory of discourse and the compositionality of language, at the end of section 2.4.

2.3. Discourse

Language is a tool for communication, an activity that involves the conveying of information as well as the directing of other people's actions. In speaking, the speaker manifests her communicative intention, but in order for those intentions to be fulfilled, she must choose words based on the likely response of the hearer.

The speaker's choice of words, then, will typically depend on two things: her purpose in speaking and the hearer's expected response to various word-choices. The speaker's expectation on the hearer will in turn depend on her past discursive experience. The speaker generalizes from that experience to form an idea of standard hearer responses and, on the assumption that the hearer approximates the standard (and duly compensating for those respects in which he is assumed to deviate from it), she chooses her words.

These two determinants of speaker word choice correspond to two traditional notions of linguistic meaning. Speaker intentions correspond to what is sometimes called "speaker meaning," i.e., what the speaker means to happen when she speaks or what she intends to convey.[39] Standard audience responses correspond roughly to what is called "conventional meaning." The speaker cannot choose the conventional meaning of a term. It is already part of the social background against which she acts.

Millikan's theory of language is largely a theory about the conditions under which symbols come to have conventional meanings. Teleosemantics is well-suited to this purpose, since it relates the semantic or intentional properties of symbols to their history. Since the conventional meaning of a term is a product, not of its present use but of factors that are already in place when the symbol is used, it is but a short step to the conclusion that conventional

[39] This is a simplification. Speaker meaning is a narrower notion than speaker intention: a liar intends to deceive her audience, but the deception is not part of the speaker meaning. These subtleties need not concern us here, since we will mainly be concerned with the other kind of meaning, conventional meaning. For the seminal discussion of speaker meaning, see Grice (1957).

meaning is determined by a symbol's *history*, a history consisting of past tokenings of the same symbol-type which have shaped the collective expectations of the speech-community.[40] If this history can be understood as a selection process, linguistic meaning can be subsumed under the general teleosemantic analysis of intentionality. To find the semantic properties of a linguistic representation, we ask how it is reproduced and what explains its past persistence.

The reproduction of a linguistic token is a process mediated by human minds. Barring flukes, a human mind must be incentivized to participate in this reproduction.[41] The primary way in which linguistic expressions incentivize humans to reproduce them is by enabling them to solve communication problems.

Paradigmatic communicative situations have at least two participants, a speaker and a hearer (or a writer and a reader).[42] In any given communicative situation, either or both of these can come away from the interaction incentivized to continue to engage in communication and to use and respond to specific words in specific ways. The speaker is incentivized if her speaking influences the hearer in ways she deems beneficial to herself. The hearer is incentivized if he gains some perceived benefit from whatever response he makes to the speaker's speech. However, if linguistic expressions are to secure their own *long-term* persistence in the speech-community, *both* participants must be incentivized. If the speaker fails to benefit, she will have less incentive to use those words for that purpose in the future. If the hearer fails to benefit, he will be less likely to respond in the intended ways in future interactions, and so make the speaker less likely to use those linguistic means to try to influence him that way.

In sum, it is *cooperative* uses of linguistic expressions that explain their persistence, uses where speaker and hearer both contribute and both benefit. Speaker and hearer rely on a set of mutually known linguistic precedents in order to produce mutually beneficial outcomes. In this process, language devices function as *coordinating convention* used by speakers and hearers to solve coordination problems (Millikan 2005a; cf. Lewis 2011). Note the implication that on the etiological theory of function defended by Millikan, conventions *qua* reproduced patterns of behavior also possess an ancestry that can confer proper functions on them.

---

[40] According to Millikan, symbol-types are individuated etiologically, rather than by their physical shape. Two symbols belong to the same type only if they belong to the same RE-family. If weather and geological processes were to form the shape of the word "dog" on some remote plain, it would not be a token of the English word "dog" (Millikan 1984, 72–73).

[41] For this reason, language forms can be compared to "memes" in the sense of (Dawkins 1976, chap. 11).

[42] As Andrew Reisner has pointed out to me, we sometimes communicate with ourselves. "Talking to oneself" is a ubiquitous phenomenon, but it is arguable whether it qualifies as communication. However, our past selves can communicate with our future selves, as when I leave a note for myself to make sure that I remember something important.

Millikan uses the term "stabilizing function" to denote that cooperative use of a linguistic expression that explains its past persistence. It is a "stabilizing" function because when it is successfully performed, it stabilizes speaker and hearer expectations and so allows the expression to go on performing the same function in the future. When the stabilizing function fails to be performed, speaker and/or hearer may adjust their expectations, and the ability of the same expression to perform that function in the future diminishes. The use of the expression may then drift (with an accompanying drift in function) or cease entirely. In the limiting case, if no expressions ever performed their stabilizing functions—ever enabled speakers and hearers to communicate for mutual benefit—people would eventually stop talking altogether (LTOBC, 31–32).

For instance, by uttering indicative sentences (making assertoric speech-acts), speakers can convey information to hearers, i.e., impart *true belief*. This, according to Millikan, is their stabilizing function (LTOBC, 53–54). Unless hearers sometimes believed what they were told, speakers would have no incentive to go on speaking, and unless speakers sometimes told the truth, hearers would have no incentive to go on believing them.[43]

Of course, it is relevant *which* true belief the hearer forms. Specific assertions, formed by uttering specific sentences, are supposed to convey specific beliefs. Correlatively, a Normal hearer (i.e., a hearer that interprets language in a Normal way, in the same kind of way that has contributed to the past persistence of the language form) will indeed form a specific belief—the belief that *P*, let's say—upon hearing the assertion. For this belief to be true, it must be the case that *P*. So this, that *P* is the case, is a Normal condition for the assertion to perform its stabilizing function to produce true belief. This is to be expected, because an assertion ought to have the same content as the belief it is supposed to produce, and since the content of a representation is part of its Normal conditions, *P* should be a Normal condition for an assertion whose function is to produce the belief that *P*.

Three things need to fit together for assertoric communication to function Normally. The first is the relation (mapping) that Normally obtains between the structure of the uttered sentence and the world. This relation corresponds to the semantic rules for the sentence, the rules that have explained past successful communication (cf. p. 48). We will see in the next section how there can be this type of Normal relation between a sentence and the world in virtue of the ancestry of the sentence's structure and components. The second is

[43] This claim, it should be noted, doesn't conflict with the fact that people sometimes make assertions with no purpose or expectation to be believed, as might happen when an actor reads a line of dialog on a theater stage or when, accused of a murder, I profess my innocence before the court although I know that the guilty verdict is already a foregone conclusion. But such uses could not, on their own, have secured the past persistence of the practice of assertion. They can exist only against the background of a practice of information-conveying assertions. They are parasitic uses serving parasitic functions (Millikan 1984, 55).

the relation that Normally obtains between the world and the belief that the hearer forms upon hearing the assertion, i.e., the "semantic" rules for the belief. The third is the way the hearer interprets the sentence, the way he performs the "translation of outer intentional signs into inner intentional signs" (Millikan 2017, 113). For the hearer's translation to be Normal he must translate the assertion into a belief that maps to the same state of affairs as the assertion. He must do this by exploiting the semantic rules of the sentence, i.e., those rules that also explain how other earlier speakers have been able to understand sentences in the same language (LTOBC, 99).

Let us turn to the relation between language and thought. Most speech-acts have stabilizing functions that involve producing propositional attitudes in the hearer. For assertions, this is typically a belief. For a command, it is typically an intention to obey the command. Later on, we will investigate forms of discourse whose function is to convey *normative* and *moral* judgment. Propositional attitudes, as we saw in the previous section, are conceptually articulated. Here we begin to see how, on the teleosemantics picture, human discourse is dependent on human conceptual thought. Not all kinds of discourse, it is true, convey propositional attitudes on Millikan's picture (Millikan 2005e, 2018). Identity-statements like "Hesperus is Phosphorous" constitute an example of a speech-act form whose function, according to Millikan, does *not* involve the conveying of propositional attitudes. Instead, she claims, it has the function to connect or merge two concepts associated with the two names flanking the identity sign ("Hesperus" and "Phosphorous") (LTOBC, chap. 12). But this stabilizing response still require conceptual capacities, and so does not vitiate the general point that language-forms can only perform their stabilizing function if the hearer has conceptual capacities.[44] When I give commands to my phone, which presumably lacks conceptual capacities, this is therefore not a Normal use of language.

At the same time, the intentionality of a linguistic representation is not *derived from* that of thought. A speech-act doesn't have its intentional properties in virtue of speaker or hearer attitudes to the speech-act. A linguistic representation acquires its intentional properties on the basis of the same content-determination principles as every other kind of representation. Linguistic meaning does not require the kind of nested speaker attitudes characteristic of Gricean views of language (Grice 1989). Nevertheless, the cooperative nature of stabilizing functions ensures that it is always a Normal condition for a speech-act that the speaker's purpose in speaking aligns with the function of the speech-act. For instance, when asserting that *P*, I Normally intend that you believe that *P* and intend that belief to be true (cf. Millikan 2005e).

[44] Some exceptions to this rule probably exist. A greeting like "Hello!" might not engage the addressee's conceptual capacities.

We are now in a position to address the fourth of the questions posed in section 1.5, "How should we explicate the intuitive sense in which the content of a representation provides a norm for its tokening?" On many occasions, Millikan strongly emphasizes that the teleological norms characterizing speech-acts and other representations are not norms in a moral or evaluative sense, i.e., agent-level norms. "The norms for language are uses that have had 'survival value' […] They are not prescriptive or evaluative norms. Their status has nothing to do with anyone's assessment. A norm is merely a measure from which actual facts can depart; it need not be an evaluative measure" (Millikan 2005f, 83).

There is a tradition from John Austin (2009) and John Searle (2011, 38) that takes discourse to be a *constitutively* rule-governed activity: speech-acts are defined and individuated by the fact that they fall under certain rules or norms. In this vein, there has been a great deal of debate in recent literature about what the "norm of assertion" is, whether truth, knowledge, justified belief, etc. (e.g. Williamson 2009, chap. 11; for review and criticism, see Pagin 2016). Millikan, however, rejects this tradition (2005a, 21–22). Instead of constitutive norms, she takes (most) speech-act types to be individuated by their stabilizing functions (2005c), i.e., teleological norms.

But it doesn't follow that the teleological norms governing speech-acts have *nothing* to do with "evaluative," agent-level norms. The correct functioning of our discursive practices is something we all have a significant stake in maintaining. It is also something we have considerable influence over. It would therefore be strange, I maintain, if there weren't—at least *pro tanto* and most of the time—norms enjoining us to take the various steps we can take to ensure that our speech-acts will function properly.

In particular, speakers and hearers can influence whether communication will be successful by influencing whether Normal conditions obtain for it. Since the obtaining of Normal conditions significantly raises the probability that a speech-act will be successful (cf. p. 53), speakers, by influencing whether Normal conditions obtain, can influence the chances that the communication episode proceeds successfully. As we have seen, one Normal condition for assertion is (trivially) that the assertion is true, and there does seem to exist, most of the time, a *pro tanto* norm enjoining speakers to try to ensure that what they say is true (exceptions will include actors on the stage, etc.)[45] Normal conditions also require that the speaker is cooperative, i.e., that she intends for the hearer to believe something true. This corresponds to a norm of sincerity which is, again, in force *pro tanto* and most of the time.[46]

[45] Note that "ensuring that what one says will be true" entails, not only ensuring that the belief one intends to convey is true, but also that the words one chooses to convey that belief are actually the ones conventionally used to convey it.

[46] Another Normal condition for assertoric success is that the hearer believes the speaker. Here, the intuition that the Normal condition corresponds to a norm for the agent is significantly weaker. It is at least not clear that *hearers* are under any particular obligation to believe

What I want to claim, then, is not that teleological norms *necessarily* correspond to agent-level norms, but that they *often* do and that there is a good reason why. In saying, this, I am not relying on any detailed first-order normative theory. I am relying on an intuitive principle whose merit I hope the reader can see, namely, that agents have a *pro tanto* responsibility not to subvert expectations on which others rely.[47]

We have now looked at the basic ideas underlying Millikan's theory of discourse and seen how it relates to the basic principles of teleosemantics. But it remains to be shown how a speech-act or sentence, formed on the basis of various crisscrossing linguistic precedents, can possess a single function and a determinate set of truth conditions. In tackling this topic we will also finally address the fifth question of section 1.5, "how does teleosemantics account for complex, productive systems of representation, such as human thought and language?"

2.4. Compositionality

We have relied heavily on waggle dances to illustrate features of Millikan's theory of intentionality, but language is disanalogous with waggle dances in several respects. For one, it is learned rather than innate. We have already seen how a learned behavior can acquire teleology and hence intentionality. It secures its persistence by incentivizing people to use it, thus engaging their (ultimately innate) psychological reward systems.

Another disanalogy lies in the productivity and recursivity of language. Since the significant features of waggle dances vary along continuous dimensions, there is in principle an uncountable infinity of them. But they can still be represented by a two-dimensional Cartesian space. Their "syntax" is simple. Linguistic representations, on the other hand, can be arbitrarily complex due to the recursive structure of language. In addition, the standing possibility of introducing new words and syntactic structures makes the family of linguistic representations open-ended in a way that waggle dances decidedly are not. There is no obvious limit to the *kind* of world state that can be represented by a sentence.

Linguistic representations, also unlike waggle dances, do not have a single ancestry. Instead, each linguistic representation is formed on the model of a number of different precedents, corresponding to individual words as

---

what they are told. On the other hand, if a speaker has taken what she perceives to be adequate steps to be credible, it is not surprising if she should be annoyed at a hearer who nevertheless refused to believe her. Moreover, consider that if a hearer forms a *false* belief, this is *also* an instance of unsuccessful communication. Hence, hearer discretion in whether or not to be credulous is consistent with the present proposal.

[47] Nor do I have (at least as yet) a theory of the nature of agent-level norms. Contributing to a theory of the latter sort is, of course, the task of this thesis.

well as syntactic structures. Each such precedent constitutes a family, and so each linguistic representation will belong to a number of different, criss-crossing families.

When I inform you that Axel is sad by uttering the sentence "Axel is sad," our successful communication depends on our mutual familiarity with a number of linguistic precedents. Some of these precedents involve the word "Axel," others the word "sad," yet others only the abstract subject-copula-adjective structure. Some of them *may* consist in tokens of the same wholesale sentence-type, but that isn't necessary for us to be able to communicate. Each precedent has its own ancestry, involving wildly diverse speech-acts.

With each such ancestry comes a stabilizing function and a Normal explanation for that function, which together have explained the past persistence of the lineage. By being acquainted with the general pattern responsible for the proliferation of each of the precedents and generalizing it to the current situation, a hearer can come to understand a novel sentence. However, a speech act doesn't have a number of distinct stabilizing functions, one for each of its lineages, which are performed in parallel. It has a single function that is somehow determined jointly by its features (LTOBC, 53, 80–81). The functions of the features must therefore be *relational* functions. They depend on the sentential context in which the feature appears (LTOBC, 80).

To get clearer about this, let us consider the types of linguistic lineages to which a speech-act can belong. I have mentioned individual words and syntactic structures, and in much of traditional linguistics (including generative grammar in the Chomskyan tradition), these are viewed as discrete and non-overlapping categories. But this is a simplification. What is reproduced when a known word is repeated in a novel speech-act context is not only the word itself but also its relationship to the sentential context. The child learns not only the word "dog" but also that "dog" can stand together with "the" in subject position, etc. Additionally, many linguistic lineages consist in idiomatic phrases with one or more open argument places, and these straddle the distinction between lexical units and syntactic structures. All these different types of structures constitute possible precedents on the basis of which a human can decode a novel linguistic representation:

> [H]uman memory must store linguistic expressions of all sizes, from individual morphemes to full idiomatic sentences (such as *The jig is up*). These expressions furthermore fall along a continuum of generality, defined by the number and range of variables they contain. At one extreme are word-like constants such as *dog* and irregular forms such as *bought*, with no variables to be filled. Moving along the continuum, we find mixtures of idiosyncratic content and open variables in idioms like *How dare* NP VP and *take* NP *for granted*. Still more general are the argument structures of individual predicates such as *dismantle* NP and *put* NP PP. Finally, at the other extreme are rule-like expressions consisting only of very general variables such as V → V-suffix and VP → V (NP). (Jackendoff and Pinker 2005, 221–22; cf. Millikan 2008)

If we take as our example an idiomatic phrase with open argument places, like "how dare NP VP," we can illustrate the principle of relational stabilizing functions. In this case, the phrase constitutes a lineage that has been used (roughly) to convey outrage over various things and has persisted because of it. *What* it has helped convey outrage *about* has varied from one deployment to the next, depending on the context of deployment (specifically, on which terms substitute for NP and VP). The function of the phrase, then, is to convey outrage about the fact that NP's referent has performed the action that constitutes VP's referent. This is a relational function. "How dare NP VP" is supposed to convey an outrage that bears a specific relation to NP and VP.[48]

Most lexical units, like nouns and verbs, can figure in a wide variety of different syntactic constructions and hence their stabilizing functions cannot be so straightforwardly defined. Nouns and verbs can both be considered, for present purposes, as referring terms (nouns typically refer to individuals or kinds, verbs typically refer to action types). It would be a mistake, however, to suppose that the stabilizing function of a noun or a verb is to *refer*. Reference is an abstract way of characterizing a term's effect on the truth-conditions (or satisfaction-conditions) of the sentences it helps compose. It is not a way for the term to intervene in the causal order, something that could explain its persistence. Rather, the stabilizing functions of a referring terms in its sentential context is to constrain in certain ways the attitude or state that is supposed to be produced by the sentence when uttered. In "how dare NP VP," the noun substituting for NP will determine whose actions the sentence is supposed to convey outrage about, and the verb substituting for VP will determine what action of NP's it is supposed to convey outrage about.

Just as the crisscrossing linguistic lineages of a sentence jointly determine its stabilizing function, they also jointly determine how the world has to be in order for that stabilizing function to be Normally performed. In other words, they jointly determine the sentence's descriptive content. As we saw on p. 48, the semantic rule (or, as Millikan prefers to say, semantic mapping function) that picks out the descriptive content of a representation is determined by the relation between representation and world that has explained the success of successful ancestral representations. Since a sentence has several crisscrossing ancestries, the rules must also be crisscrossing. There must be one rule explaining how sentences with the word "Axel"[49] have contributed to the continued use of sentences with "Axel," one rule explaining how

[48] Note that if the open positions in the sentence schema defined by this construction are not occupied by lexical units of the specified kinds—a noun phrase and a verb phrase respectively—there is no obvious thing that the construction is supposed to do. "How dare run Peter?" has no clear meaning. The relational functions of linguistic lineages are not necessarily defined for every potential relatum.

[49] Here, the word "Axel" picks out a *single* lineage corresponding to the use of this name for talking about a single individual. Cf. (Millikan 1984, 79).

sentences with "sad" have contributed to their own persistence, and so on, and these must together determine a rule for the whole sentence.

Let us stick, for ease of exposition, to indicative sentences (and the assertions made by uttering them). According to Millikan, the function of an indicative sentence is to produce a certain true belief. The rule for each of its features must therefore be one that can help explain how that feature has historically contributed to producing true belief. Jointly, the rules given by the sentence's features should explain the truth of the belief that *it* is supposed to produce. In other words, these rules should pick out the truth-condition of the belief. Conversely, the sentence should produce the belief whose truth-conditions match the world-state jointly picked out by the rules for its features. The sentence's relation to the world and its interpretation by the hearer must "match" in order to jointly explain linguistic persistence (cf. the discussion of stabilization on p. 62 above).[50]

Of course, it is only the complete sentence that relates to a *specific* possible world state, i.e., has a content. Its features, corresponding to discrete sentence lineages, do not themselves pick out specific world states. We can instead profitably think of them as picking out sets of world states, together with principles for how these sets are narrowed down by the rest of the sentential context. Analogously, we can think of each discrete lineage as corresponding to an unsaturated sentence schema, which must be saturated in order to yield a determinate content. The terminology of saturation is, of course, Frege's, but whereas Frege saw only predicates as unsaturated, while considering proper names to be "complete in themselves" (Frege 1952, 31), Millikan insists that all sentence features are incomplete:

> "Theaetetus" is just as incomplete a sign—just as much a sentence with gaps in it—as any predicate or sentence form. "Theaetetus," in fact, is so gappy that the sentence form itself needs to be filled in. (LTOBC 106)

Different ways of filling in the sentence schema constituted by a given sentence feature (recall that both words and syntactic structures are features in the relevant sense) give different contents, according to the rules that have explained the feature's past persistence. These rules correspond to what Millikan in *LTOBC* calls the feature's *Fregean sense* (LTOBC, 111) and, later, its *semantic mapping function* (2005b, 64). The term "Fregean sense" derives from the fact that the rules constrain the type of state of affairs that must obtain for a sentence with the feature to be descriptively correct. For a name like "Axel," these rules will pick out *Axel* as the person whose circum-

[50] It is an assumption of Millikanian teleosemantics that there *is* a unified explanation of the past persistence of each language device that appeals to the mutual stabilization of its relation to the world and its interpretation by hearers. It is not an assumption of teleosemantics that this explanation should take a form that vindicates any specific current theories in semantics or psycholinguistics.

stances determine whether Normal conditions obtain for the sentence, hence whether it represents correctly. In other words, they will fix the name's referent, thus playing one of the important theoretical roles of the *Sinne* posited by Frege (cf. LTOBC, 102-07). The terminology invites confusion, however, since Millikan's Fregean senses do not play all the theoretical roles of *Sinne*. Therefore, I will stick with "semantic mapping function."

As Frege also observed (1952, 13), different ways of filling in an unsaturated sentence schema correspond to different ways of *transforming* one sentence into another. The features defining a sentence jointly determine a set of transformations that can be performed on it (and others that *cannot* be performed on it, like substituting a finite verb for NP in "how dare NP VP?") Each admissible transformation also corresponds to a transformation on the side of content. The ways that the representations composing a lineage (an intersection of lineages) can be transformed into one another reveal the syntactic and semantic structure of the lineage's members.

Here, we return to the issue of how content is individuated, already broached briefly on p. 24 and p. 59. Two different representations can pick out the same content via different semantic mapping functions. Semantic mapping cuts finer than content. A waggle dance picks out the state of affairs it represents, perhaps that there is some nectar a hundred meters due east, via one set of rules. The sentence "there is some nectar a hundred meters due east" picks out the same state of affairs via a quite different set of rules. This difference is manifest in the different transformations these two representations permit.

For the waggle dance, possible transformations include changing the duration of the waggle, its angle, and the time and place of its tokening. These correspond to parallel changes in the nectar-location being described. My statement can also be transformed so as to describe other nectar-locations, by replacing the terms "100 m," "due east," and my own position when uttering it. But these, at least the first two, are not transformations of the same kind as in the waggle dance case. Rather than changes along a single continuous dimension, they consist in the wholesale substitution of discrete units. Moreover, there are possible transformations of my statement for which there are no corresponding transformations on a waggle dance, like replacing the term "nectar" with "coffee" or negating the sentence. "Theaetetus flies" and "Theaetetus instantiates flying" likewise differ in the rules according to which they map onto their joint truth-conditions, which can again be illustrated by the substitution test. By substituting "instantiates" for "hates" in the second sentence, we get a content that cannot be reached by word-for-word substitution transformations on the first sentence (LTOBC, 107-09, 2005b, 63–64). "The semantic mapping of a sentence articulates it, placing it in a logical space of contrasting possibilities. Its truth-condition is not, as such, articulated" (2005b, 64).

This means that semantic mapping, rather than content, is the appropriate level of semantic individuation at which to understand (at least some) hyperintensional phenomena in language. Supposing the world states picked out by teleosemantic contents to be intensionally individuated (i.e., to be identical iff they obtain in exactly the same possible worlds),[51] differences in content can't explain failures of intersubstitutivity in hyperintensional contexts.

As traditionally defined, a hyperintensional context is one where intensionally equivalent sentences (those that are true in exactly the same possible worlds) fail to be intersubstitutable *salva veritate*. One example of hyperintensional contexts is attitude attribution contexts, like "believes that…" So for instance, the sentences "Theaetetus flies" and "Theaetetus instantiates flying" are presumably intensionally equivalent—but they may not be intersubstitutable *salva veritate* in the context "Axel believes that…"

In this case, the difference is that Axel's belief that Theaetetus instantiates flying, but not his belief that Theaetetus flies, requires him to have the concept of instantiation. Let us recall the account of conceptual thought from section 2.3. As I mentioned there, referring terms can allow a hearer to recognize and learn about their referents. In fact, Millikan asserts that it is one function of a sentence to produce recognition of the referents of its component terms, i.e. to engage the interpreter's concepts of those referents.[52]

[51] Intensionally individuated facts are identical just in case they obtain in exactly the same possible worlds. If there are hyperintensionally individuated facts, they can be distinct despite obtaining in exactly the same possible worlds. For instance, it could be a metaphysical necessity that *A* is morally right just in case *A* is happiness-maximizing, still [*A* is morally right] and [*A* is happiness-maximizing] could be distinct, hyperintensionally individuated facts (see Nolan 2019 for an overview).

It is unclear to me whether teleosemantics can accommodate the idea that two representations can be intensionally equivalent—true in exactly the same possible worlds—yet be made true by distinct facts. That would require that these distinct, hyperintensionally individuated facts could figure, separately, in independent Normal explanations, which would in turn require a hyperintensional understanding of (causal) explanation (one that does not, for example, reduce causality to counterfactual relations defined in terms of possible worlds). I don't know how this would be done. At any rate, it strikes me as being somewhat contrary to the metaphysical spirit in which at least Millikanian teleosemantics is couched.

Hyperintensional individuation of facts is typically motivated by different concerns than those that primarily motivate the teleosemanticist, like how to account for mathematical discourse. Such issues pose independent problems for the fundamentally empiricist edifice of teleosemantics. Are mathematical statements representations? If so, how can a mathematical fact explain the success of anything? If not, what are they? These are all very good questions, and I believe a future expansion and refinement of the teleosemantic project must take them seriously. Like other forms of naturalist meta-semantics, teleosemantics works best when applied to representations of contingent, empirical states of affairs. Teleosemantics is, at bottom, a form of semantic empiricism, and like all such theories its weak points lie in its application to necessary truths, modality, and so on. However, this must be a different investigation from the present one. I will be assuming that the world states picked out by contents are intensionally individuated, without further probing the tenability of this assumption.

[52] In *LTOBC*, Millikan reserves the term "representation" for specifically this type of conceptually engaged representation (Millikan 1984, 96). The broader category which I have

Hence, to participate in Normally successful discourse, the interpreter must have concepts for the referents of the terms used,[53] as well as being able to understand what the sentence says *about* those referents. She must also know that the terms used *do* refer to those things, having stored that information under her concepts for the referents (if not, discourse can be an opportunity for her to learn this). The way a sentence or speech-act "engages" a hearer's concepts is by producing a propositional attitude that employs those concepts. In this way, then, the belief Normally produced by a given sentence will not only share its content, but also something we can think of as its conceptual or logical articulation.[54] This also explains why the two sentences behave differently in intentional contexts. Since they Normally produce different beliefs, it makes sense that they would also Normally attribute different beliefs (for more discussion of attitude attributions, see section 4.2).

Relatedly, semantic mapping, rather than content, is the appropriate level at which to understand logical or inferential relations between sentences. Sentences with the same content are not necessarily logically equivalent, as evidenced by the contrast between "Theaetetus flies" and "Theaetetus instantiates flying" (for example, the latter but not the former entails the existential generalization "something instantiates something.") But, at least to a first approximation, representations with the same semantic mapping function *are* logically equivalent.[55]

All this is to show that in Millikan's theory, it is the semantic mapping function rather than the content that plays many of the theoretical roles traditionally attributed to propositions. They determine the conceptual abilities required to understand (to be a Normal consumer of) the sentence, and they settle its logical and inferential relations (and those of the belief that is its Normal product). Since propositions are also frequently called "contents,"

---

called "representations" she calls "intentional icons." In later works, Millikan reverts to the more common terminology I have been using.

[53] Though if she lacks them, she can, as we have seen, form those concepts upon hearing a sentence that refers to them.

[54] Distinguishing sentences on the basis of their different semantic mapping function doesn't account for *all* failures of intersubstitutivity in propositional attitude contexts, however. The sentences "Cicero is bald" and "Tully is bald" have the same semantic mapping function. The strict Millianism of Millikanian teleosemantics—her view that the semantic values of referring terms are individuated by their referents—precludes any attribution of distinct semantic values to co-referring terms. To address this issue, Millikan goes meta-linguistic: belief-attributions say something, not just about the content of the subject's beliefs, but also about the names by which the subject would identify the entities that the beliefs are about. For details, see LTOBC, ch. 13.

[55] As in the previous footnote, this claim requires some qualifications with respect to co-referring names like "Cicero" and "Tully." "Cicero is bald" and "Tully is bald," at least on some ways of reckoning, are not logically equivalent. To spell out these qualifications in detail and draw out their consequences would require a sojourn into Millikan's theory of identity and of the epistemology of meaning which is outside the remit of the present thesis. I will rely on the approximation in what follows.

this fact is wont to cause some confusion unless clarified. We will have reason to return to this point later, in chapter 4.

To round off the picture, I should emphasize that not all parts of the sentence need contribute to its semantic mapping, i.e., the rules that pick out its content. A good example to the contrary is the indicative and imperative forms, which do not contribute to mapping a state of affairs but only determine how the hearer is supposed to relate to the state of affairs mapped by the sentence's other features. In the case of indicatives, the hearer is supposed to form a belief that the state of affairs obtains. In the case of imperatives, he is supposed to bring it about (LTOBC, 53–54). Hereby, teleosemantics recapitulates the traditional distinction between force and content.

In chapter 4, I will argue, tentatively, that the deontic "ought" also belongs to this class of non-representational linguistic devices that have stabilizing functions but don't help pick out any particular world state or correspond to any particular property.

This more or less concludes my overview of Millikanian teleosemantics, but before we leave this chapter I wish to briefly discuss a controversy in the teleosemantic literature concerning content indeterminacy. Since we will run into some indeterminacy problems in later chapters, especially in chapter 4, it is useful to understand how they arise within teleosemantics and what strategies are available for dealing with them.

2.5. Content Indeterminacy

So-called indeterminacy problems haunt naturalistic theories of intentionality. Probably the most famous of these is due to Saul Kripke in *Wittgenstein on Rules and Private Language* (Kripke 1984), who (following Wittgenstein 1953), presents arguments against two types of meta-semantic theory. The first and most Wittgensteinian of these strikes against what we may call "mere-similarity" theories of intentionality, according to which following the semantic rule for an expression is simply a matter of continuing to use it as before, to "go on in the same way." Against this type of view,[56] Kripke makes the simple point that there are innumerable ways of "going on in the same way." Indeed, *any* way of going on counts as going on in the same way according to *some* measure of sameness. Most of these measures will strike a normal person as impossibly gerrymandered or "gruelike"[57]—but then the

[56] To be precise, the target of Kripke's argument is the view that meaning something is a matter of *intending* to go on as before, but this subtlety is immaterial for our purposes. He who intends to go on as before intends his future use to be governed by his past use, and the problem, for us as well as for Kripke, is that the finite series of past uses cannot on their own govern anything.

[57] As the reader will recall, the property *grue*, which came to philosophical fame in Nelson Goodman's *A New Riddle of Induction* (Goodman 1983, chap. 3), is the property of being

onus is on the meta-semanticist to explain why gruelike ways of going on do not comply with semantic rules.[58]

Teleosemantics, too, maintains that representational correctness—following the semantic rules—consists in "going on in the same way," but it has a way of picking out a (more or less; see below) *unique* same way of going on that is the *right* one, namely, the one that has explained past evolutionary success. To take the standard example in the literature: a frog catches flies by detecting them visually and then flicking its tongue out towards them. Intuitively, what the frog visually represents is something roughly like *there is a fly over there*. And the visual mechanism involved has been selected because, sufficiently often, the representations it sends to the tongue-mechanism have been tokened in the presence of flies. It has not been selected because it has tokened them in the presence of *flee bees*—a predicate whose extension includes all flies and all BBs (spherical projectiles used in air guns)—even though it might in fact, on each one of the occasions that did contribute to explaining its past persistence, have tokened them in the presence of a flee bee (specifically, if there were no BBs in the frog's ancestral environment). Explanations are counterfactually supporting: the mechanism *would* still have been selected if it *had* detected flies, but it *would not* have been selected if it *hadn't* (even if it *had* detected flee bees, i.e., if the flee bees it had detected were all BBs. Presumably, the frog's digestive system is not adapted for metabolizing metal pellets). Hence, the frog represents flies and not flee bees: if it tokened the representation in the presence of a BB, it would misrepresent, even if so doing would constitute one way of "going on the same way" (Millikan 1990, 1993a).

This, at least, is the idea. In essence, to follow the semantic rule for an expression is to go on using it in the way that has explained past success. It is a quite pleasing idea (though not everyone buys it; see (Fodor 1990, chap. 3, 2008)). It accounts for the intensionality (with an 's') of semantic rules—the fact that stating them requires modal language—by tying them to explanations, another intensional phenomenon. And it also explains the "ruleness" of these rules, their normativity, at least insofar as it explains why it might be a *good idea* to follow them—explanations of past success (fallibly) predict conditions for future success (p. 53)—and suggests an account of why we expect one another to follow them (p. 64).

---

green and observed before some future time *t*, or blue and not so observed. One way for me to go on using the word "green" in the same way I have used it in the past is to use it for all and only *grue* things, which means that, when time *t* comes around, I start calling blue things "green." For discussion of the parallels between Goodman's problem and the problem Kripke attributes to Wittgenstein, see (Kripke 1984, 58–59).

[58] The second of Kripke's arguments is directed at dispositional theories of meaning, according to which the meaning of an expression is determined by the speaker's dispositions to use it. Here, Kripke's point is essentially that a disposition isn't a rule at all: we can be disposed to misuse an expression. Dispositionalism fails to account for the normativity of meaning (cf. section 1.2.1).

However, Karen Neander (1995) has advanced an influential line of argument against Millikanian teleosemantics which threatens to bring us right back into indeterminacy. Neander's criticism starts from the observation that there is always a number of different, biologically adequate descriptions of what a system is supposed to do or, equivalently, a number of different things that the system is supposed to do, ordered hierarchically by an in-order-to relation.[59] So, to return to our standard honey bee example, the waggle dance's consumer is supposed to bring bees to nectar, *in order* for them to bring back nectar to the hive, *in order* for them to feed the hive, *in order*—ultimately—to help the bees survive and procreate. All of these are valid descriptions of the waggle dance's function, but depending on which level we look at, we get different Normal conditions. A normal condition for getting the bee to nectar is that there is nectar at the indicated spot, but for the bee to bring back nectar, the spot must also be free from predators that would otherwise have eaten the foraging bee, and to feed the hive the nectar must be non-poisonous, etc.

So what is the content of the waggle dance? It looks like unless we can find a principled way of picking out *one* of this hierarchically ordered series of functions as the one relevant to content determination, we have ourselves another indeterminacy problem! But which level? If we, like Millikan (LTOBC, 100, 2004, 85), demand that descriptive content should pick out Normal conditions for *all* the consumer's functions (including, presumably, the most ultimate one: contributing to fitness), we get what appears to be an implausibly specific content, like "there is non-poisonous nectar at *x*, and also no predators, and…" Neander's own proposed solution is, in effect, to pick the *lowest* in the series of functions. Before the producer can do anything else, it has to respond to certain proximal stimuli in the environment, like a flying object projecting a path across the frog's retina. And Neander suggests that this is what the frog represents: small dark moving things, not flies (Neander 1995, 2013). But this, one may feel, makes the content *too* proximal. Don't we want to be able to say that the frog sometimes mistakes BBs for flies? And is there a principled reason not to go even deeper down, and have the frog represent (e.g.) the pattern of excitations on its own retina? (Neander 1995, 135–36; but see 2017, chap. 7; Schulte 2017)

Millikan has responded to Neander by insisting that the content of a representation must not only be a success-condition for that representation, but it must also be something the representation's producer *can actually detect* (Millikan 2004, 85–86, 2009, 404). More formally, there needs to be a Normal explanation of how the representation producer has been able to produce representations that correlate with their content, and this explanation must

[59] Sometimes the in-order-to relation is causal: the bee gathers nectar in order to feed the hive. Sometimes it is constitutive: the frog catches a fly in order to catch food. The argument, as far as I can see, is roughly the same in either case.

make reference to the fact that the producer has responded to some proximal stimuli that carry information about the content (cf. p. 49). This requirement, according to Millikan, excludes the sort of highly detailed and implausible content-attributions otherwise threatening ("there is non-poisonous nectar at *x*, and also no predators, and…") because the producer does not have access to information about these sorts of details (Millikan's response concerns specific examples, but I take it she intends her argument to generalize).

Millikan's response has been criticized by Manolo Martínez, who argues that producers *do* in general have access to information of this kind, even on the theory of information that Millikan herself favors (Martínez 2013, 337–41). And so the debate continues. It is difficult for me to make a detailed evaluation of all these arguments, and it would, at any rate, take us too far afield. And note that even if Millikan's appeal to information should work, Neander has given independent reasons to favor proximal content (see below), so we would still be left with multiple content candidates. At the end of inquiry, we might have a teleosemantic content-determination principle that is entirely determinate, independently motivated, and yields plausible contents across the full range of representations. Until that time, it is worth asking whether teleosemantics can live with indeterminacy.

We must recall that the kind of indeterminacy threatened by the functional in-order-to hierarchy *is not* the kind of indeterminacy foreshadowed by Kripke's argument. Kripkean indeterminacy is completely open-ended: there is *no* way of "going on" that can't, by some measure of sameness, be counted as a way of "going on the same way." Moreover, some of these content candidates will be mutually contradictory. "Green" can mean *green*, or it can mean *grue*, but an object observed after time *t* can't be both green and grue.

By contrast, what I will call "hierarchic indeterminacy" gives us a fairly well-behaved set of candidate contents that relate to each other in principled ways. They do not contradict one another: correctness in terms of the "higher" contents entails correctness in terms of the "lower." And these candidates all have something in common: they all give us a tool for explaining past success and predicting future success. Some of these explanations/predictions will be more detailed and more reliable. In other cases it will be easier to determine, empirically, whether the success-conditions obtain. But all will have cognitive value. All will play the predictive and explanatory role that we attributed to content on p. 53 above.

Perhaps semantic indeterminacy isn't so bad, then. What reason do we have for insisting on determinate contents? Here's one: content is supposed to give us a norm for success and failure for representations, but if content is indeterminate, there are several different candidate norms and hence no univocal standard of success or failure. On the other hand, it is not as if success itself is a univocal matter: success can often come in respects or degrees. We could therefore simply treat different content candidates as norms for different degrees or respects of success. We could even go so far as to suggest

that, rather than being indeterminate between a number of different candidate contents, teleosemantics determinately assigns a range of distinct contents to representations, where each is the norm for a different degree or respect of success.

Recall Neander's view that the frog represents small dark moving objects. We might have wanted our theory to entail that a frog that projects its tongue toward a BB misrepresents the BB as a fly and so makes a mistake. Neander, however, argues that whether or not there is some mistake involved here, it is not a mistake of the *representation producer*. The producer, we assume, is designed to produce a representation in reaction to certain proximal stimuli, and this is precisely what it does, even in the case where those proximal stimuli are caused by a BB. To attribute a mistake to the producer seems to imply that the producer is *inferential*, that it integrates information from a number of different sources (perhaps also that it can learn?). Since, by hypothesis, it isn't, attributing "high" content "bestows on the frog semantic capacities it doesn't have" (Neander 1995, 134). We thereby judge it by a norm that is not applicable to it.

On the other hand, it does seem unremarkable to say that a frog who catches a BB has failed, even if this failure cannot be blamed on its perceptual system. It is a failure due to bad luck, but no less of a failure for all that. If we want to capture this type of failure, we must assign "high" content. But as I have been arguing, there is no reason to choose: we can have both high and low content at the same time. They will just correspond to norms for different degrees or respects of success.[60]

The need for determinate content arises, I suspect, specifically in the case of *cognitive* representations, i.e., beliefs and related attitudes, and the speech-acts and sentences that convey them. These types of representations are more than just bearers of descriptive content. They have *truth-conditions*. Presumably, teleosemantics must at least assign determinate content to these representations, or it has failed to account for what is arguably its core *explanandum*, truth and falsehood.

It seems to me, however, that the prospects for assigning determinate content to beliefs—and, by extension, to the assertions and indicative sentences that convey them—are quite a bit better than for representations in general. Suppose, as we have done above, that the function of a belief is to enter into

[60] One may want to say: Intuition tells us that representations have determinate content, and the job of meta-semantics is to give conditions that pick out those contents that they intuitively have. If this line is used apropos of the sort of primitive, fairly automatic systems typically used as examples in the "teleosemantics and indeterminacy" debate, I think it is simply false. There is widespread agreement in the literature that some content-attributions are unintuitive, but I personally cannot bring myself to have strong intuitions about the exact content of a frog's fly-catching detector, or a bee's waggle dance (here, I have an ally in Papineau 1998, 5). I have the intuition that these states are *contentful*, an intuition that my proposal accounts for. But at any rate, the notion of content is a theoretical concept, and we should perhaps not expect our intuitions to be infallible guides (cf. Björnsson 2018, 272).

inference with other mental attitudes in order to allow the cognitive system to accumulate information about the referents of the constituent concepts, test the consistency of the belief set, and produce action aimed at satisfying desires and needs. In abstract terms, the function of a belief would then be to contribute to making the subject's cognitive system into an accurate map of the world and adapt its behavior to external circumstances. Speaking loosely, the function of a belief with a given content *P* would be something like *helping adapt the organism to the fact that P*. Given that the *functions* of beliefs are most perspicuously defined by reference to the condition picked out by their descriptive contents, it wouldn't be surprising if the teleosemantic analysis assigned them the content in terms of which their function is defined.

Indeed, given these assumptions about the functions of beliefs, the effectiveness of Neander's argument is diminished. There is no *further* thing that the belief that P is supposed to do, beyond adapting the organism to the fact that P. We can discuss whether the waggle dance represents *there is nectar*, or *there is non-poisonous nectar*, or *there is non-poisonous, predator-free nectar*, because for each of those added conditions, the dance will Normally work better. But my *belief* that there is nectar will not necessarily work better just because there are no predators close to the nectar. That depends, in familiar holistic fashion, on what else I believe and desire. I might be in the predator-killing business, and then I would *want* there to be predators by the nectar. Or I might believe that there are usually predators next to nectar, and then I may choose to avoid the nectar. Since a belief doesn't cause any *particular* behavior, but just adapts me in a generic way to a certain circumstance, its usefulness is contingent exclusively on the obtaining of that particular circumstance.

If this is right, we can hold out for the hope that even if the indeterminacy problem for primitive representations should prove endemic to teleosemantics, there is no corresponding problem for cognitive representations.[61]

[61] If it turns out to be wrong, we might still be able to hold out hope that teleosemantics can live with indeterminacy. Gunnar Björnsson (2018) argues, on teleosemantic grounds, that all beliefs have two kinds of descriptive content, one "concrete" and one "strategic," and suggests some strategies for dealing with the consequences (for more on Björnsson's argument and the motivation for his view, see p. 106). If, as I argued above, hierarchical indeterminacy can be profitably understood as a case of *multiple* contents rather than *indeterminate* contents, Björnsson's strategies may be applicable here as well.

To simplify, Björnsson suggests that even if a belief has several different contents, only one of these need count as its truth-conditions. If we combine this suggestion with my discussion above, we might be able to claim that though each of a belief's contents constitutes a norm for *some* kind of success, only one of its contents is a norm for a specific kind of truth-requiring success. While Millikan, as we have seen, claims that there is no specific kind of function that characterizes representations, there could still be specific kinds of functions that characterize *truth-apt* representations (not all descriptive representations are necessarily truth-apt, as I suggested on p. 25), and truth-conditions could be norms for proper performance of these truth-aptness-conferring functions.

Furthermore, it lies close at hand to tie the truth-aptness-conferring functions to *inference*, so that a truth-apt representation is one whose proper function involves interacting

2.6. Summary and Conclusion

In this chapter, I have given a summary of Millikanian teleosemantics. There is much more to be said, and I strongly encourage the reader who feels unsatisfied or wants to know more to consult the primary sources.

By way of summary, I would like to highlight some of the things I have discussed in the foregoing which will prove especially important as we proceed:

1. An entity's proper function is the way its ancestors have contributed to the historical persistence and proliferation of the lineage. A Normal explanation is an explanation for how the entity has historically made this contribution. Normal conditions are conditions cited in a Normal explanation.
2. Representations are produced by producers and serve to adapt consumers to external conditions. The descriptive content of a representation depends on the Normal conditions for proper performance of the consumer's function, as adapted to the representation. They are therefore Normal success-conditions for the representation itself
3. Speech-acts function Normally when they are cooperative, implying that speaker and hearer should both benefit. Though Normal conditions for speech-acts are not equivalent to agent-level norms, they often correspond to *pro tanto* norms for speakers.
4. Descriptive contents are individuated coarsely. Semantic mapping functions are the appropriate level at which to understand hyperintensional phenomena in language.
5. Teleosemantics is subject to indeterminacy problems. The severity of these problems is open to discussion.

The next chapter will serve as something of a bridge between this chapter's discussion of teleosemantics and the subsequent chapters' treatment of meta-ethical questions. I will discuss the phenomenon of hybrid representations, which promise to offer a *via media* between meta-ethical cognitivism and non-cognitivism. However, I will also discuss the potentially devastating possibility that teleosemantics entails that *all* representations are hybrid.

---

inferentially with other representations. If the inferential functions of beliefs are sufficient to pick out univocal norms of success for them, then these could be identified with their truth-conditions. This is of course a mere suggestion, and further investigation is needed to see if it is viable.

3. The Problem of Universal Hybridity

In the last chapter, I presented Millikanian teleosemantics, the theoretical framework I will be using in the coming chapters. This chapter introduces a problem for the theory, *the problem of universal hybridity*, which has been noted by Marc Artiga (2014). Artiga argues that teleosemantics entails that all representations are *hybrid*, meaning that they have both *descriptive* and *directive* content (directive content, the reader may recall from p. 24, is the sort of content possessed by desires and other directive attitudes).

Hybrid representations are not, as such, a novelty for teleosemantics. According to Millikan, there is a special class of representations called *pushmi-pullyu representations* that are characterized precisely by their hybrid nature: they serve at once to *describe* the world and to *direct* behavior. They "connect states of affairs directly to actions" (Millikan 2005d, 296). For an example, consider again the by-now familiar waggle dance. The waggle dance, I have said, describes the location of nearby nectar. But it also directs other bees to fly there. This is its function, the effect that has allowed past waggle dances to persist. The waggle dance doesn't give the bee mere information that it can use in any number of ways, like an assertion would. It demands action. The waggle dance is a pushmi-pullyu representation. It has both descriptive and directive content.

Other representations, presumably, are not hybrid. Propositional attitudes like beliefs and desires, as well as standard speech-acts like assertions and commands, seem to be *pure* representations. They have either descriptive content (beliefs, assertions) or directive content (desires, commands) but not both. There is no *particular* action a belief is supposed to cause, nor is there any *particular* state of affairs under which a desire ought exclusively to be tokened. This view accords with common sense and philosophical received wisdom, and Millikan is in agreement (1989a, 296).

However, if the argument presented by Artiga is sound, standard versions of teleosemantics, including Millikan's version, in fact entail that this commonsensical view is false. Teleosemantics, according to Artiga, instead entails that *all representations are hybrid*. Call this the "universal hybridity thesis," or UHT.

The purpose of this chapter is twofold. First, I want to address the problem itself. I will defend Artiga's argument against some possible objections, thus endorsing the conclusion that teleosemantics does indeed entail the UHT. But I will also defend the possibly counterintuitive claim that the UHT

is not, in fact, a great problem for teleosemantics. Teleosemantics can survive having the UHT as a consequence.

In fact, having the UHT in hand will prove a resource in later chapters, which is the second reason why I want to discuss Artiga's argument here. This is because the UHT can help us make sense of *normative judgments*. Normative judgments seem to have features reminiscent of both descriptive and directive representations. On the one hand, "they are classificatory, truth-evaluable, apt candidates for knowledge, and apt for inference" (Cuneo 2018), all properties that are reflected in the declarative form of the sentences used to express them. On the other hand, they seem to be, to some extent, motivating or action-guiding, and the speech-acts that convey them seem to possess directive force.

On the assumption that representations in general only have one content, it is difficult to give a coherent theory of normative judgments that accommodates all these intuitions (although people have certainly tried). With the notion of a pushmi-pullyu representation, teleosemantics was already committed to the falsehood of this assumption, and there have already been attempts to understand normative judgments as pushmi-pullyu representations (e.g. Sinclair 2012; Millikan 2005d, 176–78). Even so, normative judgments are propositional attitudes, so even if it is only propositional attitudes that are generally pure, a theory that treats normative judgments as hybrid will be somewhat awkward. It will require us to explain why normative judgments should be so different, and that might force us to take on some quite specific commitments regarding normative psychology. However, once we acknowledge that all representations are hybrid, the claim that normative representations are hybrid will go from precarious hypothesis to triviality, and any problems that this claim may entail are likely to be solvable with the same tools used to solve the general problems attendant upon the UHT.[62]

The present chapter is structured as follows. In section 3.1, I introduce the notion of directive content and explain how Millikanian teleosemantics accounts for it. In Section 3.2 I introduce the notion of pushmi-pullyu representations along with Artiga's argument that teleosemantics entails that all representations are pushmi-pullyu, i.e. hybrid. In section 3.3 I discuss some possible objections to Artiga's argument, and conclude that they are all inadequate. Finally, in section 3.4 I pose the question whether the UHT is such a bad consequence for teleosemantics, present a number of reasons to think that it is, and try to show why these reasons do not motivate a rejection or major revision of teleosemantics in the light of Artiga's argument.

[62] Of course, this leaves open the question what *distinguishes* normative representations from what we would normally think of as pure descriptive or pure directive representations. I will make some suggestions in the chapters to come.

3.1. Directive Content

In the last chapter, I spoke exclusively about descriptive content, the kind of content that gives a representation "mind-to-world direction of fit." But as I indicated already on p. 24, there is at least one other kind of content: *directive* content, the sort of content possessed by desires, intentions, commands, and so on, which gives them "world-to-mind direction of fit."[63] The descriptive content of a representation, according to Millikan, specifies Normal conditions for the successful performance of the consumer's proper function, as adapted by the representation. As we will see, the directive content of a representation specifies conditions for *its own* successful function performance (LTOBC, 100; cf. Papineau 1984, 562). Notably, these two analyses line up rather nicely with the traditional terminology of two "directions of fit" between mind and world.[64] First, the system must adapt to the world, in order then to be able to adapt the world to its own needs.[65]

According to Millikan, the directive content of a representation is determined by its *focused proper function* (LTOBC, 99–100). We know what a proper function is (p. 43), but what is a *focused* proper function? In general, an entity can have a number of different functions, and they are not always clearly individuated. In particular, if the representation's consumer is supposed to cause some given state of affairs, and that state of affairs is itself supposed to cause a further downstream state of affairs, and so on, all of these states of affairs are functions of the consumer. A *focused* proper function is, roughly, the last in such a series of stringed-together functions before it splits apart, i.e., before there is no longer any *determinate* thing that is supposed to happen next (LTOBC, 34–38).

The Normal explanation for how a consumer, as adapted by a descriptive representation, can perform its job makes reference to a certain relation that must hold between the representation that adapts it and the state of the environment. Likewise, the Normal explanation of how a consumer of *directive* representations performs its job makes reference to a certain relation that must hold between the representation and the outcomes it produces. In each

[63] The distinction between two directions of fit is usually attributed to (Anscombe 1985, 56), though she doesn't mark the distinction using this terminology.

[64] It would perhaps be better to call them "representation-to-environment" and "environment-to-representation" direction of fit, because a representation doesn't have to be mental, and the environment can be intra-mental.

[65] Could there be more kinds of content? If every kind of content specifies a direction of fit between mind and world, it is hard to see how: there can only be two directions between two items. However, a third possibility might be *suppositive* content, the kind of content possessed by supposings, wonderings, and pure imaginations, which are not supposed to fit anything in particular. I believe, though, that the best treatment of suppositive content within teleosemantics would be to treat it as simply the descriptive content of a representation that is "run offline," in a manner analogous to that in which assertions can be used in fiction with no intention of conveying information, while still in possession of their conventional descriptive content (cf. Millikan 2004, 81).

case, this relation specifies a semantic rule for the representation. And just like the Normal mechanism whereby descriptive representations are produced must generate a non-accidental correlation between representations and the states they map onto, so that a Normally produced representation is a natural sign of the world state it represents, so Normally produced directive representations must be natural signs of their represented, meaning that they must carry information about *future* states of affairs.

In several respects, then, the Millikanian theory of directive representations runs parallel to the theory of descriptive representations. We shouldn't be surprised, therefore, if some of the same problems afflict both theories. And indeed, the theory of directive representations seems to suffer from analogs of the indeterminacy problem for descriptive representations that I discussed in section 2.5. Below, I review three potential sources of indeterminacy for directive contents.

First, quite plainly, unless we have a principled way of picking out *one* link in the causal chain that the representation is supposed to produce as the thing it *represents*, we get an indeterminacy problem. Picking the *focused* proper function constitutes such a principle, but is there a deeper justification for this choice?

The reason to favor the focused proper function, I think, is that it is more explanatorily central than any earlier or later link. Earlier links tend to be relational. A desire combines with other desires and beliefs in a complex inferential process in order to produce a given result. The inferential process varies depending on the subject's other beliefs and desires, but the aimed-for end result remains the same. To produce this *invariant* result is the desire's *focused* proper function. Since it is invariant, it constitutes a more compact, informative characterization of the overall mechanism by which the desire Normally contributes to fitness (cf. Schulte 2019, 167). As for later links, these, by definition, tend to diverge, so to mention any *single* link gives us at most a partial view of the desire's Normal contribution to fitness, and to mention them all gives us a disjunctive mess.

But *second,* as Peter Schulte has pointed out, there are indeterminacies that are unrelated to the choice between links in the *causal* chain, but derive simply from the possibility of describing a given effect in different ways. To return to our dear friend, the frog (p. 73): supposing that the frog has a motivational state that combines with its perceptual state to cause his fly-catch response, is the focused proper function of this motivational state to bring about the catching of *a fly*, or of a *non-poisonous* fly, or of *frog nutrition*, etc.? Each of these descriptions is a description under which the resulting effect can explain the ancestral persistence of the motivational state, and hence a candidate for the focused proper function—thus the descriptive content—of the fly-catch response (Schulte 2019, 165).

Since this argument is analogous to Neander's argument about descriptive content (p. 74), we might expect Millikan's response to be likewise analo-

gous. If there is no mechanism whereby the frog could reliably *detect* non-poisonous flies, so as to raise the probability that a fly it represents is non-poisonous over the probability that an arbitrary fly is, then the frog would not represent the fly's non-toxicity. Outcomes that obtain by luck are not represented. We noted, however, that the jury was out on Millikan's response to Neander, and similar uncertainties pertain to this argument.

It is possible that we could respond to Schulte's concerns in the same way as we did with the analogous concerns about descriptive content in section 2.5: simply bite the bullet and accept that at least simple representations might have indeterminate directive content. In the case of descriptive content, the trick was to make sure that the proposal did not generalize to cognitive representations. Is there a corresponding concern here?

Imperative sentences come high on the list of directive representations that it would be good to be able to assign determinate content to. Doing so requires that we can make plausible, not only that it is a function of an imperative to produce its intuitive satisfaction-conditions (which is what Millikan argues, and which we may treat as given), but that there are no *other* functions of the imperative about which some of Schulte's indeterminacy worries could be raised. The most promising response is, again, analogous to the one we gave in the descriptive case. Just like there is no *specific* value to having a belief beyond its capacity to adapt us to a certain state of the world, there is no *specific* further point to uttering an imperative beyond having it complied with. Imperatives may have persisted because of their ability, by way of producing compliance, to satisfy various further ends that speakers and hearers have had (LTOBC, 57). But as long as the same imperative construction has been used to further *different* ends that were satisfied by compliance, these further ends are only alternative further functions, not the focused proper function of the construction. On any particular occasion, we will of course typically have a reason for uttering an imperative. But this is not a further purpose that the *imperative*, as a token of its sentential type, has in virtue of its ancestry. It is only a purpose we have in uttering it.

Can an analogous line of reasoning be applied to desires and other directive propositional attitudes? I am less sure about that. We typically attribute desires using infinitive phrases, such as "his desire to have a nice sports car," and this might create the impression that they have determinate satisfaction conditions, picked out by the semantic rules of the phrases used to attribute them. But this attribution practice may hide psychological complexity. My desire to have a nice sports car and your desire to have a nice sports car may very well be quite different, formed by different processes and having different Normal ways of producing their outcomes. In particular, if the explanation for why I desire a nice sport-car is that I have come to expect that it will impress people, whereas the explanation for your desire is your longing for the sense of speed and excitement you expect such a vehicle to provide, then perhaps these further goals or motivations that explain our

respective desires should enter into their directive contents? After all, if I get my sports car but don't impress people, the sports car itself will be of little joy to me. I may come to resent it as a symbol of my inadequacy.

These are mere speculations. I advance them only because I think that teleosemantics could survive their truth. Our practice of attributing desires using infinitive phrases would still make sense, because you and I would still share a motivating attitude that had, as *part* of its function, to make us acquire a sports car. Ascribing our respective desires using the same phrase thus captures a similarity between our respective psychologies, one that derives from the teleological features of our respective motivational systems.

With these observations about directive content in hand, let us turn to look at the UHT and Artiga's argument.

3.2. Hybrid Representations and Universal Hybridity

As we saw in the introductions, what Millikan calls "pushmi-pullyu representations" are representations that possess both descriptive and directive content: they serve both to describe the world and to direct behavior on the basis of that description. Millikan introduces pushmi-pullyu representations as a way of characterizing what is *unique* about the language and advanced cognition of humans (and perhaps some of the higher animals). In her 1989 paper "Biosemantics," she enumerates six characteristics of human mental representations that set them apart from those of lower life-forms, one of which is our possession of "indicative and imperative representations." She asserts that "Simple animal signals are invariably both indicative and imperative" and goes on:

> The step from these primitive representations to human beliefs is an enormous one, for it involves the separation of indicative from imperative functions of the representational system. Representations that are undifferentiated between indicative and imperative connect states of affairs directly to actions, to specific things to be done in the face of those states of affairs. Human beliefs are not tied directly to actions. Unless combined with appropriate desires, human beliefs are impotent. And human desires are equally impotent unless combined with suitable beliefs. (Millikan 1989a, 296)[66]

In the grand scheme of nature pushmi-pullyu representations are the norm, while pure descriptive and pure directive representations constitute the human exception.

[66] In this text, Millikan has not yet introduced the terminology of "descriptive" and "directive," using "indicative" and "imperative," respectively, in their stead. Since "indicative" and "imperative" already have a use denoting grammatical categories in language, I think the terminological change is an improvement.

Consistent with this view, the examples Millikan gives of pushmi-pullyu representations are, with a few notable exceptions (see Millikan 2005d, 176–85), of rather primitive signaling systems. We are already familiar with the waggle dance. Other examples include the food call of a mother hen to its brood (it describes the location of food, and directs the chicks to come there), animal warning calls (they signals the presence of danger and directs conspecifics to seek protection), and so on (Millikan 2005d, 173–74).

One needs not be a teleosemanticist to appreciate what is distinctive about these examples. As Millikan says, they connect a specific state of affairs directly to a specific action. For each such signaling system there is a mapping, describing the normal operation of the system, from states of the environment to specific signals and from specific signals to specific behaviors.[67] Danger at location *s*, time *t* leads to warning-call at *s*, *t* leads conspecifics to hide at *s*, *t*.

So far, so good. PPRs seem like a useful and interesting addition to our inventory of intentional types. They form a class of hybrid representations that shows, by contrast, what is distinctive about the pure representations of human talk and thought. However, an argument due to Marc Artiga (2014) purports to show that standard versions of teleosemantics, including Millikan's version, in fact entail that pushmi-pullyu representations do *not* constitute a distinct class of representations but that, on the contrary, *all representations, necessarily, are hybrid*. Call this the "universal hybridity thesis," or UHT.

A common reaction to the UHT is that it must be false.[68] If so, then either Artiga is right that teleosemantics entails the UHT, and then teleosemantics must be revised or abandoned, or he is wrong, and then we should show why. Indeed, Artiga himself takes this to be the lesson of his argument, and proposes a revision of teleosemantics in response. I, however, disagree. I believe (*contra* Millikan) that Artiga is right: teleosemantics does entail the UHT (or, at least, a weakened form of the UHT that is still sufficiently strong to be worrisome (see section 3.3.1 below)). And I also believe (*contra* Artiga) that this state of affairs requires neither the revision nor the abandonment of teleosemantics.

Artiga's argument in itself should not be difficult to grasp at this point. Artiga begins by recapitulating the teleosemantic principles of content determination.[69] These, he claims, in formulations that should be familiar by

[67] Since a system of representations is always contrastive, even the absence of an overt signal may count as a signal in the relevant sense. If a beaver splashing its tail at time *t*, position *s* means (describes) something like "danger at *t*, *s*", a beaver *not* splashing its tail at *t*, *s* means something like "no danger at *t*, *s*"—at least provided the beaver is in a state where Normally, it would splash if there were danger (i.e. not sleeping, etc. I'm grateful to Gunnar Björnsson for stressing the latter point).

[68] That seems to be Artiga's own reaction, and I have had similar reactions from many of those with whom I've casually discussed the possibility.

[69] Like me, Artiga focuses specifically on Millikanian teleosemantics (2014, 546).

now, identify the directive content of a representation with its focused proper function and its descriptive content with the most proximate Normal conditions for successful consumer response (Artiga 2014, 550, 552).

Now Artiga makes the simple observation that for every function, there are Normal conditions for the successful performance of that function; and for every Normal condition, there is a function *for which* they are the Normal conditions. That much just follows from the definition of Normal conditions (p. 44). He then makes the further observation that if everything that has a function has Normal conditions and vice versa; and if the function of the consumer as adapted by the representation determines the representation's directive content, and the Normal conditions for the adapted consumer response determines its descriptive content; then everything that has descriptive content must also have directive content, and vice versa. Hence, teleosemantics entails the UHT (Artiga 2014, 554–56).

Can it really be that simple, or has Artiga missed something? Below I will consider a couple of potential rejoinders against Artiga's argument on behalf of Millikanian teleosemantics.

3.3. Potential Rejoinders

3.3.1. Desires

Consider first the argument as applied to the case of desires. As Artiga acknowledges, not just any Normal condition will be part of the descriptive content of a representation, but only the most proximate Normal conditions (cf. LTOBC, 99). That is not sufficient to avoid the conclusion that desires have descriptive content, however, since for any set of Normal condition, if the proximity metric is to have any meaning at all, there must be some that are *most* proximate.

But as we have seen (p. 49), being part of the most proximal Normal conditions for successful functioning of a representation is not sufficient for *P* to be the content of that representation. In addition, the representation must have been produced by a producer that Normally performs its function by maintaining a certain relation (mapping) between the representations it produces and the world, a relation that obtains for the present representation just in case *P*. And this requires the producer to Normally be causally sensitive to natural signs of whether *P* obtains.

Could it be that this constraint is not met by desires? If so, the ancestral persistence of the desire-forming mechanism could not be due to its sensitivity to conditions that have explained the success of the desires it has produced. In effect, it would have produced desires at random, and they would

just have happened to be produced under conditions that made their effects beneficial sufficiently often to explain its persistence.

As a claim about desires, this seems wildly implausible. Intuitively, desires adapt us to pursue beneficial ends (beneficial in an evolutionary sense, of course), and when they are successful, it is Normally because the ends they make us pursue are indeed beneficial. It would be very strange to learn that the desire-forming mechanism is not, in fact, Normally sensitive to whether the outcomes promoted by the desires it forms would be beneficial. Indeed, one may think that the entire weight of behavioral psychology for the last hundred years is that our motivations are formed by mechanisms designed by evolution to exploit correlations between behavioral outcomes and certain hard-wired primary reinforcers (like food, sex, etc.) that presumably possess this status because of their past contribution to fitness.

This observation about how desires in fact work is not sufficient, of course, to establish that there *couldn't be* directive representations formed by a process insensitive to their Normal success-conditions. Hence, it is not sufficient to establish that the UHT—the thesis that all representations *necessarily* are hybrid—follows from teleosemantics. What it *does* establish—that desires have descriptive content—is perhaps bad enough. But can we avoid the stronger conclusion?

Suppose there were some device that generated various items at random—a "guessing system" (Price 2001, 94)—which items then in turn produced, by interacting with further systems, various behaviors, a specific behavior for each item. Suppose further that these behaviors then went on to yield evolutionary success sufficiently often for the device to be reproduced. The guessing system is introduced by Carolyn Price (2001, 94 ff.) to make the argument that Millikanian teleosemantics has to assign *descriptive* content to the items produced by it, a claim Millikan explicitly rejects. Millikan's rejoinder, to put it briefly, is that the guessing system would not have been selected because the items that it produced carried information discriminately about the conditions that explained their success, which is required for a system to be descriptively representational (cf. p. 49) (Millikan 2007, 444–45). But it follows, of course, that unless there is an analogous reason to deny *directive* content to the guessing system's products, those products are pure directive representations.

It seems to me, however, that the guessing-system would not have been selected because the items it produces carry information about the behavioral outcomes they engender. It *does* carry information about those outcomes, provided they produce them by a reliable mechanism, but it doesn't follow that this fact has contributed to the explanation of the system's past persistence. This informational relation would seem to be wholly incidental to the past success of the guessing-system, since its successes have been attained at random. Hence, it seems, the products of the guessing-system do not directively represent their causal outcomes.

If these arguments are right, the verdict remains that Millikanian teleosemantics is committed to the UHT. However, even if I am wrong and guessing-systems can produce directive representations, the upshot is only that the theory avoids the UHT in the strong form I have articulated. It isn't *necessary* that all representations are hybrid. This is poor comfort, since it is highly implausible that any of the directive representation kinds we are typically interested in and intuitively take to be examples of pure directive representations—desires, intentions, commands, etc.—are produced by guessing systems. Even if we could thereby save teleosemantics from the strong UHT, we will still have to contend with its weakened version: that necessarily, all representations *except* those that are produced by guessing systems are hybrid. This weakened thesis may seem sufficiently bad to merit continued pursuit of an argument to refute Artiga.

3.3.2. Beliefs

Turn to beliefs. Can it be that beliefs lack directive content because they lack *focused* proper functions? Recall that the focused proper function of an entity is the *last single determinate* thing it is supposed to do. Every subsequent effect is just one alternative function among many. But beliefs don't seem to have *any* single thing they are supposed to do. Having a true belief is good for a lot of things. It allows you to plan your actions in ways that potentially yield all kinds of benefits. But any plan you make with the belief's help, and any benefit you secure thereby, is merely one alternative use you can make of it. Is there really a *single thing* accomplished by successful beliefs that gives rise to all further alternative outcomes? If there isn't, then we could conclude that beliefs possess no focused proper functions, and hence no directive content.

Artiga considers a related rejoinder on behalf of the teleosemanticist. Perhaps there is no *particular* thing that a belief is supposed to do, or help its consumer do: it is just supposed to generically help its consumer. In that case, there would be nothing in particular that could qualify as the belief's directive content (Artiga 2014, 557). To this, Artiga replies that if beliefs are to have *descriptive* content, there *must* be some particular things they are supposed to do. Since descriptive content depends on Normal conditions for successful consumer response, if there is no particular thing that counts as a successful consumer response for the belief, we could assign no determinate descriptive content to it either (2014, 558). This is a convincing reply to the rejoinder as Artiga formulates it, but it may seem to miss the stronger rejoinder in the vicinity: that even if beliefs have determinate functions, these functions are highly relational: there is no *single* outcome that the belief is supposed to produce, regardless of circumstances. Descriptive content can still be assigned to beliefs on the basis of their relational functions, but since

they are not in the business of producing a single determinate outcome, they don't have directive contents.

But a specification of the function of a belief doesn't necessarily have to take disjunctive form. In our discussion in the last chapter, I supposed that the function of a belief is to enter into inference with other mental attitudes in order to let the subject learn about the referents of the constituent concepts, test the consistency of the belief set, and produce action aimed at satisfying desires and needs. I said that in abstract terms, the function of a belief is to contribute in its own special way to making the subject's cognitive system into an accurate map of the world and adapt its behavior to external circumstances. Could we not say, then, that the *focused* proper function of a belief with a given content *P* is something along the lines of *contribute thus-and-so to adapting the organism to the fact that P*, where "thus-and-so" is to be filled in with details about the particular mechanisms whereby a belief accomplishes this task, e.g. testing the likelihood of other beliefs and the viability of action plans via inferential processes? To call this a *proper function* of a belief that *P* seems unobjectionable, and it correctly predicts that the descriptive content of the belief will indeed be *P*, since something that adapts the organism to *P* will presumably tend to be successful, barring a fluke, only if *P*. To go a step further and call this a *focused* proper function of the belief that *P* will perhaps require a certain liberality with what we count as a "single effect." But it seems difficult to find a plausible principle for individuating effects that would allow us to deny this conclusion.

This objection, nevertheless, is stronger than the previous one. It is not inconceivable that one could articulate a principled and independently motivated reason to deny that beliefs have a focused proper function, or give alternative accounts of directive content that avoid the conclusion (see Shea 2018, 189 for one proposal). Here, I propose to leave this issue to the side. Even if we could show that beliefs lack directive content, we would still have to deal with the fact that teleosemantics seems to assign descriptive content to desires and other motivational states. And if we can assuage worries about the UHT, we may not need to answer Artiga's argument.

3.3.3. Pre-Content Type-Individuation

Is there another way to evade Artiga's conclusion? Hitherto, we have been assuming that there is one type of entity, "representations," and a set of content-determination principles that guarantees that if a token entity of this type has a given (historical) property it will thereby *ipso facto* have content of a given kind. But perhaps this is the wrong way of reading what Millikan says. Consider the following passage from *LTOBC:*

> There are *two* paradigms of intentionality: an indicative paradigm and an imperative paradigm. And these display important differences as well as striking simi-

larities between the ways "the sign" is properly and Normally related to "the signed." (LTOBC, 86)

Here, Millikan seems to suggest that we should *first* distinguish types of signs and *then* articulate content-determination principles for each type. And indeed, an assumption of this sort structures her presentation of Chapter 5 of *LTOBC*. Millikan begins by distinguishing the two types of representations, indicative and imperative ones, and then gives content-determination principles for each one of them.

It would seem possible for teleosemantics to avoid committing itself to the UHT if there were some way of defining the categories of descriptive and directive representation *independently of* the kind of content they have and then define the two kinds of content as the ones paradigmatically possessed by descriptive and directive representations, respectively. For instance, we could try to distinguish descriptive representations from directive ones based on the *kind* of function they have. This is a natural thought: it does seem as though (what we would intuitively identify as) directive representations have functions that are more directly tied up with bodily behavior and with concrete effects on the environment. But note how *ad hoc* the solution seems. Both kinds of representations will have functions, so how do we motivate the principle that function only yields directive content for one of them, other than by the *post hoc* reason that it would make the account line up better with our intuitions? And since the proposed distinguishing mark—the extent to which the function concerns bodily behavior and the external environment—is a vague affair, either concept-possession would have to be so as well, or we would have to draw a line somewhere, on a basis that couldn't help but be arbitrary.

There might be some other way of distinguishing descriptive and directive representations antecedently of content-determination, but it is not evident what that would be. In the absence of a suggestion, I must conclude that content-determination principles ought to apply equally to all kinds of representation.

3.4. Learning to Live with the UHT

I tentatively conclude that Artiga's argument is successful: teleosemantics entails the UHT, at least in its weakened version.[70] This raises two questions: 1) is that so bad, and 2) if it is, what should we do about it? Below, I enumerate a number of problems that the UHT could be thought to entail for

[70] I have not mentioned all the objections against the argument that Artiga himself considers or his replies to them. As far as I can tell, these replies are successful. A reader interested in saving teleosemantics from the UHT may want to take a look at them, however.

teleosemantics. I have tried to be exhaustive, but there may certainly be some problem I have missed.

First, the thesis means that we lose an elegant and intuitive way of distinguishing different kinds of representations. We might have thought that the distinction between descriptive and directive content was a natural way to capture the difference between beliefs and desires, or between assertions and commands, but the argument shows this to be mistaken. Unless we want to fundamentally revise teleosemantics, then, we must find some other way to account for these differences.

Second, the UHT means that the notion of *pushmi-pullyu representations* is necessarily coextensive with that of representations *simpliciter*, and so cannot play the theoretical role envisioned for it, i.e., to account for what is special about primitive representations and so also, by contrast, what is special about advanced human representations. Again, to save teleosemantics from major revisions, we need another way to account for these differences.

Third, and relatedly, we intuitively think of beliefs as "descriptive," desires as "directive," and the kinds of primitive representations Millikan calls "pushmi-pullyu representations" as "hybrid." What explains these intuitions, if representations of all three kinds are hybrid?

Fourth, we identify beliefs and desires via that-clauses ("the belief that it is raining") or infinitive clauses ("the desire to dance in the rain") that supposedly pick out these attitudes by their content, i.e., their *unique* content. What explains this practice, if beliefs and desires have several different contents? Why can't we pick out a desire by its descriptive content?

Fifth, if desires and commands have descriptive content, doesn't that mean that they have truth-conditions? If so, why can't we assign truth or falsehood to them?

In the face of these problems, some will want to consider the option of revisionism: reformulating teleosemantics so that it doesn't entail the UHT. What might such revisionism look like? Though Artiga doesn't endorse any particular way for teleosemantics to deal with his argument, he suggests in passing that the theory might be revised so as to countenance descriptive content only (2014, 552). This, however, strikes me as an unsatisfying solution. It generates similar problems as the UHT itself: it leaves us without tools for accounting for the difference between descriptive and directive representations, it defeats the theoretical purpose of pushmi-pullyu representations, and it makes a mystery out of our practice of identifying desires by infinitive clauses (because surely, these infinitive clauses do not identify desires by their *descriptive* content?).

There may be other revisionary options, but I will not further entertain the possibility. As stated, I believe that the UHT is an acceptable consequence and that no revisions are necessary. To support this contention, I will explain how teleosemantics can deal with each of the five problems listed above.

The first three problems concern how teleosemantics can account for the intuitive difference between descriptive and directive representations (beliefs and desires, assertions and commands), as well as between these advanced cognitive attitudes and more primitive forms of representation. If the UHT is true, we lose the ability to account for them by the *type* and *number* of contents that define the different categories. Luckily, there are many other options available for teleosemantics.

Even without the UHT, we would have needed ways to individuate representation-types that went beyond the types of content they have. On the conventional picture, assertions and beliefs are both descriptive, but they belong to different types. And it is a straightforward matter to account for the difference. They are produced in different ways, are instantiated in different media, and, crucially, have different functions and different Normal ways of performing these functions (an assertion is supposed to produce a belief, but a belief is supposed to help guide inference and behavior).[71] Likewise, intentions and desires—in the colloquial, rather than the generic philosophical sense of "desire"—are both directive mental representations. What distinguishes them presumably has something to do with the way they are produced (intentions are formed on the basis of desires, rather than vice versa) and the way they interact with the rest of the cognitive machinery in order to perform their functions (intentions have a more direct access to action than do desires).

If there is a problem here, it has to do with why we intuitively think of beliefs and assertions as "descriptive," desires and commands as "directive," and some representations (like those Millikan originally identified as PPRs) as "hybrid." This was the third problem listed above.

It may seem fairly obvious why we would think of a belief as descriptive, even if it has a focused proper function roughly like *contribute thus-and-so to adapting the organism to the fact that P*. This, as we have already pointed out, is a very generic function. Even to be able to pick it out using a relatively compact description, we had to describe it *in terms of* the belief's descriptive content *P*. Indeed, it makes sense to say that for a representation to *be* descriptive *just is* for it to have the function of adapting the organism, in a generic manner, to a certain world affair.

Likewise, even if a desire has a descriptive content, this descriptive content is likely to concern states of affairs that are directly relevant to action in a way that the descriptive content of beliefs typically are not. If the desire-forming mechanism, as seems plausible, has been selected due to its sensitivity to the benefit (in an evolutionary sense) attendant upon various action alternatives, then the descriptive content of a desire is, as it were, a mere reflection of its practical import.

[71] Cf. Millikan, "to be a belief involves having certain kinds of proper functions" (1984, 138).

But ultimately, it isn't the generic nature of a belief's directive content that makes it descriptive, or the practical significance of a desire's descriptive content that makes it directive. We could form the *belief* that an action alternative would be beneficial, and this belief would not for that reason be a directive attitude. The descriptiveness of beliefs and the directiveness of desires must, I think, be explained by reference (again) to their functional role, their Normal way of interacting with other attitudes and the rest of the cognitive machinery. The belief that φ-ing would be beneficial can be used to *reason*: to draw conclusions about φ-ing and about benefit, to test the consistency of the belief-set and of the concepts of φ-ing and benefit themselves. This reasoning can be engaged with in a dispassionate frame of mind, one that doesn't directly engage us to go out φ-ing. In contrast, the desire to φ can presumably not be used to reason about φ-ing or about benefit. It may not actually engage the *concept* of benefit at all, even if recognizing and reasoning about its descriptive content requires this concept.

All this is to say that our intuitions about what is descriptive and directive may best be understood as deriving from person-level phenomena, pertaining to the whole ecology of mental life, and are as such badly understood in terms of notions like those of descriptive and directive content which, by design, are applicable to the whole range of intentional phenomena, including sub-personal cognitive representations, simple animal warning calls, and so on.

Similarly, the simple representations that we intuitively think of as "hybrid," that Millikan tries to capture with the notion of pushmi-pullyu representations, may be better understood in terms of the ecology of the simple creature's mental life, its relative lack of dedicated inferential processes that operate on those representations, leading to a relatively direct tie between tokened representation and behavior.

Let us now turn to the fourth problem, how to explain our practice of attributing attitudes using phrases that pick out one, but not the other, of their contents. The question is what the alternative to this practice would be. Does the UHT lead us to expect that we should be able to attribute desires, for example, via indicative sentences that pick out their descriptive contents, so that we could say "he desires that φ-ing would be beneficial" meaning that he desires to φ? For this practice to make sense, we would need to *know* that the descriptive content of the desire to φ is *φ-ing would be beneficial* (if indeed it is). However, there is no reason to expect that we should know that, especially not if the descriptive contents of desires, as argued above, are not conceptualized. As the next chapter will illustrate, deriving the descriptive content of directive attitudes is far from trivial.

Moreover, it is plausible that our attribution conventions reflect the linguistic means we conventionally use to influence each other's attitudes. Indicatives are used to convey belief, and they are also used to attribute belief. According to Millikan, an embedded sentence attributes precisely the atti-

tude it is supposed to produce (LTOBC, 212; cf. section 4.2). If we had possessed indicative forms that Normally performed their stabilizing function by producing desires, we would probably also be able to attribute those desires using the same indicative forms. But we don't.

This leaves us with a single remaining issue: if directive representations have descriptive content, why can't we attribute truth and falsehood to them? I will propose two explanations for this: one relatively superficial, the other somewhat deeper. The superficial reason is simply that directive representations are attributed using, not full sentential complements, but infinitive phrases: "the desire to run," "the command to get me a glass of water." These infinitive phrases do not embed under sentential operators like "it is true that…," and neither do their infinitival complements:

(1) # It is true to bring me a glass of water.

(2) # It is true that bring me a glass of water

Desires are sometimes attributed using optative constructions like "the desire that he give me chocolate on Valentine's Day," but these aren't felicitously embeddable under "it is true that…" either:

(3) # It is true that he give me chocolate on Valentine's Day.

Certainly, we can attribute truth to something without having that something represented by a full sentential complement, as we do when we use a construction with a demonstrative or a description plus "is true." One might therefore ask why the following exchange should not be felicitous:

Axel: "Hey, Ingemar! Bring me a glass of water!"

Konrad: # "That is true."

But it seems plausible that in order for a use of one of these constructions to be felicitous, it must be possible to give an answer to the question "*what* is true?" that substitutes a full sentential complement for the "what." In dialogs like the one above, it is unclear what that would be. The context does not supply an obvious candidate.

This is my superficial explanation of why we can't attribute truth to desires. It may feel a bit too superficial, however. It leaves unaddressed *why* our language should be constructed in such a way as to make these constructions infelicitous. That is clearly a very complex question, so I will settle for outlining a tentative answer. So long as there *is* a candidate explanation of this fact about language compatible with the UHT, the latter remains in a dialectically good position.

Truth is intimately tied to inference. This is evidenced by the fact that the truth-apt sentences coincide with those that can be embedded in the antecedents of conditionals. A conditional conditionalizes the consequent on the *truth* of the antecedent. In Michael Dummett's words, "the notion of truth is tied to the condition which the antecedent expresses" (Dummett 1973, 351). An imperative or an infinitive clause, however, can as little be embedded in the antecedent of a conditional as under "it is true that…":

(4) # If bring me a glass of water, then I'll be happy.

In itself, this observation only kicks the bucket down the road: *why* can't they, if they (attribute attitudes that) have descriptive content? To answer this, let us bring on board Millikan's Ryle-inspired theory of conditionals. Her proposal is that a conditional has as its stabilizing function to produce a disposition: a disposition to token the attitude Normally conveyed by the consequent (unembedded) *contingent on* her harboring the attitude Normally conveyed by the antecedent (unembedded) (Millikan 2018, 236).

Suppose this is correct. Suppose further that this disposition is one that is Normally implemented by the inferential machinery of the cognitive system, i.e., whatever mechanisms in us are responsible for testing the consistency of our belief-set, drawing conclusions from our existing beliefs, and combine these beliefs with desires in order to produce adaptive behavior.

Now, the conventional understanding of desires is that insofar as they enter into mental inferences, it is only as practical premises or conclusions. It may, in other words, be that the functional role of desires precludes that the cognitive system should be sensitive to the intentional relations of their *descriptive* contents. It may be that while desires possess descriptive content, this content is inferentially inert. This would be consistent with my suggestion above that the descriptive content of a desire is not conceptualized.

Imperatives, we may suppose, Normally perform their stabilizing function not by producing beliefs, but by producing motivating states such as desires. If the above is correct, our cognitive system would simply be resistant to producing what we would otherwise expect the stabilizing outcome of a sentence like (4) to be: a disposition to infer from the desire Normally conveyed by the antecedent to the belief Normally conveyed by the consequent. This would explain why (4) is infelicitous.

More generally, it is plausible that in evaluating the truth of an assertion, we typically rely in part on the same inferential operations on the belief that it Normally produces in us (or, more precisely, the state that constitutes *entertaining* this belief). But if a command Normally produces not a belief, but a motivating state, and the cognitive system is insensitive to the intentional relations of the descriptive content of motivating states, that could explain why we do not find commands truth-apt.

This explanation does not preclude that there *could* be attitudes that resembled motivating states in some respects—their directive content would not be generic like that of a beliefs, but directly concerned with action—yet whose descriptive content were inferentially engaged. If there were, we might expect that our conventional linguistic means of expressing these attitudes would also permit us to express our reasoning *about* them. In other words, we might expect them to have declarative syntactic form, so that they could be felicitously embedded in conditionals and so that truth and falsehood could be attributed of them. In the next chapter, I will defend a theory of certain normative judgments that identifies them with motivational attitudes akin to desires. As we know, the sentences conventionally used to express normative attitudes do have declarative form.

3.5. Summary and Conclusion

I have evaluated Artiga's argument and concluded that—with a few reservations—it seems sound: Millikanian teleosemantics does entail the universal hybridity thesis. I have also tried to show that this conclusion should not be cause for worry. There are solutions to the problems it gives rise to.

My defense of Artiga was in part motivated by a specific agenda. I want to use his conclusion in my analysis of normative thought and language. In particular, I want to use the idea that *directive* representations have *descriptive* content in order to defend a teleosemantically informed version of the non-cognitivist idea that normative judgments are directive attitudes. It is to this project that we now turn.

4. Descriptive Content and Normative Truth

One of the strongest points in favor of non-cognitivism about normative talk and thought is the fact that many normative judgments seem very much like motivational, practical, or action-guiding-states, and the statements that express them very much like directives.

A statement like "you ought to get a haircut," for instance, strikes us as having roughly the illocutionary force of "get a haircut!" A normative judgment in the first person—my judgment about what *I* ought to do—seems akin to a directive attitude, one that is directly involved in motivating me to make certain decisions. Even the corresponding judgment in the third person, my judgment about what some other person *A* ought to do, seems to imply something about my motivations, i.e., about what I *want* him to do, what I would be moved to attempt to make him do.[72]

But there is also good reason to be skeptical to this idea. In many respects, normative judgments behave like regular descriptive beliefs. In Terence Cuneo's words, already quoted above, "they are classificatory, truth-evaluable, apt candidates for knowledge, and apt for inference" (Cuneo 2018). Their aptness for inference, in particular, is manifest in the fact that the sentences that conventionally express them have the same syntactic and embedding properties as regular descriptive sentences (but see Franzén 2018, chap. II). For instance, they can figure in the antecedents of conditionals (cf. Geach 1960, 1965; Searle 1962).

So there are some intuitively compelling reasons to think that normative judgments are motivational attitudes rather than cognitive ones, but also strong pressure against this idea. But perhaps we don't have to choose! In the last chapter, I argued that teleosemantics entails the *Universal Hybridity Thesis*, a result that, if correct, would entail that even intuitively motivational attitudes have descriptive content. Perhaps, then, we could hope to explain some of the "cognitivist" features of normative judgment in terms of those

[72] Granted, as soon as we move away from statements and judgments on this simple form, things look less clear. Consider past-tense judgments. What is the motivational import of my judgment that Napoleon ought not to have invaded Russia? Napoleon is dead, and so is his Russian campaign. What is the motivational import of my judgment that Playstation 4 is the best gaming console, or that capital punishment is wrong? However, even statements and judgments like these can be imagined to possess motivational import. It's just that their exact motivational function is more difficult to articulate than for the simple cases mentioned earlier in the paragraph (cf. n. 92, p. 129).

descriptive contents? This chapter is an extended attempt to evaluate the prospects for this simple idea.[73] It will turn out to be not-so-simple after all, and to make it work, we will have to further rethink the relationship between descriptive content and truth, a project I began in section 3.4.

To evaluate this idea we need, first of all, some notion of what the descriptive content of normative judgments might be. This is what I will attempt to provide in section 4.1. I begin by trying to determine what descriptive contents we should attribute to cognitively simpler motivational states like desires. I then go on to look at the descriptive contents of more deliberative descriptive attitudes like intentions, before considering the class of ought-judgments conventionally expressed by sentences of the form "*A* ought to φ," like "Axel ought to get a haircut." My arguments in this part of the chapter are intended to illustrate broad principles rather than giving definitive accounts. The ideas in the rest of the chapter will survive significant revisions in details.

In section 4.2 I observe an interesting consequence of my account: that tokens of the same ought-judgment in the heads of different subjects can have different descriptive contents, and that the sentences used to express them in general have no descriptive content at all, hence do not qualify as representations. For the latter reason, I call the resulting view "discursive non-descriptivism." I go on to explain in what sense two token attitudes can count as "the same" even though they have different descriptive content, and show that the relevant sameness-relation overlaps to a considerable extent with the traditionally important relation *sameness of propositional content*.

The rest of the chapter is dedicated to discussing consequences of discursive non-descriptivism. In sections 4.3 and 4.4 I present an account of how attributions of truth, correctness, and speaker competence are applicable to statements that lack determinate truth-conditions. I suggest that the truth-predicate and related terms sometimes have the function to convey, not that

[73] Various versions of the idea that normative or moral judgments are hybrid attitudes, often called "besires," already exist in the literature. The term "besire" originates in (Altham 1987, 284–85), who broaches the idea only to reject it. For further critical discussion, see (Smith 1995, 116–20).

The besire-idea bears a close affinity to the family of meta-ethical views known collectively as "hybrid expressivism" (see e.g. Barker 2000; Copp 2001; Ridge 2006; Boisvert 2008). Mark Schroeder (2009) has written an extensive review and criticism of hybrid expressivism in its various forms, though as far as I can tell, none of the positions discussed by Schroeder corresponds to the one I will develop here. I don't think the latter view can properly be thought of as a form of expressivism at all, at least not if expressivism is understood as a type of view that tries to understand the semantics and pragmatics of its target vocabulary in terms of the type of attitudes they are used to *express* (where "expression" can be understood, roughly, as the relation between a speech-act and the state the speaker has to be in for it to be permissible for her to make it (cf. Schroeder 2010, 28–31)). Following Millikan, I instead try to understand the semantics and pragmatics of ought-statements in terms of their functions, i.e., in terms of the attitudes they have as their stabilizing function to *produce* in the listener. To what extent this constitutes a genuinely alternative approach will, regrettably, have to be determined by future research.

the statement *itself* is descriptively correct, but that the attitude Normally *produced* by that statement in a given subject's head is descriptively correct. I further illustrate this account by comparing it with John MacFarlane's theory of assessment-sensitive truth in (MacFarlane 2014).

In section 4.5 I discuss why ought-statements often appeal to common interests. Section 4.6 is devoted to a brief sketch of a program for the compositional semantics of normative sentences compatible with the view defended. Section 4.7, finally, discusses some indeterminacy issues that will have arisen in section 4.1.

In this chapter, things start to get somewhat speculative, and many of the views defended are tentative or function as placeholders for more complete accounts. I hope the main ideas defended are sufficiently interesting to justify some promissory notes of this kind.

4.1. Descriptive Content of Directive Attitudes

In this section, I will look at the descriptive content of directive attitudes and try to figure out what the descriptive content could be of statements and judgments of the form:

> *A* ought to φ

(where *A* is either a name referring to an individual, or a personal pronoun like "I" or "you"). I will begin from the assumption that such statements convey, and such judgments are, directive attitudes, whose function is to bring it about that *A* φs, and which Normally perform this function by motivating the subject of the judgment to make *A* φ (or, in the case where the subject herself is *A*, by motivating the subject to φ).

The proposal I will end up defending is the following: a subject *S*'s judgment that *A* ought to φ has the descriptive content given by

> (GENERAL NORMATIVE) *A*'S φ-ing has the highest expected benefit for *S* out of the alternatives available to *A*.

Before I go on to show how I reach this conclusion and what it entails, let me briefly dwell upon what kind of conclusion it is. The reader might worry that a content specification of this kind falls afoul of Moore's open question argument (Moore 1903, para. 13). Isn't it possible to accept that *A*'s φ-ing has the highest expected benefit for *S* out of the alternatives available to *A*, yet wonder whether *A* ought to φ, and be rational in thus wondering?

Here, we must recall that sameness of descriptive content doesn't entail sameness of logical or inferential role (p. 59). Hence, sameness of descriptive content is not something that we should expect to be recognizable *a*

priori. Sameness of descriptive content is like sameness of reference in not being epistemically transparent to a rational, linguistically competent subject.

But this admission raises questions about what kind of explanatory role descriptive contents *do* have, and how hypotheses about them can be tested. On p. 95 above I suggested that the descriptive content of desires may be inferentially inert, meaning that it is not part of the Normal operation of the cognitive system to test the descriptive consistency of the desire-set. Whether this is true, it seems clear that ought-judgments *are* inferentially engaged. This is one of the cognitivist features of normative thought I hope to explain by assigning them descriptive content. Normally, when we draw valid inferences, it is presumably because we *recognize* that the contents of the attitudes involved in the inference bear the appropriate modal relation. But in keeping with Millikan's dictum that "intentionality and rationality are *not* two sides of the same coin" (LTOBC, 140; cf. p. 60), these modal relations are not necessarily epistemically transparent to the subject. Psychological assumptions—about how and to what extent the cognitive system is Normally wired up to *reflect* (p. 31) the modal relations between the descriptive contents of these attitudes—must be adduced to explain the inferential engagement of our normative judgments.

In what follows, I make no serious attempt to spell out and defend a psychological theory of normative inference of this kind. Instead, I will rely in a somewhat unprincipled manner on the assumption that people are, to some extent and Normally, but far from infallibly, sensitive to the modal relations between the descriptive contents of their normative attitudes as well as between those and the descriptive contents of other attitudes like beliefs. I assume that they will tend to avoid tokening attitudes whose descriptive contents are inconsistent, tend to token an attitude whose content is entailed by one they already harbor, and so on; but that various factors can intervene to make these tendencies less than perfectly reliable. "To some extent and Normally" is a theoretical promissory note that must be cashed in by future refinements of the current proposal.

Though this lacks in concreteness, I think it is sufficient to allow our intuitions about the inferential relations among attitudes to serve as rough guides to evaluating proposals about their contents. In particular, if we take two attitudes to be inconsistent, we should assign descriptive contents to them that are inconsistent, and if we take one be inferable from the other, then we should assign content to one that is entailed by the content of the other. Otherwise, it becomes difficulty to explain why our cognitive system should be equipped with the inferential dispositions that underlie these intuitions.[74]

[74] Note that Moore's open question argument trades, not on the recognition of an inconsistency or entailment, but on the *absence* of any such recognition. It seems plausible to me that "active" inferential dispositions, such as a disposition to treat two attitudes as inconsistent

A second heads-up about the discussion to follow: if the descriptive content of normative judgments is to be able to account for their cognitivist features, then, one may think, that content has to be *determinate*. I suggested on p. 76 that although content indeterminacy may be acceptable for primitive, non-conceptual representations, it is not so for beliefs and other truth-apt representations. This, I suggested, is because the descriptive contents of those representations are supposed to coincide with their truth-conditions. It will therefore be disheartening to learn that the analyses to follow reveal a number of potential sources of indeterminacy for directive attitudes, including normative judgments. But this chapter will also further call into question the straightforward relationship between descriptive content and truth that I began to tease apart in section 3.4. For this reason, I will just flag those problems here, and return to discuss their implications in section 4.7 after we have the rest of the account in hand.

With these preparatory remarks out of the way, let us get to the topic at hand. What is the descriptive content of a directive attitude? There are several types of directive attitudes: desires, intentions, longings, brute urges, etc. In the philosophical literature, "desire" is often used as a generic term to cover them all, but as I hope to illustrate below, it is unlikely that a generic characterization can be given of the descriptive content of a directive attitude in general. To give us somewhere to start, and as a way to introduce the principles and assumptions I will be relying on, I begin by discussing "desires" in a sense closer to the colloquial one. I mean something like wants or goals, something that informs one's decision-making while bearing a less direct relationship to action than, say, intentions (to be discussed further down).

As I suggested on p. 87, it is plausible that desires are produced by mechanisms that are Normally sensitive to whether the behaviors they motivate will yield beneficial outcomes for the organism. Following this suggestion, let us consider, as a first very rough approximation of the descriptive content of the desire to φ:

(DESIRE) φ-ing will be beneficial.

"Beneficial" should ultimately be cashed out in terms of fitness. The desire-forming mechanisms must, after all, have been selected for motivating the organism to pursue historically fitness-enhancing outcomes. But the organism has no means of tracking all possible means of enhancing its fitness. It has to rely on certain proximal signs. Food is a sign of fitness, and so is sex,

---

or to infer one from the other, are stronger predictors of content relations than "passive" inferential dispositions, such as a failure to treat two attitudes as inconsistent or infer one from another. The former, it seems, requires an active engagement by the cognitive system that stands in need of positive explanation, whereas the latter can be chalked down to a mere failure to do something, a product of inertia.

and so, in the human evolutionary past, were loftier things such as the esteem of others and the satisfaction of curiosity. These all count as kinds of benefit in the relevant sense.

It is unlikely, however, that Normal desire-formation indiscriminately tracks aggregated benefit. More plausibly, different kinds of desires are formed under the influence of expectations of different kinds of benefit. The details must be supplied by empirical psychology, but we can make some rough educated guesses.[75] Through experience, the subject comes to associate certain outcomes with more fundamental goods, and through explicit means-ends reasoning she draws conclusions about the most efficient means to attain these outcomes. I may desire to go to a good school because experience suggests that this is a good way of getting a good education, and getting a good education is a good way to impress people, i.e. to satisfy the more basic need for social esteem.

According to behavioral psychology, behavior is built on the basis of an inventory of "primary" or "unconditioned reinforcers," including all the usual suspects like food, sex, and pain, but presumably, at least in humans, also things like social esteem, intimacy, accumulation of new information, and other outcomes of a more socially or intellectually valuable sort. These act like proximal signs of fitness for the organism. Outcomes that are not in themselves primary reinforcers can come to be associated with primary reinforcers, for instance through classical conditioning, and these "secondary reinforcers" will then also function as psychological rewards or punishment, driving learning and decision-making. Money is the standard example of a secondary reinforcer. Plausibly, the function of the learning mechanisms whereby humans are equipped with secondary reinforcers is to make us pursue outcomes that are actually good predictors of primary reinforcers. The secondary reinforcers Normally act like natural signs of primary reinforcers, which in turn are Normally natural signs of fitness.

Plausibly, the desire-forming machinery is designed to motivate pursuit of outcomes that are either themselves primary reinforcers or have acquired the status of secondary reinforcers because they are natural signs of primary ones. Different desires can have been designed by processes that Normally track different kinds of primary and secondary reinforcers. For instance, I can harbor a desire to eat a large greasy hamburger and also a desire to lose weight. The first might have been formed by a process that tracks outcomes that yield calories, the other by a process that tracks outcomes that yield social esteem and sex. Perhaps, then, rather than assigning descriptive content to desires in terms of undifferentiated benefit, we should assign to each desire a descriptive content in terms of the specific kind of benefit—the specific primary reinforcer(s)—Normally tracked by the process that have

[75] Nothing argumentatively rides on the details of these guesses. They are to illustrate the general approach.

formed it. So to my desire to go to a good school we could assign a descriptive content along the lines of *Going to a good school will get me social esteem*.

Problems of a Neandrian kind start to emerge at this point (cf. section 2.5). Suppose going to a good school really is a good way of getting a good education, but it is *not* a good way to get me social esteem (those around me resent the perceived self-importance and pedantry of educated people). In this case, my desire-forming machinery has successfully identified what my motivational system treats as a proximal sign (getting a good education) of a more basic good (social esteem), but my getting a good education is in fact *not* a sign of these more basic goods. On one level the desire-forming machinery has been successful. On another, more basic level it has failed. But the failure doesn't inhere in *this* particular employment of the machinery. It successfully identified a means to get a good education. The failure was embedded in the machinery already before the deployment, the result of abNormal functioning in earlier learning. Should we say, then, that this desire of mine is "correct" or "incorrect," that it does or doesn't correctly describe the world? This is the first of the indeterminacy problems that I warned about above. Let us keep it in mind for later.

Reflecting on the Normal process of desire-formation gives us some analytical tools for tackling normative judgments. But desires are, in many respects, bad models for normative judgments. If normative judgments are directive attitudes, they are probably less like desires and more like *intentions* in that they are *deliberative* states formed by explicit means-ends reasoning (or at least sensitive to such reasoning), taking into account the relative desirability of various outcomes and the likelihood that these outcomes will occur. There are differences as well, that I will discuss below. But the similarities are compelling enough to merit a discussion of intentions.

Intentions differ from desires in several respects. I can have multiple incompatible desires, some of which can even be impossible to satisfy, and presumably, there need be nothing abNormal about this: there is a point to keeping desirable outcomes present before the mind, even if circumstances prevent me from attaining them, in case circumstances should suddenly change. Though I might not currently be able to afford a sports car, I could win the lottery tomorrow.[76] Intentions, however, are recipes for action, either imminent action or action under specific contingencies. To function Normally, they must presumably be sensitive to my actual scope for action. If I am fiscally constrained, I can satisfy my desire to have a sports car *or* my desire to have a savings account, but not both. If I nevertheless intend to satisfy

[76] More than one ancient school of wisdom teaches that the key to happiness is to get rid of vain desires, and this may indeed constitute wisdom for some people, but seeing how hard it is to actually accomplish, it's a good guess that it does *not* constitute evolutionary wisdom.

both, there is presumably something wrong with my intention-forming machinery.[77]

As a basis for discussing likely candidates for the descriptive content of intentions and of normative attitudes, let us assume that the following, rather conventional model of Normal intention formation is true. When forming an intention, one compares a number of different available action alternatives and estimates their likely consequences. One weighs these consequences against each other by some metric, and forms the intention to pursue that course of action which yields the highest estimated benefit. On this model the intention-forming machinery Normally tracks action alternatives with the *highest expected* benefit out of the range of alternatives. On the basis of this model, I propose the following descriptive content for the intention to φ:

> (INTENTION) My φ-ing will produce the highest expected benefit for me out of the available alternatives.

Now let me list three potential problems for this proposal. *First*, what constitutes "most benefit" is subject to the same indeterminacy worries as the proposal for desires, DESIRE, discussed above. This is aggravated by analogous worries pertaining to the hypothesized weighing mechanism: to the extent that this weighing is a product of learning, there can be malfunctions in earlier learning episodes that yield maladaptive intention-formation later in life, and the question returns whether we should attribute these mistakes to the intention-formation episode itself.

Second, there is the issue of what constitutes the "available alternatives." Nobody, I assume, forms her intentions by considering *all* alternatives that are, in a physical sense, available to her. Most options are never considered at all. Testimony from therapeutic practice, conveyed to me in personal conversation, suggests that it is sometimes sufficient to make a patient aware of the alternatives available to her in order to completely change her life: she had persisted in dysfunctional behavior due only to a lack of imagination. Is such a patient's intention-forming machinery operating abNormally? How much consideration of alternatives should we demand from a person before we attribute Normally produced intentions to her? Is it sufficient that she considers the alternatives that are starkly before her conscious mind?

Third, my proposal entails that intentions descriptively represent the estimated *likelihood* of different outcomes given different courses of action. Since intentions are not traditionally assumed to have descriptive content, there is no direct intuitive support for this feature of the proposal. But we can support it indirectly if we assume, as I am about to do, that intentions have descriptive content similar to that of first-person ought-judgments. Many

[77] Unless, of course, I intend to steal the car. I hope the principle I wish to illustrate is clear.

people do have strong intuitions that the truth of a first-person ought-judgment depends, in most cases, not on what will actually happen if we act according to the judgment but on what can be expected to happen. Whole theories on the semantics of "ought" have been built on the assumption that its truth-conditions are sensitive to the probability estimates available to the subject (e.g. Finlay 2014). If I judge that I ought to buy a sports car, it seems unduly harsh to accuse me of a *mistake* in normative judgment just because the car gets totaled in a freak accident the same day it rolls out from the dealership.

If this is correct, ought-judgments are akin to epistemic modal judgments (like the judgment that there might be rain tomorrow) in that the correctness or truth of the latter is also, intuitively, contingent on the evidence available to the subject. The question, however, is whether "estimated likelihood" or "evidence available to the subject" can actually serve to characterize a Normal success condition for an attitude. There is reason to suspect that they cannot. Consider the following example: I smell a weird odor, and, suspecting that a poisonous gas may have infested the university building, I conclude that I ought to go outside. I then act accordingly. In this case, it seems as though the *success* of my ought-judgment and the action it prompts is contingent only upon whether there is *actually* a poisonous gas in the building, not on whether my evidence supported this surmise. If there was, I have just avoided a nasty death, but if there wasn't, I have had to interrupt my work and go outside for no reason.

The "ought" that is made true or false not by the expected benefit of our actions but by the outcomes those actions actually produce is what philosophers often call the objective "ought" (see e.g. Broome 2013, chap. 3). Perhaps, then, the teleosemanticist can say that there is no "ought" save for the objective "ought." My freak accident means that I actually ought not to have bought a sports car, and my judgment to the contrary was false. I might have been faultless in making my judgment, because the freak accident was not something I could have predicted, but it was false nonetheless.

The problem with this strategy is that there are possible circumstances where we are in a position to determine that an ought-judgment, if understood to be an objective ought-judgment, would likely be false, and yet this ought-judgment seems intuitively like the correct one to make. In such cases we can't appeal to the agent's lack of information to account for the mismatch between truth and our intuitions about correct judgment. John MacFarlane illustrates this point with the following example:

> Suppose you buy three rubber duckies for your child, and later learn that one out of every hundred rubber duckies from this manufacturer leaches out toxic chemicals. What should you do? […] [W]e know that it is highly probable that all three of the duckies are safe, and hence highly probable that we ought, objectively, to keep them all. Despite that, we decide to throw them out, and rationally so—we

would not risk a child's life for the price of three rubber duckies. (MacFarlane 2014, 282)

If the ought-judgments that inform our decision-making were objective ought-judgments, we would tend to make decisions that seem, in fact, to be unreasonable. A cognitive system working on this principle would, on average, produce less aggregated benefit over time than a system employing subjective ought-judgments.[78] The former kind of system would be sensitive only to which action-alternative is likely to produce a better outcome, not to the *relative amount* of benefit produced by different action-alternatives given different possible circumstances (cf. Gibbard 2005, 345).

An analogous problem for teleosemantics has been observed in the case of epistemic modal judgments by Gunnar Björnsson (2018), and perhaps we can take guidance from his proposed solution to that problem in our present predicament. Epistemic modal judgments—such as the judgment that *p* is likely, that it might be that *p*, that it must be that *p*, and so on—would seem to represent the subject's epistemic circumstances if anything does. But as Björnsson observes, it seems that in this case, too, the success-conditions for actions guided by epistemic modal judgments are concrete facts about the world, not facts about the subject's evidence. In order to account for the intuitive truth-conditions of such judgments, Björnsson proposes that epistemic judgments have as their function not only to guide concrete behavior, but to help implement behavioral *strategies* "such as that of performing the action alternative that is ranked best when possible risks and benefits of each alternative are given weight in proportion to how strongly the [subject's] evidence suggest that they would accompany that alternative" (Björnsson 2018, 268–69). Presumably, following this strategy is Normally successful only if the subject's epistemic judgment contributes to ranking action alternatives according to the evidence the subject in fact has. Hence, we can conclude that the epistemic judgment descriptively represents facts about the subject's evidence.

What would the analogous solution in the case of ought-judgments be? Can we say that the function of an ought-judgment is, not to produce concrete action but to implement a strategy of acting on those action-alternatives that yield the highest expected benefit? This doesn't seem altogether implausible—but at the same time, it would be strange if an ought-judgment (or an intention) did not *also* have the function of producing a concrete action.

A feature of the view that Björnsson defends is that epistemic modal judgments have not one, but *two* descriptive contents: one *strategic*, deriving from the Normal success-conditions of the epistemic strategy they help im-

[78] This, at least, is the case on the assumption that these system, as it were, operates on a policy of trying to make those ought-judgments most likely to be true—which seems like a fair assumption to make.

plement, and one *concrete*, deriving from the concrete actions those judgments guide the subject to perform (Björnsson 2018, 270–71).

> On the one hand, we have good reproductive explanations pointing to concrete effects (predator avoidance, rehydration) of actions guided by various utterances, beliefs and judgements, and corresponding representational contents. Call these contents "concrete". On the other, we have good reproductive explanations pointing to the strategy implementation function of these same actions, and corresponding contents. Call these contents "strategic." (Björnsson 2018, 271)

Björnsson spends the rest of the paper defending the feasibility of the idea of dual contents. If the same idea can be applied to ought-judgments as well, we would have a picture according to which each of these judgments are, in a sense, *both* a subjective *and* an objective ought-judgment. I cannot evaluate Björnsson's defense of the idea of dual contents here. For our purposes, the dual content problem can very well be thought of as another form of content indeterminacy (or, conversely, indeterminacy problems might be better thought of as symptomatic of features of representations that could be more fruitfully approached in terms of multiple contents; cf. p. 75). As far as I can tell, the approach to the indeterminacy issues that I offer in section 4.7 is different from Björnsson's dual contents approach, though there may be deeper affinities (cf. n. 61, p. 77).

With these rather serious reservations in mind, let us consider the prospects for generalizing something like INTENTION to the case of normative judgments. We will begin with first-person normative judgments, i.e. those conventionally expressed by statements of the form

> I ought to φ

As already mentioned, this type of judgment seems to have much in common with intentions. They are, if not products of deliberative practical reasoning, then at least sensitive to such reasoning. We could therefore try to simply assign them the *same* descriptive content as the corresponding intentions, so that my judgment that I ought to φ gets assigned the descriptive content:

> (FIRST-PERSON NORMATIVE) My φ-ing has the highest expected benefit for me out of the available alternatives.

There are some problems with this proposal, however. To begin with, it seems conspicuously self-centered. We might have thought that judgments about what I *ought* to do were often motivated not by considerations of what would benefit *me*, but rather of what would benefit people in general or perhaps what obligations I have. On the other hand, if we are serious about our naturalism, we must acknowledge that the capacity for making ought-

judgments must have paid its evolutionary dues *somehow*, and this means that the explanation of their past persistence must appeal to their ability to motivate actions that are *somehow* beneficial for the agent, in the sense of "beneficial" we have been employing hitherto: either historically fitness-enhancing, or something the cognitive system takes to be a proximal sign of a historically fitness-enhancing outcome.

There need be no conflict between "beneficial for me" and "beneficial for others," or between "beneficial for me" and "my duty." A social creature's welfare is often tied up with the welfare of others and with the opinions others have of her, which can in turn depend on whether she reliably discharges her duties. But FIRST-PERSON NORMATIVE suggests that the judgment that I ought to φ can be descriptively correct even when the benefit I accrue from φ-ing is eminently anti-social.

Another issue with FIRST-PERSON NORMATIVE, which serves to nuance the above dialectic somewhat, is the fact that normative judgments—in contrast to intentions—are usually taken to come in different *kinds* or *flavors*: prudential normativity (what one ought to do in the light of prudential considerations), moral normativity (what one ought to do in the light of morality), epistemic normativity (what one ought to do in the light of epistemic considerations) and so on. An intention, on the other hand, is typically formed in the light of all considerations—or at least all considerations the subject is aware of at the time.[79]

This also helps to explain another intuitive difference between intentions and first-person ought-judgments: the former are more directly tied to action. Even if ought-judgments are motivational states, the failure of rationality that consists in not acting on one's normative judgments is intuitively less severe than failure to act on one's intentions. This makes sense if normative judgments Normally track only certain categories of benefit. It would then be rational and consistent to simultaneously harbor ought-judgments in favor of different courses of action, as long as those judgments have different flavors, and the cognitive system would need to be equipped with a mechanism for weighing these against each other and choosing between them before issuing in action. The intention-forming machinery seems like a good candidate for playing that role.

Philosophers, it is true, also sometimes talk about "all things considered" normativity: what one ought to do in the light of all considerations. Perhaps

[79] As a sort of extreme version of the idea of different normative flavors, Stephen Finlay (2014) has proposed an *end-relational* semantics for "ought" (as well as for "good" and "reason"). On Finlay's view, every use of the deontic "ought" indexes a specific goal determined by the context, and the truth-conditions of "ought(*p*)" are, roughly, that *p* is that possibility out of a range of alternatives (also contextually determined) that makes attainment of the goal most likely (Finlay 2014, 73). If Finlay is right, then instead of a limited range of different normative flavors, we have an "ought" that can be relativized to any conceivable goal—including highly anti-social ones.

an all-things-considered ought-judgment, at least, shares the descriptive content of an intention. Indeed, maybe a (first-person) all-things-considered ought-judgment simply *is* an intention. If not, there are some grounds for questioning their psychological reality. For suppose all-things-considered ought-judgments stood at one remove from action the way other ought-judgments do. Then I, having made such a judgment, would still have to form an intention to act on it. And either my intention-forming machinery could Normally form an intention to act *contrary to* my all-things-considered ought-judgment (perhaps favoring my judgment about what I ought *morally* to do instead), or it could not. In the former case, it is unclear in what sense the all-things-considered ought-judgment *is* a judgment about what I ought to do in the light of all considerations, as opposed to just another flavor of ought-judgment standing on the same level as the others. And in the latter case, it is unclear why the cognitive system would need an additional discrete attitude to mediate between particular ought-judgments and intention-formation.

It seems likely, then, that first-person ought-judgments, rather than tracking indiscriminate aggregate benefit, Normally track particular *kinds* of benefit—at least insofar as they are distinct from intentions. In the discussion to follow, I will nevertheless rely on the formulation of FIRST-PERSON NORMATIVE to simplify the discussion. As already stated, general evolutionary considerations together with the assumption that ought-judgments are directive attitudes would seem to require that these judgments Normally correlate with the potential for *some* kind of benefit for the subject. The reader should keep in mind, however, that ought-judgments are unlikely to track benefit *indiscriminately*. I will return to these issues in section 4.5.

Let us now see if we can generalize these suggestions to the third-person case. Let us look, in other words, at judgments conventionally expressed by statements of the form

A ought to φ

where "*A*" (for "agent") is the name of a person. If the function of my first-person ought-judgment is to make me φ, then it seems like the function of my third-person ought-judgment should be to motivate me to make *A* φ. But there is clearly a difference between my judgment that *A* ought to φ and my judgment that *I ought* to make *A* φ. The former is not reducible to the latter.

Intuitively, when we are considering what *we* ought to do, we are comparing *our* alternatives. The judgment that I ought to make *A* φ favors this course of action—making *A* φ—over any alternative courses of action available to me. When we are considering what *A* ought to do, by contrast, we are comparing *A*'s alternatives. The judgment that *A* ought to φ favors *A*'s φ-ing over any alternative actions available to *A*. In light of this, we might try to assign the following descriptive content to a subject *S*'s judgment that *A*

ought to φ (In the case of first-person normative judgment, $A = S$, and GENERAL NORMATIVE reduces to FIRST-PERSON NORMATIVE):

> (GENERAL NORMATIVE) *A*'S φ-ing has the highest expected benefit for *S* out of the alternatives available to *A*.[80]

Here, however, the problem of self-centeredness returns with a vengeance. Shouldn't a judgment about what *A* ought to do, if it concerns anybody's benefit, concern either *A*'s benefit or else perhaps the benefit of people in general? GENERAL NORMATIVE, however, implies that such a judgment is concerned only with the judger's benefit.

Add this problem to the problems of indeterminacy already generously discussed, and we may be tempted to abandon the whole attempt to understand normative judgments as motivational attitudes along the lines laid out here. But we shall soldier on. I will assume that GENERAL NORMATIVE is the correct—or close to the correct—account of the descriptive content of "*S* ought to φ"-judgments. In the next three sections I will be discussing a consequence of this assumption: that different tokens of the same ought-judgment in the minds of different judges have *different* descriptive contents. This consequence may itself seem sufficient to doom the project of accounting for the "cognitivist" features of normative discourse in terms of the descriptive content of directive attitudes, but I will argue that it does not, and understanding why not will also allow us to understand features of normative *discourse* which will help us allay the self-centeredness concern (and perhaps the indeterminacy concern as well).

4.2. Discursive Non-Descriptivism

In this section, I will discuss the consequences of GENERAL NORMATIVE for ought-*statements*. A statement of the form "*A* ought to φ" will, on any plausible teleosemantic theory of such statements, have as its stabilizing function to produce the *judgment* that A ought to φ in the head(s) of the addressee(s). Need more be said?

[80] If GENERAL NORMATIVE gives the descriptive content of third-person ought-judgments, they would seem to be less like intentions and more like mere desires. Since they don't Normally track the relative expected benefits of the action-alternatives available to the judgment's subject, but only those available to its object, they are ill-suited to serve as directly action-guiding states. True, judgments that Normally track relative expected benefit for another person might be the product of what we may think of as "vicarious" practical deliberation, putting ourselves in the shoes of the other and considering the various action-alternatives available to her. But since GENERAL NORMATIVE predicts that ought-judgments only represent benefit for the judger, and not for the judgee, such a judgment, it seems, could not function as more than an idle wish unless the judger had some direct way of influencing the judgee's behavior.

Let us note a curious consequence of GENERAL NORMATIVE. If it is correct, then *my* judgment that *A* ought to φ and *your* judgment that *A* ought to φ, although intuitively "the same judgment," in fact have *different* descriptive contents. This is because, according to GENERAL NORMATIVE, the descriptive content of an ought-judgment is relativized to the judgment's subject on two counts: it concerns the *subject's* benefit and the *subject's* expectations. And note that this remains true no matter how we resolve the various indeterminacy issues discussed above. If ought-judgments are objective ought-judgments, their descriptive content will still be relativized to the subject's *benefit*. And no matter how we interpret benefit, it will still be the *subject's* benefit that is in question.

At first look, this may not seem all that curious, although worrying for GENERAL NORMATIVE. It may seem, simply, as an expression of the metaethical position we can call *indexical contextualism*, or just "indexicalism." Indexical contextualism about some normative term is the view that its reference depends on the context in which they're uttered, much like common indexicals such as "I" and "here." If some version of indexicalism is true, we would expect that when I say "A ought to φ" my statement has different truth-conditions than when you say "A ought to φ." Though using the same sentence to express our respective attitudes, these sentences—and the attitudes they express—in fact have different descriptive contents. If we assume that "ought" indexes the property of being (expectedly) beneficial for the speaker, we may seem to get the semantics predicted by GENERAL NORMATIVE, a view akin to what is often called "speaker subjectivism."

On closer scrutiny, however, this is not in fact what is going on. To see why, we have to talk briefly about attitude attributions. Generally, indexical expressions, when embedded in attitude attributions, get their referents in the context of utterance. When I say "Axel believes that I'm here," the belief I attribute to Axel has the descriptive content that *I'm* here, not that *Axel's* here; and that I'm *here*, where I actually am, not that I'm wherever Axel is.

We can expect that if "ought" were indexical, it would behave in attitude attributions the same way as other indexicals. But then, if "ought" indexes some property pertaining to the *speaker's* expected benefit, then when I talk about *S*'s judgment that *A* ought to φ, the attitude I attribute to *S* should be the judgment that, roughly, *A*'s φ-ing would benefit *me*, not that her φ-ing would benefit *S*, contrary to the account I have offered.[81]

Millikan's own account of attitude attributions vindicates this point. She argues that in such constructions, roughly, the embedded sentence indexes the attitude that the sentence, uttered unembedded *in the same context as the*

[81] This argument specifically shows the incompatibility of indexical relativism with GENERAL NORMATIVE, but arguments appealing to problems with attitude attributions (and attributions of indirect speech) can be employed as a general strategy against indexicalist views. See e.g. (Schroeder 2009, 285–86).

attribution, would have as its stabilizing function to *produce* in *S* (LTOBC, 211–13). But as we saw for assertions on p. 62, the descriptive content of a speech-act depends not on the attitude it is used to *express*, but on the attitude that it has, as its stabilizing function, to *produce*. That is just an application to the case of discourse of the general principle that the descriptive content of a representation depends on the Normal conditions for the consumer response. So if indexicalism were true, "*A* ought to φ" as uttered by me would have as its stabilizing function to produce in you an attitude with a descriptive content roughly like *A's φ-ing would benefit Karl Bergman.* And that would also be the judgment I attribute to you with the phrase "your judgment that *A* ought to φ," *contra* GENERAL NORMATIVE.

So if indexical contextualism is true, and given this general principle about how indexicals behave in attitude attributions, GENERAL NORMATIVE is false. By contraposition, if GENERAL NORMATIVE is true, indexicalism isn't. So what is? Something slightly more curious. The hypothesis, recall, is that when I utter "*S* ought to φ," this statement has the stabilizing function to produce in my hearer the judgment that *S* ought to φ, which according to GENERAL NORMATIVE has a descriptive content pertaining to what benefits the hearer. So if my statement's descriptive content depends on the attitude it produces in the hearer, it too should be about the hearer's benefit. When I'm telling you that *A* ought to φ, I'm not talking about what would benefit me, but about what would benefit you.

This also looks like a form of indexicalism, although one according to which "ought" behaves more like "you" than like "I," indexing the addressee rather than the speaker. But again, this isn't really what's going on, and again, some observations about attitude attributions help bring the point out. If I tell you that Axel believes that you're a fool, the attitude I'm attributing to Axel has the descriptive content that *you're* a fool. So if "ought" behaved like "you," and I told you that Axel thinks that *A* ought to φ, we should expect the attitude I'm attributing to Axel to be one whose descriptive content pertains to the benefit *you* can draw from *A*'s φ-ing. But again, this is not what GENERAL NORMATIVE says. GENERAL NORMATIVE says that the descriptive content of Axel's ought-judgments pertain to *Axel's* benefit, *regardless* of the context in which they are attributed.

There are other disanalogies between how "you" and "ought" behave, if GENERAL NORMATIVE is true. "You," at least the singular "you," indexically picks out a *specific* addressee. If I exclaim "you're a fool!" in front of a group of people, there must still be a particular person I'm addressing myself to in order for my statement to have a determinate content. But there is no similar constraint on "*S* ought to φ." I can address that sentence indiscriminately to a crowd. If the stabilizing function of the sentence is to produce, *in each listener*, the judgment that *S* ought to φ—and this judgment has a dif-

ferent descriptive content for each of them—then it looks like the sentence would also have to have a different descriptive content for each listener.[82]

So GENERAL NORMATIVE does *not*, it seems, allow us to treat "ought" as an indexical. And that might be for the best, because indexicalism has well-known problems dealing with disagreement. If I say

(1) Axel ought to get a haircut

and you say

(2) Axel ought not to get a haircut

We would seem to be disagreeing. But if "ought" picks out different properties in different contexts of utterance (whether in terms of the speaker's or the addressee's benefit) then the property attributed to Axel's cutting his hair by (1) is not the property denied of that same action by (2), so it is hard to explain the appearance of disagreement.[83]

How, then, to account for the semantics of "ought," if not indexically? It starts to look as though, if it is the stabilizing function of "ought" to produce attitudes with the descriptive content given by GENERAL NORMATIVE, "ought"-sentences themselves can't have any descriptive content at all. They are not supposed to adapt the addressee(s) to any *single* state of affairs, even a contextually determined one, but are supposed to adapt *each* addressee to a *different* state of affairs: that so-and-so action on the part of *A* would be beneficial for *that* addressee. But since some of these states of affairs can obtain while others fail to obtain, there is no single set of Normal conditions for the "ought"-statement, hence no single descriptive content. Rather, it will in a sense have different descriptive contents for each addressee—and in another sense, no descriptive content at all.

Both of these interpretations introduce some complications into our picture. The latter implies, at least if "ought"-statements are representations in the first place, that the universal hybridity thesis is false. But perhaps the former interpretation is the more natural one. The way teleosemantics paradigmatically makes use of the producer-representation-interpreter setup in explaining the nature of representation presupposes that there is only one interpreter for each producer, but in the case of discourse that is simply not

[82] There is of course also a plural "you" that can be used to address a crowd. But the singular and the plural you are distinguished by various syntactic markers (I say "you're fools" rather than "you're a fool" when addressing a group), and, in most languages other than Standard English, by distinct words. There are no similar distinctions in the case of "ought."

[83] There are a number of moves available to the contextualist in response to this argument. She can reject the intuitions that disagreement obtains, or she can explain the disagreement in terms of a clash of non-cognitive attitudes or other pragmatic considerations. See (Zeman 2017) for an overview.

true, at least not in the general case. One speech-act can have several different interpreters, and there is no principled reason why Normal conditions for one interpreter response should always equal Normal conditions for another.

There is, however, a third, more conservative possible interpretation that I will make use of for now, namely, that "ought"-sentences are not representations at all. "Ought," though its stabilizing function is to produce *attitudes* that are representations, is itself a "word that doesn't represent" (cf. Millikan 2018). In what follows, I will rely on this assumption in discussing the consequences of GENERAL NORMATIVE. We will need a name for the view, and I will call it "discursive non-descriptivism," since it implies that "ought"-statements, but not ought-judgments, lack descriptive content. Since "discursive non-descriptivism" is a mouthful I'll just write "DND" for short.[84] (In section 4.6 I will sketch a compositional semantics for "ought"-sentences that, I think, shows that DND is not much more than a notational variant of the view that ought-sentences have different descriptive contents for different addressees.)

Can DND account for the pattern of attitude attribution implied by GENERAL NORMATIVE? To answer this question, we need a theory of attitude attributions that accommodates the embedding of nonrepresentational sentences. Luckily, Millikan has a theory of just this kind. She writes:

> What the [embedded sentence] does is to index the type of mental state that it is a focused proper function of the expression type of which it is a token to produce. That is, rather than indexing its own type, the filling indexes a mental-state-type that is so-related to, namely, properly caused by tokens of, its type. (LTOBC, 212)

Can this principle be applied to attributions using "ought"-sentences, to account for GENERAL NORMATIVE? The immediate problem here is that if GENERAL NORMATIVE is right, then at least according to one way of typing attitudes, tokens of "*A* ought to φ" are supposed to produce *different* attitudes in *different* speakers. There is no *one* type of attitude that sentences of this type are supposed to produce. But that, of course, goes for any kind of sentence. Even a straightforwardly descriptive sentence is supposed to produce a different type of attitude in me than in you, namely, a belief-of-mine rather than a belief-of-yours. And at the same time, we can always find a way of typing attitudes that makes it the case that "*A* ought to φ" is supposed to produce *the same* attitude in me as in you. Indeed, there is a quite intuitive sense in which they are supposed to produce the same attitude in us, namely, the judgment that *A* ought to φ.

We can circumvent these difficulties if we read Millikan's principle as saying that the embedded sentence indexes the attitude that it is the function of unembedded tokens thereof to produce *in the subject of the attribution*. So

[84] This has the added benefit of evoking an awesome game.

when I say "Ingemar thinks that *A* ought to φ" I attribute the attitude that tokens of "*A* ought to φ" would have as its function to produce *in Ingemar*.

Now this principle implies a specific way of typing *sentences*. As we saw above, if I tell you that "Ingemar thinks that you're a fool," I'm not attributing to Ingemar the attitude that unembedded tokens of the *orthographic* type "you're a fool" are supposed to produce in Ingemar. These, if anything, are supposed to produce in Ingemar the belief that *he's* a fool. Rather, I'm attributing to Ingemar the belief that unembedded tokens of the same *referentially individuated* type as the embedded occurrence of "you're a fool" are supposed to produce in Ingemar. Two sentence-tokens belong to the same referentially individuated type if, roughly, they have the same structure and the referring terms occupying each given place in that structure has the same (possibly contextually supplied) referents in both sentences. In other words, when I say to Axel, "Ingemar thinks that you're a fool," I am attributing a belief that I could produce by saying, to Ingemar, "Axel's a fool," or that Axel could produce by saying "I'm a fool."

This necessary qualification is already accounted for by Millikan, who adds to her principle that "in intentional contexts *all* expressions are typed by reference […] only" (LTOBC, 213). If DND is correct, however, "ought" is not a referring expression, and it must thus be typed some other way. Stabilizing function lies close at hand. Millikan suggests that "surely, we can assume that the stabilizing function […] of the most basic syntactic form of the whole sentence in intentional contexts is considered for purposes of typing it" (LTOBC, 212). If "ought" is indeed non-representational, it can very well be considered a feature of the "most basic syntactic form" of the sentence.

Let us recapitulate. A sentence embedded in an intentional context attributes to the subject the attitude which unembedded tokens of the *same* sentence is supposed to produce in that subject. Two sentence tokens count as belonging to the *same* type when 1) their referring expressions refer to the same things and 2) their non-referring features have the same stabilizing function.[85] Since, by hypothesis, the stabilizing function of "ought" is to produce attitudes with different descriptive contents in different addressees, it can also be used to attribute attitudes with different descriptive contents in different addressees. So GENERAL NORMATIVE is vindicated.

[85] Note that this is not a *different* principle from the one governing attributions of descriptive beliefs, but simply a generalization of that principle. Applied to attributions of regular descriptive beliefs, the generalized principle generates, as a special case, the principle that the embedded sentence indexes a belief with its own descriptive content.

4.2.1. Attributive Types

With DND, we have an account of "ought" that looks to be compatible with GENERAL NORMATIVE. But now the question is whether this package (which I will also call "DND") can actually account for the "cognitivist" features of ought-discourse, which is what we were hoping for. The prospects, at first sight, look bleak. Indexical relativism suffers from the problem of disagreement, and DND looks no better in this regard. If I think that *A* ought to φ, and you think that *A* ought not to φ, we seem to be disagreeing—but according to DND, our respective attitudes attribute different properties to *A*'s φ-ing. Furthermore, DND seems incapable of explaining why we attribute truth and falsehood to ought-claims, since it predicts that these claims lack descriptive content.

But I think this package can in fact account for the "cognitivist" features of ought-judgments and the statements used to communicate them, while at the same time accounting for some of the ways in which such judgments and statements notoriously elude treatment as purely descriptive representations. Most of the remaining chapter will be devoted to my case for this claim, starting with the issue of disagreement.

Let us begin by looking at one specific issue. GENERAL NORMATIVE entails that when you and I both think that *S* ought to φ (that is, when our respective token attitudes are both correctly attributed using the locution "he/she thinks that *A* ought to φ") these respective attitudes have different descriptive content. It is natural to think that the situation described by "you and I both think that *A* ought to φ" is also one that could be described as "you and I both think the same thing" or "you and I both share (tokens of) the same judgment (type)." But if DND entails that these token judgments have different descriptive contents, can it account for this intuition?

Clearly, there are many respects in which our respective judgments are "the same," many types they are both tokens of. One that comes immediately to mind is that they both have descriptive contents that *relate the same way* to their respective subjects: they both say that their respective subjects can expect to benefit from *A*'s φ-ing. By the same token, they have *functions* that relate the same way to their respective subjects: to motivate that subject to make *A* φ. But we may wonder why our intuitions about what constitutes "thinking the same thing" would track *these* relationally defined types when, in the case of regular descriptive beliefs, they straightforwardly track sameness of descriptive content.

Another thing they have in common is that both of them are attributed by and (as per Millikan's theory of attitude attribution) Normally produced by tokens of the same sentence type. Recall that we said that two sentences belong to the same type, for purposes of attitude attributions, when 1) their referring expressions refer to the same thing and 2) their non-referring features have the same stabilizing function. Now, suppose that *this*—being at-

tributed and Normally produced by the same sentences—is the sameness-relation that our intuitions about when two people "think the same thing" track. This would make immediate sense, since it allows us to describe the state of affairs where two people think the same thing by saying "they both think that *P*." And it would also be a general principle that subsumes both sameness of *belief* and sameness of ought-judgment.

We could call the attitude type that two people share tokens of, when their attitudes can be attributed using the same sentence, an "attributive type." *Why* would our intuitive sameness-judgments track sameness of attributive type, rather than, say, sameness of descriptive content? A first pass at an answer might go like this: people who harbor tokens of the same attributive type are the same *for purposes of discourse*. More precisely, if I utter a sentence, two listeners will both count as having responded in the Normal way—as doing their part in fulfilling the stabilizing function of the sentence—insofar as they come to instantiate token mental states of the same attributive type. Moreover, the attitude Normally *produced* by a sentence will also typically be the one Normally *expressed* by that same sentence. By "expression," let us understand the relation that obtains between an utterance of a sentence and the state of mind the speaker must Normally be in when uttering that sentence. At least for descriptive assertions, identity statements, and ought-statements, the state of mind *expressed* seems to belong to the same attributive type as the state of mind Normally *produced*. And there is good reason to suppose that this pattern generalizes. If I am told something, then insofar as this episode of communication proceeds Normally, I should be able to go on to Normally say that same thing again, expressing the attitude that was just conveyed to me, by uttering the same sentence.

Let us posit that *attributive types are what statements, in general, are designed to make people share tokens of.* Different types of statements are designed to make people share tokens of differently-individuated attitude types. This should not be surprising. Sometimes, in speaking, we are interested in changing how our interlocutor represents the world. Sometimes, we are interested in making him share our goals. This is indeed the assumption behind the standard Millikanian account of the difference between indicative (assertoric, descriptive) and imperative (directive) discourse. The notion of attributive types give us a way of talking about "what gets shared" in successful discourse that abstracts from these differences.

As Gunnar Björnsson has suggested, relations of sameness of attributive type can also be used to account for intuitions about *agreement* and *disagreement* (Björnsson 2015). Intuitively, what happens when two people come to respond Normally to the utterance of a sentence is that they come to *agree*—with the speaker and with each other. Conversely, when a subject harbors an attitude that would be rationally incompatible with responding Normally to an utterance, she intuitively *disagrees* with the speaker. Björnsson calls his view "discursivism" about disagreement, because it explains

agreement and disagreement in terms, roughly, of whether subjects can participate together in episodes of Normally successful discourse.

I think these observations are sufficient to make sense of DND's implication that our intuitions about what constitutes "thinking the same thing" often track sameness of attributive type rather than sameness of descriptive content. In virtue of sharing tokens of the same descriptive type, subjects are, as it were, potential allies in a joint discursive endeavor. This, by the way, is an account of "thinking the same thing" that I believe has merit regardless of whether DND itself turns out to be correct. It makes sense that we, as social beings, should sort ourselves as "same" and "different" based primarily on our potential as allies in cooperative endeavors.

Next, I want to look at whether this same framework can be used to make sense of our practices of attributing *truth* and *falsehood* to ought-claims.

4.3. Truth and Assessment

The notion of descriptive content is supposed to explain our attributions of truth and falsehood, correctness and incorrectness, and so on, as well as the phenomena these attributions supposedly pick out. But according to DND, token normative statements do not have determinate descriptive contents, so the prospects for explaining our attributions of truth-values to these statements seem dim. Token normative judgments *do* possess descriptive contents. But we do not, in fact, assign truth or correctness to attitudes on a token-by-token basis.

According to philosophical received wisdom, the "primary truth-bearers" are *propositions*. Propositions are not token representations. They are either themselves contents or akin to contents, in that they are features of token representations or abstract entities to which token representations bear relations. According to the received view, when we say things like "what he said was true" or "what he believes is true," we are not attributing truth to the token statement or belief as such but to its propositional content.

Though I haven't emphasized this issue thus far, teleosemantics is often discussed in terms that seem to commit it to the contrary view that contents are not themselves truth-bearers, but merely features of token representations in virtue of which the latter are true or false. But this might have seemed like a fairly minor, perhaps even a purely notational distinction. After all, if representations with the same descriptive content are true or false together, it is but a short step to the view that the "primary" truth-bearer is an abstract entity to which these jointly true or false representations all bear a relation.

But this conciliatory strategy won't work if DND is correct. Suppose I tell you:

> Ingemar and Konrad both think that Axel ought to get a haircut, but that's false.

Here, I'm presumably not attributing falsehood distributively to Ingemar's and Konrad's respective token judgments. If I were, why would I refer to the truth-bearer using the singular "that" rather than the plural "those"? Nor, if DND is correct, am I attributing it to any abstract entity these token judgments are both related to *in virtue of* sharing the same descriptive content. By hypothesis, they do *not* share the same descriptive content. The conventional answer is that I'm attributing it to the *proposition* that is the joint object of Ingemar's and Konrad's respective beliefs. But if so, and if DND is correct, sameness of propositional content cannot track sameness of descriptive content.

What the above observation *is* consistent with is that sameness of propositional content tracks sameness of *attributive type*. This is an intriguing observation, but so far, we don't know what to do with it. The question remains: how can attributive types determine truth-values if they are not descriptively individuated?

We need not remain wedded to the assumption that a truth-value is something a representation has in virtue of its descriptive content. There are many theories of truth that challenge the idea that truth is essentially a matter of describing the world correctly: quasi-realism (Blackburn 1993), minimalism (Wright 1994; Horwich 1998), and other related "deflationary" accounts come readily to mind. To abandon the correspondence theory of truth may seem like a very un-Millikanian move. But we should keep in mind that the data we are asked to explain here concern, primarily, the use of terms like "true" and "correct" in natural language, and it would seem as though the appropriate teleosemantic approach to data like these would be to inquire into the *stabilizing function(s)* of the terms in question.[86] And indeed, Millikan suggests (albeit in passing) an account of one stabilizing function for "true" that has a rather deflationary ring:

> "It's true that p" is, I believe, typically just redundant, but often used when the speaker also, or still, or decidedly, believes that *p* or is urging acceptance that *p* against actual or possible doubt. Speaker confidence tends to produce hearer confidence, so that may be a stabilizing function that "true" has. (Millikan 2018, 240)

What *would* not only be un-Millikanian, but would also threaten to call into question the whole point of the teleosemantic enterprise, is a theory that severed *all* ties between descriptive content and truth. Surely, at least in the case of straightforward descriptive representations, attributions of "truth" and "falsehood" Normally follow descriptive correctness.

[86] I'm grateful to Gunnar Björnsson for pressing this point on me.

I said above that it seems as though sameness of propositional content tracks sameness of attributive type. Attributive types, I also said, are what statements are designed to make people share ("designed" in the etiological sense, of course). Assertoric statements are designed to make people have the same beliefs (i.e. beliefs with the same descriptive content). Ought-statements, we might say, are designed to *direct people the same way*. If sameness of truth-value follows sameness of propositional content, which in turn follows sameness of attributive type, we should expect "truth" to have something to do with this purpose of getting people to share something. And indeed, if Millikan is to be trusted, assertoric statements are specifically designed to make people share, not any old beliefs, but the *true* ones (p. 62).

If "true" in the assertoric case denotes a property characteristic of those beliefs that assertoric statements are designed to share, then we might expect the same to hold in the normative case. *Which* way are ought-statements designed to direct people? A natural answer is: the good (or right) way. But this answer employs unanalyzed normative notions, and moreover, there is no reason to suppose that there is a *single* way that ought-statements are designed to direct people. If normative discourse is cooperative in the way descriptive discourse is, their function should be to direct the addressee in a way that is good for *him*, and there is no reason to believe that a way to be directed that is good for one person must be good for another.

I will try to make sense of all of this by hypothesizing that there are *two* distinct ways in which we can assess a statement, *both* of which have some claim to being thought of as assessments of the statement's "truth" or "correctness." In the case of regular descriptive beliefs, these two forms of assessment agree, in the sense that we can never Normally assess as *true* in one way what we assess as *false* in the other, or vice versa. But due to the non-descriptive individuation of normative attributive types, the two come apart in the normative case. The two are:

1. Assessing a statement for purposes of determining whether we personally should accept it (what I will call "assessment-for-acceptance").
2. Assessing a statement for purposes of determining whether the speaker, in making it, evinces discursive competence (what I will call "assessment-for-speaker-competence").

I will explain these in turn, and how they can be thought of as ways of assessing a statement's truth, starting with assessment-for-acceptance.

Consider the situation of an *Assessor* who is being courted by two distinct groups of people *A* and *B*. The members of *A*, Adam, Alice and Art, want *Assessor* to think that Axel ought to get a haircut, whereas the members of *B*, Barbara, Bonnie, and Bert, want her to think that Axel ought not to get a haircut. The two judgments are mutually inconsistent, and *Assessor* can rationally adopt at most one of them. Both groups are trying to make *Assessor*

share something (what I have called an attributive type), and *Assessor* must assess *what they are trying to make her share*. What she assesses, then, is not the token judgments in each of Adam's, Alice's, Barbara's etc. heads, but the types themselves: *whether* she should adopt a token of either of those types and, if so, *which* type she should adopt a token of.

I would be natural for *Assessor* to describe her assessment and eventual choice in terms of correctness, agreement, etc. In making her choice, *Assessor* is trying to determine which group to agree with (this was Björnsson's point, as reported above). She could report her decision by saying, about one of the groups, "they're right" or "what they say is true." This, I submit, is one stabilizing function for "true," "correct," and related terms. They convey acceptance of a claim to its speaker (and anybody else who cares to listen), where acceptance consists in forming a token of the attributive type that the claim is supposed to make her share. When "true" is used in this way, it reports the result of the first form of truth-assessment: assessment-for-acceptance.

Now, when the opposing claims are regular descriptive ones, each of the token attitudes belonging to each member of A and B respectively, as well as the statement whereby they convey those attitudes, all have the same descriptive content. In this situation, *Assessor* will Normally base her assessment on which of these two descriptive contents obtain. It is therefore natural to think of her assessment as an assessment of these descriptive contents themselves. Here, then, truth *qua* the standard that guides the assessment follows descriptive correctness.

But in the ought-case, by hypothesis, neither claim is uniquely associated with any *single* descriptive content. So on what basis should *Assessor* make her assessment? On the basis, I submit, of whether *her own token*, the token judgment she is invited to adopt, would be descriptively correct. That, after all, is what is relevant to whether accepting the claim would be a Normally successful move for her.[87] If this is correct, the descriptive content of her own token functions as a standard or norm for her truth-assessment.

What have we concluded? Truth-assessment as assessment-for-acceptance entails a notion of truth as a standard for evaluating attributive *types*, tokens of which may or may not share descriptive content. But descriptive content is still *relevant* for the assessment, since it is based on the descriptive content of the token judgment that constitutes the assessor's acceptance of the claim. This account has two virtues: it explains why truth-ascriptions follow attributive types even when these are not descriptively individuated, and it

[87] Again, we could say the same thing about regular descriptive claims. In evaluating whether to believe a descriptive claim, I should consider only whether *that* belief, the token belief that would constitute *my* acceptance of the claim, would be descriptively correct. In the case of descriptive assertions, the descriptive content of *my* token will be the same as the descriptive content of the tokens in the speakers' heads and of the statement itself, but at bottom the structure is the same.

preserves the desired connection between truth-assessment and descriptive content in the case of regular descriptive claims while also showing how this connection generalizes to the normative case. But it buys these virtues at the price of making the truth of normative claims *assessment-relative*.

Although this is a non-standard view, it is not without proponents. John MacFarlane (2014, chap. 11) has defended the view that deontic statements (among other kinds of statements) are what he calls *assessment sensitive*: their truth-value is sensitive, not only to the context in which they are uttered, but to the context in which they are assessed. MacFarlane is not a teleosemanticist, and he doesn't explain assessment sensitivity by appeal to the theoretical machinery of descriptive content, stabilizing function, and attributive types. But his meticulous defense of the coherence and explanatory force of the idea of assessment sensitivity lends some independent support to a teleosemantic theory that predicts the same phenomenon.[88] As we will see, however, MacFarlane's theory, though it may help illuminate the picture I have presented, also differs from it in ways that will be relevant to the discussion to follow.

MacFarlane's theory implies that the truth-predicate, when applied to assessment-sensitive claims, indexes two contexts: a context of utterance and a context of assessment. Such a sentence is never just true: it is "true as used at c_1 and assessed as c_2," where c_1 and c_2 range over contexts of use and contexts of assessment respectively. We can readily make sense of a sentence's truth-value being relative to a context of use: an indexical expression has different referents in different contexts of use, hence pick out different truth conditions. But what does it mean for a sentence's truth-value to be relative to the context of assessment?

We have suggested that truth can be understood as a norm for assessing a sentence for potential acceptance. MacFarlane, too, suggests that truth can be understood as supplying *norms*. This, explicitly, is not to be understood as a *definition* of truth (2014, 98), but only as a characterization that helps illuminate the role that truth plays in thought and language. The first normative role for truth that MacFarlane considers is not, however, truth as a norm for *acceptance*, but as a norm for *assertion*. MacFarlane suggests that the following rule governs the practice of assertion:

> **Truth Rule.** *At a context c, assert that p only if p is true at c*. (MacFarlane 2014, 101)

MacFarlane draws on the analogy of Truth Rule in order to illuminate the idea of assessment-relative truth. Note that if Truth Rule were the only nor-

[88] It should be noted, however, that in his discussion of "ought," MacFarlane focuses exclusively on data that suggest they are sensitive to the information available to the assessor. He does not, as I have done, suggest that they might also be sensitive to what would benefit the assessor.

mative role that truth played, it would be hard to understand why truth should be assessment-sensitive. Indeed, Truth Rule makes no distinction between a context of utterance and a context of assessment.

MacFarlane suggests that truth supplies *two* norms, a norm for *assertion* and a norm for *retraction*. The Reflexive Truth Rule is a generalization of Truth Rule to doubly relativized truth:

> **Reflexive Truth Rule** *An agent is permitted to assert that p at context c_1 only if p is true as used at c_1 and assessed from c_1.* (MacFarlane 2014, 103)

Reflexive Truth Rule, in other words, permits assertion only if the assertion would be true as used *and assessed* from the same context, the context of assertion itself. Since the context typically involves the speaker herself, this means that she is permitted to make an assertion only if that assertion would be true *as assessed by herself*.

If Reflexive Truth Rule were the only normative role that truth played, that would still be insufficient to explain assessment sensitivity. Assessment from *other* contexts than the speaker's own have no role to play here. MacFarlane thus proposes a second rule governed by truth, a *retraction* rule governing the circumstances under which an agent is obligated to retract a statement previously made (MacFarlane 2014, 108). More resonant with our concerns, however, is the alternative approach MacFarlane also suggests: to understand assessment-relative truth as a standard for *rejection*:[89]

> **Rejection Rule.** *An agent in context c_2 is permitted to reject an assertion of p made at c_1 if p is not true as used at c_1 and assessed from c_2.* (MacFarlane 2014, 110)

Here, c_1 is the context of assertion, and c_2 is the context of assessment. Given that "rejection" is just the converse of "acceptance," Rejection Rule seems to hew very close to the normative role for truth I have proposed above.

There are some compelling parallels between MacFarlane's account of assessment-relative truth and my own idea of truth-assessment as assessment-for-acceptance. But there are also important differences. First, MacFarlane suggests that the Reflexive Truth Rule and the Rejection Rule are *con-*

[89] Though MacFarlane makes nothing more of this suggestion, and builds the rest of his theory on the retraction rule approach, he doesn't see it as a rival of the latter but as compatible with it. The reason MacFarlane favors the retraction approach is because he thinks of his rules as norms for *speech acts*—presumably because he takes it that norms must be norms for *acts* rather than, say, relations to propositions—and "it is clearer that there is a speech act of retraction than that there is a speech act of rejection" (MacFarlane 2014, 111). But we can conceive of acceptance as an *act* without conceiving of it as a *speech act*: it is the act of forming an attitude in response to a speech-act, thereby doing one's part in fulfilling the pattern that constitutes the Normal performance of the speech-act's stabilizing function.

stitutive of the speech-act of assertion. In other words, he holds that something is an assertion only insofar as it is governed by these two rules (MacFarlane 2014, 101). We have already seen (p. 64) that Millikan rejects the view that speech-act types are defined by constitutive rules. Relatedly, MacFarlane formulates his rules in terms of the deontic notion of permission. I, however, have only claimed that an assessor will Normally be guided by the truth, understood as her standard for assessment-for-acceptance, of the claim in deciding whether to accept it. Nothing necessarily follows about the deontic status of the assessor. As we saw already on p. 64, the relationship between Normalcy and deontic norms for agents is complex at best. If I described truth as a "norm" for acceptance above, this should therefore be understood in the looser sense that also accommodates teleological norms.

These differences are surely symptomatic of the fact that MacFarlane is skeptical of the possibility of giving a definition of truth (MacFarlane 2014, 98). His rules are not intended as indirect definitions of truth, but merely as ways of illuminating the nature of truth by indirectly tying it to certain normative notions. Within the teleosemantic framework, however, we already possess a definition of truth as descriptive correctness: the obtaining of the most proximate Normal conditions for a representation's proper function. This is a definition that makes no reference to agent-level normative notions. My suggestion, to understand truth-assessment as assessment-for-acceptance, is essentially in the same spirit as this standard teleosemantic truth-definition.

There is another difference between MacFarlane's account and my own that stands out. I have, as of yet, said nothing that could be interpreted as an endorsement of MacFarlane's view that truth also provides a norm for *assertion* reminiscent of Reflexive Truth Rule. And it is to this issue that I would now like to turn my attention.

4.4. Assessing Speakers

Consider again the Reflexive Truth Rule. What sort of assessment is involved when I assess whether an agent has abided by that rule? By MacFarlane's light, I am assessing whether the agent was permitted to make his claim. But I am not, of course, trying to determine whether the claim is acceptable *to me*, i.e., whether I should accept it. If MacFarlane is right, I could very well reject a claim while holding the speaker faultless for having uttered it.

I believe there is indeed a phenomenon like this, something that is sometimes discussed under the heading of "faultless disagreement." I disagree with a claim (or at least refuse to accept it), yet do not fault the speaker for having uttered it. But there are difficulties with accounting for this phenomenon within the present framework.

When assessing-for-acceptance, I assess whether the token judgment that would constitute *my* acceptance of the claim would be descriptively correct. It is tempting to think that the other kind of evaluation, what we may think of as a "disinterested" evaluation of the speaker's performance, is an evaluation of whether the judgment the speaker *expresses—her* token of the attributive type—is descriptively correct. That would produce an account pleasingly parallel to MacFarlane's: MacFarlane's Rejection Rule tells us to reject judgments that would be true in the context of *our* assessment, while the Reflexive Truth Rule tells us that speakers are permitted to make judgments that are true in the context of *their own* assessment.

However, on p. 64, I expressed the tentative view that insofar as agents have a deontic responsibility to uphold Normal conditions for speech-acts, they derive from the cooperative nature of communication: a speaker's responsibilities derive from an underlying responsibility to be a good cooperative partner, and this typically involves ensuring that Normal conditions for discourse obtains. Now, the fact that the speaker's judgment token is descriptively correct is no guarantee that the communicative episode as a whole will go off Normally. Not every Normally produced judgment is a judgment that it would likewise be Normal for others to accept (a token of the same attributive type as). It seems like *insofar as* there is a responsibility on the part of the speaker to contribute to Normal communication, it should primarily consist in a responsibility to ensure that Normal conditions obtain for the token judgments she is trying to make her *audience* harbor—not so much that they obtain for her own token judgment.

If we assume, with Millikan and *contra* MacFarlane, that speech-act types are defined not by the constitutive rules that govern them but by their functions, we also do not need to find a *single* rule or set of rules governing assertion or any other speech-act. Consequently, we can endorse the natural idea that there are a number of different standards against which a claim can be evaluated, which will be appropriate under different circumstances and for different purposes. For instance, in assessing whether the speaker's judgment token is correct, we are assessing whether her own capacity for normative judgment is in good order, which can often be useful. The considerations in the last paragraph suggest, however, that the most salient standard for assessing a speaker is typically by how well she adapts to her audience.

In the cases that we are usually most invested in, that audience will include ourselves. When this is the case, assessing whether a speaker has adapted to her audience *involves* assessment-for-acceptance. We may also, however, assess a speaker from the perspective of *other* members of the audience to see if the speaker has managed to successfully adapt to *them*. We can expect this vicarious mode of assessment to become particularly salient when we are not ourselves part of the intended audience, such as when we make a retrospective evaluation of a past speech-act. To assess speakers in this way can be useful to us, even if we are relatively unconcerned about

whether we ourselves should accept the claim made, because it helps us keep track of who is a competent and responsible speaker, one who can successfully participate in maintaining cooperative discourse.

This form of vicarious speaker assessment constitutes the second type of assessment I mentioned on page 120 above as a form of truth-assessment, i.e., assessment-for-speaker-competence. By "competence" I mean something like the willingness and ability to participate responsibly in cooperative communication. Note that in the case of descriptive assertions, there is no real difference between "vicarious" assessment and assessment-for-acceptance: if a descriptive claim is acceptable to another, it is *ipso facto* acceptable to us, because all tokens of an attributive belief-type have the same descriptive content. In the case of ought-claims, which are non-descriptively type-individuated, they can come apart.

I suspect that this phenomenon could be partly responsible for widespread relativist intuitions about normative claims: the sense many of us have, philosophers and laymen alike, that, for example, normative claims made by members of cultures very different from ours should not be judged by our standards. Since we are very far from being the intended audience for these claims, it is natural that assessment-for-speaker-competence should be more salient when evaluating them than assessment-for-acceptance. Of course, such speakers can be highly competent in adapting to their *intended* audiences even if *we* feel no temptation to accept their claims.

In any case, I believe these observations about the importance of a competent speaker adapting to her audience allows us to address an issue that has haunted GENERAL NORMATIVE since section 4.1, namely the issue of its conspicuous self-centeredness, the fact that it specifies a descriptive content for ought-judgments that concerns only the judger's own benefit. Let us turn to this issue now.

4.5. Communities of Judgment

Even granted that I have managed to account for truth-attributions to normative claims given GENERAL NORMATIVE, the latter still leaves some worries open. Many statements and judgments of the form "*A* ought to φ," if they are about anybody's interests, seem to be neither about the speaker's nor about the hearer's interests but rather about either *A*'s interests or the interests of people in general, the common good. How can we account for this intuition given GENERAL NORMATIVE?

The answer to this questions inheres in the fact that Normal directive discourse—and we are supposing that normative discourse is a species of directive discourse—requires a certain *commonality of interests* among speaker and hearer.

I suggested above (p. 120) that ought-statements can be understood as means to "direct people the same way." To direct people a certain way, I must typically give them a reason to be directed that way or otherwise trust that they will appreciate the reasons they already have (I can also try to deceive them, of course). In this respect, ought-discourse—as can be expected, given our assumptions—is akin to *imperative* discourse as Millikan understands it:

> It is worth specifically noting some common alternative further functions of the imperative that directly but alternatively reinforce hearer acts of interpretation and compliance. Most of these functions are of interest to the speaker too, the speaker and the hearer both aiming, as it were, to get the same thing out of the imperative. This usually happens because one of these partners takes an interest in the interests of the other. In such cases, these functions count as further alternative *stabilizing* functions of the imperative, since the performance of these functions helps to account both for the proliferation of speaker uses and for the proliferation of cooperative hearer responses to imperatives. Alternative hearer motivations for complying with imperatives, when viewed as purposed or as ostensibly purposed by *speakers*, correspond to certain so-called "illocutionary acts" that may be performed by speakers uttering imperative mood sentences. Not all of these functions are stabilizing functions. But three out of the four that come most readily to mind are. Described as "illocutionary acts," these four are *giving orders, making requests, giving advice,* and *giving directions* ("turn left here," "add the sugar slowly," etc.) (LTOBC, 57)

Let us begin by looking at the simplest case, namely statements of the form "you ought to φ," which are most immediately akin to imperatives. For my utterance of a token of "you ought to φ" to be able to perform its stabilizing function Normally, you and I must *both* have a stake in your φ-ing. My own stake is more or less implied by the fact that I choose to speak (though I can, of course, misjudge where my own best interests lie). But to be a cooperative speaker, I must try to ensure, before I speak, that you have a stake as well.

Of the four kinds of illocutionary acts Millikan mentions, the two that seem closest to what we usually do with ought-statements in the second person is "giving advice" and "giving directions." About advice-giving, Millikan says:

> When advice is delivered in the imperative mood […] the speaker purposefully advocates ends that the hearer does well to adopt in the hearer's own interests. Furthering the hearer's interests in this manner is an alternative stabilizing function of imperatives. (LTOBC, 58)

About directions:

> When directions are given in the imperative mood, the hearer is, Normally, already motivated towards a goal for which the imperative supplies a means. The

> speaker (or writer) is also motivated to help the hearer (or to help hearers or readers generally) to reach this goal. (Ibid.)

In both these cases, Normal communication requires that speaker and the hearer share some interest. So if the attitude that the speaker expresses describes *her* own interests, it is at the same time only in virtue of the fact that those are also the hearer's interests that Normal communication can be accomplished. In other words, Normal normative communication requires *commonality of interest*.

Let us observe that commonality of interest can come about in several ways. It can be that the speaker and hearer both have an independent stake in some outcome. It can also be that the speaker, though she has no personal stake in the outcome, has a stake in furthering the hearer's interests because of altruistic concern, because she has need of the hearer, or both.[90]

The speaker's and hearer's commonality of interest can also be due to the fact that they both have a stake in the welfare of some third party or larger group. This can again be because of altruistic concern, because they need those people or are themselves members of that group, or all of the above. They may be parents sharing an interest in the welfare of their children, soldiers sharing an interest in the welfare of their platoon, or citizens sharing an interest in the welfare of their nation. Or they may just be human beings sharing an interest in the welfare of human beings generally, because they sympathize with other humans and because they themselves are human.

About requests, Millikan says the following:

> In the case of Normally and properly functioning requests, the hearer must already be motivated to further the speaker's interests as such, either as an end or as a means to further ends. (LTOBC, 57)

Requests, then, are the converse of advice. In making a request, the speaker Normally relies on the concern the hearer is presumed to have for *her*. Intuitively, this type of speech-act is less commonly performed using constructions with "ought." There is something strangely insincere and self-serving about using "ought" when you are speaking as a representative of your own interests. The same is true of orders:

[90] It can be hard to distinguish between these two. We tend to feel more concern for those who are close to us, and those are also the ones upon whom we most depend for fulfilling our emotional as well as our physical needs. According to one plausible theory of the evolution of human altruism, our propensity to feel altruistic concern even for non-kin evolved as we came to adopt the subsistence strategy of *obligate collaborative foraging*, requiring the cooperation of several individuals to bring down large prey. For creatures dependent on this subsistence strategy, their (inclusive) fitness will be tied up not only with the welfare of their genetic kin, but also with that of non-kin group members whose assistance is required in foraging. That would have produced selection pressure in favor of altruistic concern for non-kin (Tomasello 2015, 44 ff.)

> In the case of Normally and properly functioning orders, it is within the speaker's control to invoke or apply sanctions to be sought or avoided by the hearer. (*Ibid.*)

Orders are the outlier among the four speech-act types discussed by Millikan, since they involve the speaker *producing* reasons for the hearer to comply rather than relying on his pre-existing interests, by directly dangling the promise of rewards or the threat of punishment over him.[91] In such cases, using "ought" has a darkly ironic ring to it.

How do we explain this contrast? Perhaps ought-statements simply have more narrow stabilizing functions than imperatives, ones that Normally let them be used to perform some of the speech-acts that can be performed using imperatives but not others. But perhaps we can give a more principled explanation.

In addition to second-person ought-statements, which are naturally interpreted as conveying directions or advice, there are also third-person ought-statements, which are not. If a judgment of the form *A ought to φ* has a motivating function, I have said, it is probably to motivate the judger to contribute to *making A* φ. But clearly, we often express third-person normative judgments under circumstances where neither we nor our audience have any direct influence over the actions of *A*, and there seems to be nothing inappropriate or infelicitous about that, as there would be, under those circumstances, about a direct command to make *A* φ.

But even if the participants in the discourse have no direct control over *A*'s behavior, they may still have an *indirect* control over it simply in virtue of the fact that they are social beings inhabiting a shared social word who have the power, *using ought-statements*, to direct people the same way as themselves by spreading the judgment that *A* ought to φ. Though my addressee may not have direct influence over *A*, he may still be able to spread the judgment to further people, and they to yet further people, until someone is reached who *does* have direct influence over *A*—perhaps even *A* herself.[92]

The chances of this strategy being successful increases the more people can be roped into the resulting *community of judgment*. This is for two reasons. One is the aforementioned, namely that the larger the community of

[91] Millikan also suggests that orders do "not correspond to a stabilizing function for the imperative" (1984, 57), but I am not sure what her basis for this claim is. If you were consistently rewarded for obeying orders, or punished for disobeying them, this might very well come to stabilize your disposition to obey them.

[92] Of course, this account only takes us so far. Some people are categorically outside the reach of even indirect influence, for instance, because they are dead or because the actions evaluated lie in the past. If we are to explain the felicity of a statement like "Napoleon ought not to have invaded Russia," we must understand the motivating function of the conveyed judgment some other way than as a motivation to make Napoleon not invade (or not have invaded) Russia. Perhaps this statement is an indirect way to convey a more general practical principle, a policy not to behave like Napoleon did when invading Russia (I take it that a view of this kind is the one defended by Allan Gibbard in (Gibbard 1990)). If we followed this approach, we would have to acknowledge a certain amount of polysemy in the term "ought."

judgment, the greater the chance that one of its members will have power over *A*. The other is that a large community of judgment will have a certain momentum of its own: the larger the community, the more pressure on undecided people to join it so as not to antagonize its members. We will be discussing these kinds of dynamics in some detail in the next two chapters.

For a third-person ought-judgment to be able to perform its function in this manner, it seems like the prescribed state of affairs, i.e. that *A* φs, must Normally be a source benefit not only for the speaker and her immediate audience, but for others as well. There are no hard and fast rules for which others, and how many, must be potential members of the community of judgment. This will vary with the context and the content of the prescription. Irrespectively, the speaker must often look forward, not only to what benefits and motivates her audience, but to what can be expected to benefit and motivate some arbitrary members of the larger population around her. In many cases, she will *not* be able to rely either on her ability to threaten these arbitrary others with punishment or entice them with reward, or on their concern for her, the speaker (they might not even know that she originated the judgment). She will have to rely on more universal motivational principles: the concern these others have for their own personal welfare, for the welfare of *A* herself, or for the advancement of some cause they all have in common. She will have to assume a "general point of view." Possibly, this can explain why ought-statements cannot be felicitously used to make requests or give commands, even in the second-person case.

DND entails that, as a *general* matter, ought-statements have no determinate descriptive contents and are assessed-for-acceptance on standards that are idiosyncratic to each assessor. But the above considerations mean that these statements will nevertheless often acquire a semblance of intersubjectively valid truth-conditions, at least locally. In addition to their immediate stabilizing function of conveying ought-judgments, such claims also have the more ultimate function of making it the case that *A* φs. When *A* is beyond the direct influence of the speaker, the latter must rely on sufficiently many others sharing her assessment in order for this further stabilizing function to be performed. This, I posit, Normally requires that *A*'s φ-ing would be beneficial for all of these sufficiently many others. In other words, it requires that *A*'s φ-ing contributes to the "common good," at least if the latter term is understood in a broad way, as including the good of whichever community would be sufficient to make *A* φ.

Do these observations suffice to answer the questions with which we began this section? In part, yes. They explain why, when someone *expresses* an ought-judgment, we typically expect her to have in mind not just her own interests, but those of the judgment's object, her audience, and often enough, the larger community as well. But it leaves open the possibility that she can harbor a Normal third-person ought-judgment prescribing a course of action

that benefits only herself, even if she could not Normally express that judgment using an ought-statement.

We may ask, however, in what sense an attitude really is an *ought*-judgment if it cannot be Normally expressed by a statement employing the term "ought" or its synonyms. Perhaps we should deny the status of ought-judgment to attitudes that are not formed under the influence of considerations about what would also be acceptable to others and could secure the support of a community of judgment. This would also make it clearer what distinguishes a *first person* ought-judgment, a judgment about what *I* ought to do, from a mere desire or intention. In trying to determine what I ought to do, I deliberate on my own course of action with an eye to what others could also accept.

Here, I think, we encounter a recognizable picture. Intuitions about normativity in general and morality in particular are pulled between two opposing poles. On one extreme, we find a picture of normativity as pure, arbitrary exercise of will (perhaps by the sovereign, or God, or each individual moral judge). On the other extreme, we find a picture of normativity as a set of universal, impartial principles that exercise authority over every rational soul. The present account situates the truth somewhere in between. But more than that, it provides an insight into the dynamics, for lack of a better word, that pull human judgment between these two opposing poles. The human will, to realize itself in the world, must secure the assent and cooperation of other wills, and to do so it must be sensitive to the demands that others place on it. In the process it is transformed into something more universal.

With these somewhat impressionistic reflections, we are entering territory that lies outside the remit of the present work. Similar motifs will, however, return in the next chapter, when I discuss the function of moral judgment.

I will end this chapter with brief meditations on two topics I have neglected so far: ought-sentences in unasserted contexts and indeterminacy.

4.6. Some Notes on Unasserted Contexts

We began this chapter with the ambition to account for some of the "cognitivist" features of normative discourse on the non-cognitivist assumption that normative judgments are directive attitudes, by exploiting the last chapter's conclusion that all attitudes, including directive ones, have descriptive content. In sections 4.3-4.5, I have discussed how we can account for the seeming truth-aptness of ought-statements on this basis. But so far, I have limited my account to attributions of truth to *atomic* ought-statements, whereas a complete account of the cognitivist features of ought-discourse should also account for what Cuneo calls its "aptness for inference," the fact that ought-sentences (and the judgments they express) stand in determinate logical relations to other sentences. These typically constitute syntactic transforms of

atomic ought-sentences (like negations thereof) or are produced by embedding the sentence in larger ones, such as conditionals.

Peter Geach influentially argued that there can be no exhaustive account of the meaning of a certain class of sentences purely in terms of the type of speech-acts they are used to perform when uttered unembedded, as long as those sentences can also appear in unasserted contexts such as the antecedents of conditionals. The antecedent of a conditional is not asserted, but it has the *same* meaning as a free-standing token of the same sentence, as evidenced by the fact that *modus ponens* is a valid inference, which presumably requires that the unembedded minor premise and the antecedent of the major premise have the same meaning (Geach 1960, 1965).

The point that Geach is pressing can be put in the following terms. The "meaning" or (to use more modern vocabulary) semantic value of a sentence must fill three theoretical roles. First, it should help predict how the sentence behaves when tokened unembedded, i.e., its pragmatics. Second, it should predict how the sentence contributes to the semantic value of larger sentences it helps compose, such as when it stands as the antecedent of a conditional. Third, it should help explain the sentence's inferential role, i.e., the inferential relations the sentence bears to other sentences, such as the fact that an unembedded token of a sentence, together with a conditional where that same sentence figures as the antecedent, can be used to perform modus ponens. Traditional non-cognitivists such as Ayer (1952), Stevenson (1944), and Hare (1952) supplied theories of the meaning of ethical terms that only accounted for the first of these.[93]

For regular descriptive sentences, the simplest theory of their semantic value is one that identifies it with the content of an unembedded utterance of it or, more precisely, with a function from contexts of utterance to contents.[94] If discursive non-descriptivism is true, ought-sentences and ought-statements have no content, if by content we mean *descriptive* content. So a theory like the one above, with "descriptive content" substituted for "content," cannot be the correct one for the semantic value of an ought-sentence. But that fact is not as such very discouraging, because Millikanian descriptive contents were never good candidates for semantic values of indicative sentences in the first place. Since two sentences can have the same descriptive content

[93] In recent decades, the expressivist semantics developed by Allan Gibbard (1990, 2003) has significantly improved the non-cognitivist's position in this regard.

[94] This theory is probably *too* simple, for reasons having to do with the "shiftiness" of certain sentential operators (Lewis 1981). For instance, the statements formed by uttering, respectively, "It's raining" and "it's raining here" in the same context presumably have the same content (the same truth-conditions), but they are not intersubstitutable *salva veritate*—as evidenced by the sentence-pair "No matter where Anna goes, it's raining" and "No matter where Anna goes, it's raining here"—and so they cannot have the same compositional semantic value (the example is from Packalén 2016, 14). But something in the vicinity is probably true. The semantic values of sentences, if not identical to functions from contexts to contents, are at least closely related to them.

while their inferential role and embedding properties differ (p. 71), descriptive contents cannot play the last two explanatory roles above. And they can't really play the first one either, since two sentences with the same content can have as their function to produce different beliefs (i.e. beliefs with different conceptual articulation).

The thing in the Millikanian arsenal that *does* seem to fit the bill is the *semantic mapping function* of the sentence. A sentence's semantic mapping function picks out its content and determines, at least jointly with the stabilizing function of its mood marker, how it is supposed to be translated into belief. It also determines which syntactic transforms it can be subject to and what semantic transforms they correspond to. As I pointed out in section 2.4, a waggle dance can have the same descriptive content as a sentence in human language, but only the latter can be negated or embedded in the antecedent of a conditional.

But clearly, the same problem arises here: if there is no particular state of the world the sentence is supposed to map, then there is no particular way it is supposed to map it. Yet we cannot for that reason conclude that there is *no* relation to the world that explains the past persistence of an ought-statement. All we can conclude is that this relation would have to be defined, in part, over the particularities of each individual consumer/assessor. Hence this relation could obtain in the case of one assessor but fail to obtain in the case of another, for one and the same sentence token. This is tantamount to saying that the sentence's semantic mapping function picks out different world states for different assessors. Which is of course more or less what we concluded earlier: for one and the same sentence token, it could be that Normal conditions obtain for its acceptance by one assessor but not for its acceptance by another.

These reflections suggest a DND-compatible semantic program for "ought": instead of assigning semantic values to ought-sentences that pick out (perhaps together with the context of utterance) specific descriptive contents, we assign semantic values that pick out (together with context, etc.) *functions* from assessors to descriptive contents, functions that map each assessor to the descriptive content of the judgment that would constitute her acceptance of the sentence.

Let us try to evaluate this suggestion. A program for supplying a DND-friendly compositional semantics for ought-sentences should meet the following constraint:

> (CONSTRAINT) For all potential assessors, if the ought-sentence *O* Normally produces the judgment *J* in that assessor, the attitude *F*(*J*) Normally produced in the assessor by a syntactic transform *f*(*O*) of *O* should have a descriptive content that stands in a modal relation to *J*'s descriptive content that explains the inferential relation in which *f*(*O*) stands to *O*.

In formulating CONSTRAINT, I have relied on the same assumptions I have already discussed on p. 100 above. Though modal relations between the contents of representations (like inconsistency and entailment) do not necessarily manifest themselves in epistemically transparent inferential relations between those representations, the inferential relations that *do* obtain between representations can be expected to reflect, and be explained by, modal relations between their contents.

For example, consider a *negated* ought-sentence "it is not the case that *A* ought to φ." The inferential relation in which it intuitively stands to "*A* ought to φ" is, of course, that of being its contradictory. What would *explain* this contradictory relation is if the descriptive contents of the two sentences, for any given assessor, were each other's modal complements, i.e. one obtains just in case the other doesn't. So if the descriptive content of *S*'s judgment that *A* ought to φ is, in accordance with GENERAL NORMATIVE,

> *A*'S φ-ing has the highest expected benefit for *S* out of the alternatives available to *A*,

then the descriptive content of *S*'s judgment that it is not the case that *A* ought to φ should be

> It is not the case that *A*'S φ-ing has the highest expected benefit for *S* out of the alternatives available to *A*.

My suggestion, to assign semantic values to ought-sentences that pick out functions from assessors to descriptive contents, seems capable of meeting CONSTRAINT. All that is required is that a syntactic transform on an ought-sentence can perform its standard operations distributively on the descriptive contents in the function's range, yielding a new function from assessors to descriptive contents that mapped every assessor to the relevant transform of the original descriptive content.

This is the merest suggestion for a program, but the approach seems promising to me. It leaves open, however, the question *which attitude* is Normally produced by the syntactically transformed sentence. We know that the sentence "it is not the case that *A* ought to φ" should Normally produce an attitude in *S* whose descriptive content is (assuming GENERAL NORMATIVE) *it is not the case that A's φ-ing has the highest expected benefit for S out of the alternatives available to A*. But which attitude is that? It doesn't seem to be, in itself, a motivational attitude. If anything, it seems to be the state of having discounted a possible motivational attitude. Here, further investigations are needed.

4.7. Indeterminacy Again

By way of concluding this chapter, I would like to return to the issue of indeterminacy that arose in section 4.1. Let us remind ourselves of the various potential indeterminacy problems unearthed by that discussion:

1. *Neandrean indeterminacy,* deriving from uncertainties over whether motivational states Normally track primary reinforcers or outcomes that the cognitive system treats as proximal signs of primary reinforcers.
2. What we may call *deliberative indeterminacy*, deriving from uncertainties over what counts as "available alternatives" for purposes of assigning descriptive content according to INTENTION and GENERAL NORMATIVE.
3. Indeterminacy deriving from uncertainties over whether teleosemantics is required to assign dual contents to ought-judgments on the model of Björnsson's (2018) account of epistemic modal judgments.

How worried should we be about these indeterminacies? That depends on what we want our theory to do. If we want it to yield determinate answers to questions like "what ought I to do," we have reason to be worried. But note that these worries are unlikely to be allayed within the present framework in any case. On many of the different possible ways to resolve the above indeterminacies, the descriptive content of an ought-judgment is what we may call "deliberatively sensitive," in that they can be descriptively correct when first entertained, but on further deliberation and investigation (perhaps revealing action-alternatives we had not previously entertained, or information that changes the expected benefit of an outcome) they become descriptively incorrect. If the descriptive content of an ought-judgment is deliberatively sensitive in this way, its descriptive correctness may not in fact be the sort of final arbiter of action we might have hoped for. If, on the other hand, we resolved the indeterminacies in such a way as to avoid all deliberative sensitivity, we would get impossibly strict descriptive correctness-conditions, requiring an action to be the objectively best one out of all possible alternatives. This move would sever the link between descriptive correctness and Normal success, as it is unlikely that people's normative judgments have very often, in the ancestral past, met these steep requirements—or have had to do so in order to persist.

We should therefore not expect that an appeal to the descriptive content of directive judgment should help us answer first-order normative questions. Rather, the account I have presented is an attempt to explain how people in fact use normative judgments and, in particular, why so many aspects of that use evinces "cognitivist" features. In particular, I have tried to explain why they are truth-apt and why they are expressed using indicative sentences that

embed like regular descriptive sentences. Will indeterminacy threaten our account of either of *these* phenomena?

Let us first consider truth-aptness. I have said that the descriptive correctness of an ought-judgment constitutes a standard used by an assessor to determine whether to accept that judgment and that a stabilizing function of "true" and "correct" is to signal such acceptance. The account only tries to explain why people *in fact* treat ought-judgments as truth-apt. It doesn't try to give determinate conditions for when they are entitled or obligated to accept such judgments (thereby answering first-order normative questions). Hence, it can survive some indeterminacy in the standards people use. It is not, after all, as if people in general are clear about the exact criteria they do or should use in evaluating normative judgments (or moral philosophers would be out of a job). What this doesn't allow us to do is to give determinate conditions for when an act of acceptance is Normally successful. However, I have already insisted (p. 75) that Normal success is something we should be able to think of as coming in degrees or respects. A normative judgment that picks out the best action-alternative out of some range, while ignoring an even better alternative because the subject was ignorant of the possibility, is clearly successful to *some* degree.

Next, let us turn to embedding. In the preceding section, I suggested that discursive non-descriptivism could assign functions from assessors to the descriptive contents of ought-judgments as semantic values for ought-sentences. But if ought-judgments have no determinate descriptive content, won't this strategy fail?

As formulated, yes. But we may be able to modify the account slightly and still get semantic values capable of playing all three theoretical roles mentioned on p. 132. For instance, supposed we picked one content candidate, i.e., one possible way of resolving the above indeterminacy worries, as the elements of the range of the function we assign as a semantic value to a normative sentence. Call it "content$_A$." As long as the content$_A$ of an ought-judgment is sufficient to uniquely specify that judgment, this allows our semantic value to serve its first theoretical role: to help predict the pragmatics of the attitude by predicting which ought-judgment it is supposed to produce according to its stabilizing function. Indeed, any principle that assigned determinate contents to ought-judgments out of the available candidates could be used to construct semantic values that would serve that theoretical role. We could pick one at random.

A problem arises only if two such content-candidates, content$_A$ and content$_B$, used to construct semantic values M_A and M_B for an ought-sentence in the above manner, would yield semantic values for *composite* sentences that could not, plausibly, be considered content-candidates for *the same* attitude. If so, the semantic values M_A and M_B would predict different stabilizing functions for the composite sentences. We would then need a principle for

excluding either content_A or content_B as a content candidate admissible for constructing semantic values.

But this problem, if it arises, may also solve itself. Provided that we have an intuitive grasp of what attitude the composite sentence is supposed to produce, we can exclude content candidates that fail to entail that it does.

So, for instance, suppose we have an assessor *A* and a sentence *S* whose stabilizing function when addressed to *A* is to produce in *A* the judgment *J*. Suppose we use content_B to construct a semantic value for *S*. Then the principles sketched in 4.6 predict that the sentence that results from negating *S*, ~*S*, will have the stabilizing function to produce an attitude *J** in *A* whose content_B is the modal complement of *J*'s content_B. Now, if that content_B cannot plausibly be taken to be the content_B of the judgment that is, intuitively, the contradictory of *J*—the judgment that ~S—then we can exclude content_B as a candidate for constructing a semantic value for *S*.

This is assuming a lot, and I have sketched the solution in only the most abstract of terms. At the same time, it is unclear whether the problem it is a solution to arises in the first place. As far as I can see, there is no particular reason to fear that we wouldn't be able to assign semantic values to normative sentences that performed the theoretical jobs we want them to perform, even though they would not correspond to determinate descriptive content for the corresponding judgments. But here, too, further investigations are needed.

4.8. Summary and Conclusion

In this chapter, I have attempted to explore and defend the view that ought-judgments are motivating or directive attitudes, and to account for their "cognitivist" properties—their truth-aptness and their embeddability and inferential aptness—by exploiting the universal hybridity thesis to assign descriptive content to them.

I think we must, in all fairness, conclude that the result has been, at most, a mixed success. Serious questions remain concerning indeterminacy and the semantics of normative sentences.

If there is a take-home message from this chapter, it is the apparatus of discursive non-descriptivism, with its ideas of attributive types as approximate counterparts for traditional propositions and of truth-assessment as assessment-for acceptance, as well as the observations in section 4.5 about the need for a commonality of interest for Normally successful normative communication. These ideas, it bears emphasizing, are independent of the *specific* proposal about the descriptive content of ought-judgments encapsulated in GENERAL NORMATIVE.

In the following two chapters, I will return to some of these ideas as I attempt to supply an analysis of *moral* thought and talk. I will split this analy-

sis into two parts, corresponding roughly to thought and talk respectively. In the next chapter, then, I give an account of the function and Normal operation of moral judgments that can serve as the basis of an account of moral discourse analogous to the one I have given for ought-discourse in this chapter.

5. Proper Function of Moral Judgment

A teleosemantic treatment of moral judgment and language presupposes that moral judgments possess proper functions. As the reader may recall from section 2.1, the proper function of an entity corresponds, roughly, to effects that its ancestors have produced that have explained the persistence of the lineage and hence the existence of the current entity.

Identifying the proper function of an entity is therefore an empirical matter. Since the actual causal processes that determine an entity's function are typically unavailable to direct observation, lost as they are in the past, the function must be inferred indirectly on the basis of its present activities. An account of an entity's past contribution to its own persistence, based on its present structure and activity, will of necessity be of a provisional nature. It should be accorded some credence, but we should be prepared to abandon it if it can be shown that it doesn't cohere, or if a better account can be given.

An exhaustive discussion of the proper function of moral judgments should therefore take the form of a comparative discussion of each candidate account. Here, I do not assume this task in its entirety. Rather, I will present *one* hypothesis about the function of moral judgment and discuss the consequences if it is true. I have opted for a rather conservative hypothesis, one that bears a close resemblance to established theories about the evolution of morality. I find it plausible, and I believe its consequences for a teleosemantic theory of moral discourse (discussed in the next chapter) are interesting and deserving of attention. But the reader should be mindful of its provisional nature.

The hypothesis I will defend is one I call the "Coordination of Responses Hypothesis," or CoRH. In summary, the CoRH states that the function of a moral judgment is to bring about or help maintain general conformity, within a group, to some pattern of behavior. Its Normal way of doing so is through coordinated sanctions, mediated by characteristic emotional states (like anger and guilt). I will present this view in more detail in sections 2 and 3.

When discussing the functions of moral judgments, it is important to distinguish their *invariant* from their *adapted* proper function. As the reader may recall from p. 46, representational systems typically have invariant functions: stable outcomes that they are supposed to produce across variations in external circumstances. The waggle dance, for instance, has the invariant function to bring nest mates to nectar. But each member of a family of representations also has the function to adapt the consumer to its own

content, a function that is specific to each (syntactically individuated) representation-type in the family. Each waggle dance has the function to make nest mates fly to a *particular* location, which is a function of the dance's shape. The invariant function of a representation can come apart from its adapted function, but under Normal conditions they will coincide.

Insofar as moral judgments are produced by a single psychological faculty with a single invariant function, they all have the same invariant function: that of ensuring that the faculty's function is fulfilled. But different moral judgments also have different adapted functions. For instance, the judgment conventionally expressed by

φ-ing is morally wrong

has the function to produce or help maintain a specific behavior, namely—if the view I will be defending is correct—the behavior that consists in abstention from φ-ing, punishing φ-ers, etc.

Identifying the invariant evolutionary function of moral judgments would be of central importance for determining their Normal conditions, hence their descriptive content, since Normal conditions include conditions that must be in place if a representation is to be able to perform its invariant function *by* performing its adapted function. Consider again the waggle dance: the function of a particular dance is to lead nest-mates to a specific location *x* (adapted), *and* to lead them to nectar (invariant). Under Normal conditions, these two functions will permit of joint performance, since there is nectar at *x*. And as we know by now, that is also the descriptive content of the dance: *there is nectar at* x.

As we shall see in section 5.4, one common proposal is that the moral faculty has evolved to produce and maintain cooperation in human groups. If this is correct, then that should also be the invariant function of the judgments it produces for that purpose. Hence, by analogy with the bee-dance example above, the descriptive content of a moral judgment of the form "φ-ing is wrong" ought to be something like *φ-ing is bad for cooperation*. We will have reason to get into much more detailed discussion of how to assign descriptive content to moral judgments in the next chapter. In preparation for that discussion, I consider the cooperation thesis and other candidate theories about the invariant function of moral judgment in section 5.4.

The CoRH as such is not, however, a claim about the moral faculty's invariant function, but about the adapted function of moral judgments and how they Normally perform that function (namely, by coordinating responses). Since the coordination of responses can theoretically be leveraged to steer behavior in all manners of directions, the CoRH as such is fairly neutral with respect to different proposals about the invariant function of morality. By partitioning my claims along these lines, I hope to be able to separate the "how" of morality from the potentially more controversial "why," and to

offer a toolbox with the help of which the latter question can be explored within a teleosemantic framework. In particular, by remaining neutral on the "why," the CoRH remains neutral on whether morality is *pluralistic*, i.e., on whether there is more than one way the moral faculty can secure its own persistence. The possible pluralism of morality bears on more traditional ethical questions: whether morality is objective and absolute or relative and contingent; whether it is a social construction or an endowment of ethical intuition; etc. These questions, transposed into the teleosemantic register, will be the topic of chapter 6, with preparatory remarks in section 5.4.

The rest of the chapter is devoted to presenting, elaborating, and defending the CoRH. In section 5.1 I clarify that the CoRH is a theory about, specifically, *paradigmatic* moral judgments. I motivate the decision to focus on the paradigm by reference to the methodological difficulties pertaining to the analysis of moral judgment deriving from the inherent vagueness of the notion. In section 5.2 I present the CoRH, explain what I mean by "coordination" and "responses," and apply the account to specific forms of moral judgment. In section 5.3 I further develop the CoRH by discussing the two main kinds of responses, overt sanctions and emotional responses, and discuss their interrelation. Section 5.4, as already mentioned, is devoted to a discussion of the moral faculty's invariant function.

5.1. Characterizing Moral Judgment

Characterizing the nature of moral judgment is not a meta-ethically neutral endeavor. A cognitivist will hold that moral judgments are defined by how they descriptively represent the world. A non-cognitivist will instead insist that a moral judgment is a type of non-cognitive attitude defined by its particular motivational profile. After the last chapter, it will come as no surprise that I will defend a view that combines these two perspectives: moral judgments are both motivational and descriptive. But that, in and of itself, is saying very little. How do they motivate, what do they describe, and why?

First, let us delimit the area of inquiry. I will be concerned with "thin" moral judgments, the kind that (according to philosophical received wisdom) are conventionally expressed using "thin" moral predicates like "right" and "wrong."[95] Specifically, I will be focusing on the sorts of judgments conventionally conveyed by sentences of the form:

φ is *M*

[95] The terminology of "thin" and "thick" is introduced in (Williams 2011, 143–45). For review, see (Roberts 2013).

Where φ stands for a predicate or description referring to an action or an action type, and *M* stands for what I will call a "deontic predicate," a category that I take to include such predicates as "(morally) right," "wrong," "obligatory," "permitted," "forbidden,"[96] etc. Since the semantics of natural languages is not my main focus here, I will not attempt to tease out the exact inferential relationships between these terms or the exact circumstances under which they do or do not express specifically moral judgments. Nor will I concern myself with investigating their relationship to other conceivable linguistic tools for expressing moral judgments, such as modal auxiliaries like "ought" and evaluative predicates like "good." The discussion to follow is sufficiently abstract and exploratory that, I hope, these details will not become relevant.

A problem inherent to the task of analyzing moral judgment is that the category of "moral judgment" is unlikely to be sharply delimited. Moral judgments are likely to form a clump around certain paradigmatic instances, slowly tapering off into more and more marginal cases. This means that the apparent failure of a given analysis can always be blamed on the inherent vagueness of the notion. What fails as an account of moral judgments in a demanding sense may yet succeed as an analysis of moral judgments in a more attenuated sense, and vice versa. Richard Joyce, in *The Evolution of Morality*, offers a list of no less than seven features of moral judgments, including such things as motivational role, categoricity, inescapability, implications of desert and justice, and so on (Joyce 2007, 70). Do all these features have to be present for something to qualify as a moral judgment, or can we dispense with one or two of them and still have something that qualifies? The question has no obvious answer, and that in itself presents a methodological problem for any attempt to offer a theory of moral judgment.

In light of these realities, what seems to me the most promising approach is the one suggested by Gunnar Björnsson and Tristram McPherson (2014). They offer an account of *paradigmatic* moral judgments, characterized by five features that jointly constitute a "homeostatic property cluster" in the style of Richard Boyd (1988). A homeostatic property cluster is a set of properties that tend to be co-instantiated due to "homeostatic" mechanisms, i.e., mechanisms that increase the probability of the presence of one property contingent on the presence of each of the others.

Focusing on the paradigm, rather than attempting to strictly delineate the entire category, allows us to bypass fraught definitional questions. Moreover, conceiving of the paradigm as a homeostatic property cluster allows us to investigate the homeostatic mechanisms involved to gain a more system-

[96] These predicates have both moral and non-moral uses. I will ignore this complication and use "right," "wrong" etc. without the qualifier "moral," intending throughout the specifically moral uses of the term. I briefly discuss non-moral uses of "right" and "wrong" in section 6.5.

atic understanding of when and why they fail, giving rise to various non-paradigmatic cases. In this way it may be possible, not only to account for conflicting intuitions about the nature of moral judgment by means of an overarching theory, but also to show how these intuitions relate to one another and what makes them compelling.

Björnsson and McPherson offer no detailed account of the specific homeostatic mechanisms responsible for clustering their features, but the suggestion they give is one that should appeal to the teleosemanticist: *function*. If moral judgments have a function, the features characteristic of paradigmatic moral judgments could all be expressions of a structure specifically adapted to performing this function, one that has been shaped and is maintained through a process of selection.

Following this idea, I will offer a functional analysis of moral judgment, a view I have dubbed the *Coordination of Responses Hypothesis* (CoRH). This view consists of two sub-claims, one about the function of moral judgment, the other about the mechanism whereby that function is Normally performed. Neither claim is particularly novel. They reflect common views on the nature of moral judgment, though reframed within the teleosemantic framework. Most immediately, I owe the view to Neil Sinclair (2012), who is one of my few predecessors in teleosemantic meta-ethics.

The first claim of the CoRH is that the *function* of a moral judgment of the form "φ is *M*" (specifically its adapted function; see introduction) is to produce or maintain a state of affairs characterized by conduct defined in relation to the action φ, as specified by the deontic predicate M, that holds *universally* or at least broadly within some group of which the judger is a member. For instance, the judgment that killing is wrong has the function to produce a state of affairs where nobody kills. This reflects the idea, defended by R. M. Hare (1952, 188–92), that moral judgments and statements can be analyzed as a kind of generic imperatives, directed not at any particular individual but to people in general.

The second claim concerns the mechanism whereby moral judgments Normally perform this function. In part, it does so by motivating the subject to act in accordance with her judgment. But it also motivates her to respond in certain ways to her own and others' actions depending on whether they accord with the judgment. These involve emotional responses (such as guilt and anger, pride, esteem and gratitude) but also overt sanctioning responses (rewards and punishments) partly motivated by these emotions. Finally, it motivates her to make others *share* her judgment, which produces a *coordination* of responses that further contributes to making its adapted function—to produce broad conformity to a pattern of action with the group—be performed.

In the next section, I will present the CoRH in more detail, trace its origins with my immediate predecessors, and show how it applies to specific types of moral judgments.

5.2. The Coordination of Responses Hypothesis

As I mentioned above, I take the CoRH most immediately from Neil Sinclair (2012). Sinclair, in turn, takes it from Alan Gibbard (1990), so I will begin by discussing Gibbard's view.

Gibbard's main *analysandum* in his seminal work *Wise Choices, Apt Feelings* is not specifically *moral* judgment, but judgments about what is *rational* or what it *makes sense* to do. According to Gibbard, such judgments have as their function[97] to enable us to solve *coordination problems* in social groups (Gibbard 1990, 61). Moral judgments are a special case of normative judgments, whose function is to coordinate certain emotional responses.

A *coordination problem* is a decision situation involving two or more agents, who have to coordinate their actions in order to attain a mutually beneficial outcome. By definition, the agents all prefer successful coordination to failure of coordination, but it may be more or less a matter of indifference exactly how the coordination is attained. For instance, if you and I want to meet up in town, it may be a matter of relative indifference exactly where we meet, as long as we both go to the same place. Similarly, if a population of motorists wants to avoid traffic accidents, it is largely indifferent whether they drive on the left or the right side of the road, as long as they all pick the same side. Solving a coordination problem means finding a way of ensuring that agents coordinate on a single solution (that you and I go to the same meeting-place; that drivers drive on the same side of the road).

Indifference is a matter of degree. What characterizes all coordination problems is that each of the participants prefers coordination to failure of coordination, but within those constraints, some coordination solutions may be preferable to others. If the church is closer than the river to both our houses, we both prefer to meet at the church. Sometimes, different agents favor different coordination solutions. If you live closer to the river and I live closer to the church, then you might prefer to meet by the river and I by the church. In those cases, we are dealing with a *bargaining* problem (Schelling 1960), which brings the agents into potential conflict: although all agents favor coordination, the fact that they each favor a different way of coordinating may prevent a solution from being found.

Alan Gibbard's proposal, in *Wise Choices, Apt Feelings*, is that normative judgments (which he identifies with judgments about what it makes sense to do) have the function to solve coordination problems by shaping mutual expectations in a group and motivating people to act in predictable ways (1990, 61). Judgments of moral wrongness, on Gibbard's view, are (roughly) judgments about what actions it makes sense to feel guilty about and to be angry at (Gibbard 1990, 40–45). Their function, then, is to *coordinate* feel-

[97] Gibbard subscribes to a view of functions that is essentially the etiological view defended in this thesis.

ings of guilt and anger-responses. Specifically, they are supposed to ensure that I am angry at your actions just in those cases when others are, and just in those cases when you feel guilty about them.[98]

At first blush, this theory might seem unnecessarily indirect. A more straightforward theory would be that the function of moral wrongness-judgments is simply to ensure that people avoid certain actions. Neil Sinclair's view, which builds on Gibbard's, diverges from it by directly ascribing to wrongness-judgments the function to dissuade the target behavior:

> The function of a judgement of the form 'Φ is wrong' for example, is to produce a stable pattern of co-ordination where no one Φ's, and relatedly, where people disapprove of Φing and disapprove of those who fail to disapprove of Φing. (Sinclair 2012, 653)

But note that on this view, moral judgment *also* retains the function of promoting the coordination of responses to the target actions, here described generically as "disapproval" rather than in terms of specific emotional responses, like guilt and anger.

What should we take to be the relation between these two elements of the function of moral judgments, the direct promotion of action and the coordination of responses? If moral judgments were exhaustively defined by the function of influencing action, they would be difficult to distinguish from, e.g., judgments of mere prudential advisability. Moral judgments, I posit, are characterized not only by the ultimate ends that it is their (adapted) function to produce, but also by the specific *mechanisms* whereby they Normally produce them. These mechanisms consist precisely in the coordination of people's responses to the target behavior.

The social responses by means of which moral judgments Normally perform their function, I further posit, include not only the emotional responses that Gibbard emphasizes (guilt and anger), but also *overt sanctions*, i.e., substantial rewards and punishments. Plausibly, the function of anger is to motivate behaviors that have a punishing effect on the target of the anger. Similarly, positive moral emotions like gratitude have the function to motivate positive sanctioning behavior (Fessler and Haley 2003). On my view, both emotional responses and overt sanctions have a role in the mechanism whereby moral judgments Normally perform their function. We can reserve Sinclair's terms "approval" and "disapproval" as generic terms for social responses to actions, positive and negative respectively, that cover both emotional reactions and overt sanctions. In section 5.3 I will discuss the relationship between these two types of responses.

"Sanction," as used here, can mean both rewards and punishments, and rewards and punishments can take either a positive form (actively interven-

[98] Others who have defended similar views include Blackburn (1998b, chaps. 2–3), Mameli (2013).

ing to benefit or harm an agent) or a negative form (withholding harm or benefit from them). "Punishment" need not involve overt punitive acts, like violence or economic sanctions. A common form of punishment is ostracism, which divests the transgressor of gainful opportunities for cooperation (Baumard, André, and Sperber 2013). As Sinclair also stresses (2012, 651), a mere public display of the *willingness* to sanction can itself serve a punitive function. Such a display contains the threat of more substantial punishment, which is often sufficient to discourage action (Fessler and Haley 2003).[99] Overt displays of disapproval unaccompanied by material sanctions have a psychologically punitive effect by themselves (Masclet et al. 2003). Matteo Mameli suggests, plausibly, that "this intrinsic aversion to disapproval has evolved as a result of—and in order to avoid—the negative effects, reputational or otherwise, of disapproval" (2013, 911; cf. Blackburn 1998b, 205).

What is sanctioned, and why? In principle, any type of behavior can be promoted through sanctions (Hirshleifer and Rasmusen 1989), but it is typically assumed in the literature that the mechanism of collective sanctions has developed specifically to encourage cooperative behavior in situations where the absence of a sanctioning mechanism would incentivize defection. These situations have the structure of a prisoner's dilemma, where each agent stands to gain by unilaterally choosing a noncooperative strategy, resulting in an outcome that is worse for everyone. By punishing defection and rewarding cooperation, thus raising the cost of unilateral defection, the prisoner's dilemma can be transformed into a coordination problem with cooperation as one equilibrium (Bicchieri 2006). For more on this, see section 5.4.

As we have seen, Sinclair proposes that a judgment of the form "Φ is wrong" has the function to produce a state of affairs "where no one Φ's, and relatedly, where people disapprove of Φing and disapprove of those who fail to disapprove of Φing" (2012, 653). If we take "disapproval" to be, either a form of punishment (as in overt displays of disapproval) or a mental state whose function is to motivate punishments, we notice that according to Sinclair, the function of moral judgments involve not only first-order sanctions (punishments of actions judged wrong) but also second-order sanctions (punishments of failures to punish actions judged wrong). This type of higher-order punishing behavior is sometimes called precisely "moralistic punishing" in the literature (e.g. R. Boyd and Richerson 1992). Higher-order sanctions are often in fact part of the social mechanisms that uphold moral norms (Martin et al. 2017). If I fail to be appropriately angry at what others perceive as a moral transgression, I will myself become suspect and a potential

[99] This last point leads us back to the question why sanctions would need to be coordinated in the first place. We may suppose that sanctions fulfill their function of influencing behavior more reliably and effectively if it is predictable and consistent. If there is to be a credible threat of punishment then clearly punishment should be fairly consistent across occasions of interaction. Moreover, by pooling their sanctioning efforts, individual agents lower the cost of sanctioning for themselves.

object of rebuke or ostracism. Conversely, if I join in a sanctioning endeavor, my comrades' esteem for me will grow.

There are good reasons why a mechanism for shaping the behavior of people that relies on coordinated sanctions must also include higher-order sanctions. The coordination of sanctions presents its own collective action problem. Sanctions can be costly, in terms of missed opportunities if not actual resources, and there is accordingly a temptation to defect: to benefit from the good effects of punishment without shouldering any of the costs. In this way, coordinated sanctions present a prisoner's dilemma type problem, which can be solved by instituting higher-order sanctions (R. Boyd and Richerson 1992; Oliver 1980). Since the second order of sanctioning similarly presents a collective action problem, there may need to be a third order of sanctioning, and so on, for what could in principle be an arbitrary number of higher orders.

Is this a problem for the account? Do we have reason to suspect that if the n^{th} order of sanctions could always in principle require an $n+1^{th}$ order of sanctions, sanctioning would never get off the ground? I'm not certain, but I don't think there is any reason to worry. As long as agents are aware, however dimly, of the *possibility* that $n+1^{th}$ order-sanctions could be put in place, this may be sufficient to motivate them to sanction on the n^{th} level.

This, then, in outline, is the hypothesis, the CoRH, that I will be defending: The function of a moral judgment is to steer behavior in accordance with the judgment, and its Normal mechanism of accomplishing this function is by coordinating responses (and responses to responses, etc.) of approval and disapproval, which include both characteristic emotional responses and overt sanctions. Let us look at how this rough view can be refined for a number of more specific judgment types.

Moral wrongness judgments: Following Sinclair, I will assume that the function of a judgment of the form "φ is (morally) wrong*"* is to discourage φ-ing, i.e., to make it the case that people do not φ. It Normally does so by coordinating responses (including higher-order responses). Sinclair, in the above quote, mentions specifically *disapproval* responses, but on our generic definition of approval and disapproval, moral wrongness judgments can just as well be said to coordinate approval *away* from φ-ing, so that people approve of abstentions from φ-ing. In Normal cases, these approvals will primarily take the form of negative sanctions: withholdings of punishments, rather than positive rewards.

Moral requirement judgments: The function of a judgment of the form "φ is (morally) obligatory" is to make it the case that people φ. Its Normal way of doing so is by coordinating responses (including higher-order responses), specifically by making it the case that people approve of φ-ing and disapprove of failures to φ.

Judgments of supererogatoriness: A supererogatory action is an action that goes "beyond the call of duty": one that, though not required, is morally commendable. The function of a judgment of the form "φ is supererogatory," I posit, is to make it the case that people φ. It differs from moral requirement judgments in that its Normal way of bringing about this end involves *only* making it the case that people approve of φ-ing, *not* that they disapprove of failures to φ.

Moral permissibility judgments: The function of a judgment of the form "φ is (morally) permitted" is a less straightforward case. Its function, intuitively, is neither to encourage nor to discourage φ-ing; rather, they serve as the contradictories to moral wrongness judgments. Plausibly, the function of permissibility judgments directly concerns the management of responses. The judgment that φ-ing is permitted is to produce a state of affairs where φ-ing is *not* punished. In this way, they serve Normally to cancel the effects of moral wrongness judgments.

5.2.1. Moral Judgment about Token Actions

I would now like to discuss a problem for the CoRH as I have defined it. The problem is this: the view suggests that the function of a moral judgment is to encourage or discourage an action *type*, i.e., to make it the case that people in general perform or avoid actions of that type. But moral judgments can be both type-directed and token-directed. The predicate or description that substitutes for φ in the schema "φ is *M*" can refer to both action types and token actions. *Killing is wrong* is an example of the former; *what you're doing is wrong* is an example of the latter. However, it clearly makes no sense to ascribe to (say) a token-directed wrongness judgment the function of producing a state of affairs consisting in the *general* avoidance of the particular token it is about. A token action is performed or avoided by one person at one time.

There are two options for how to extend the CoRH so that it covers token-directed judgments in addition to type-directed ones. According to the first option, the function of token-directed judgments is the direct discouragement of the token action in question, plus the coordination of responses directed at the action (if it is nevertheless performed). This solution, however, seems hard to square with the existence of token-directed judgments *in the past tense*, some of which concern actions far in the past, performed by long-dead agents. If their function were to discourage the target action itself, they would be constitutively incapable of performing their function, and we would therefore have to conclude that they are defective and abNormal. But they seem perfectly sensible.

The other option is to maintain that a token-directed wrongness judgment, no less than a type-directed one, has the function to discourage an action

kind to which the token action belongs. Indirectly, by discouraging this kind, the judgment also discourages the token action itself *qua* member of the kind, at least insofar as this is possible (i.e. the action is not already performed).[100] I will follow this second strategy. More specifically, I will maintain that the function of token-directed wrongness judgments is to bring about or help maintain a state of affairs where the types of actions not performed, disapproved of, etc. include the target token (and *mutatis mutandis* for the other kinds of moral judgment). It should be noted that since any token action belongs to a large number of different types, there are many different states of affairs meeting the aforementioned desideratum, thus counting as teleological success for the judgment.

5.2.2. Summary of the CoRH

I will now sum up the CoRH as I have described it so far:

> (COORDINATION OF RESPONSES HYPOTHESIS) The adapted function of a moral judgment conventionally expressed by a sentence of the form "φ is *M"* is to bring about or help maintain a state of affairs characterized by conduct defined in relation to φ that holds universally, or at least broadly, within some group of which the judger is a member, and where the relation between the conduct and φ (compliance, abstension) is determined by the deontic predicate *M*. Its Normal way of performing this function is by producing coordination of approval and disapproval, including sanctions as well as emotional responses like guilt and anger.

This formulation requires some commentary. I say, somewhat vaguely, that the function of a moral judgment is to produce or maintain conduct that holds "universally, or at least broadly, within some group of which the judger is a member." First of all, I leave it deliberately open *which* group it is that a moral judgment has to produce universal conduct in to count as teleologically successful. I will return to this issue the next chapter. For the moment, let us think of it as the judger's group, tribe, society, etc.

Second, the formulation should not be taken to mean that moral precepts cannot be conditionalized on possession of some attribute or membership of some subgroup. The moral precept that it is wrong for police to abuse their powers, for instance, can be considered universally obeyed if no-one satisfies the conjunction *is a police and abuses their powers*, i.e. if no police abuses their powers.

Third, the formulation "universally, or at least broadly" is meant to indicate that the success of a moral judgment can come in *degrees*. There is rea-

[100] This solution is roughly analogous to the analysis Hare (1952, 156–58) offers for past-tense ought-judgments.

son to suspect that few moral precepts have ever been *universally* complied with in any groups of significant size for any length of time. I assume that the moral faculty has been able to pay its evolutionary dues by producing, often enough, sufficiently broad conformity.

If this is granted, the next natural question is: how is the coordination of approval and disapproval actually accomplished? It seems that two factors must be involved. First, the moral judgment must dispose the judger *herself* to harbor the relevant emotions and motivate her to dispense the relevant sanctions. Second, it must motivate her to induce others to harbor these emotions and dispense these sanctions as well—to actually bring about coordination—and others must be receptive to her attempts. In section 5.4, I will discuss how coordination on the social level can be accomplished. Before that, in the next section, I will discuss the emotions and sanctioning dispositions harbored by the individual moral judge, with a particular eye to how these two forms of response relate to each other, how they reinforce each other, and how they contribute to the functional unity of paradigmatic moral judgments.

5.3. Emotions and Sanctions

There is certainly no dearth of literature on the topic of what role emotions play in moral judgment and to what extent a judgment can qualify as "moral" without a characteristic emotional or motivational component. Most people would probably deny that somebody who acted morally, or punished the moral transgressions of others, *purely* out of the promise/threat of (higher-order) sanctions would qualify as acting out of genuine moral conviction. Whatever psychological states motivate his conduct either do not qualify as moral judgments, or if they do, they are at least not of a paradigmatic kind. And the moral judgments of other people working through him (by influencing his actions) do so in a deficient way.[101]

For our purposes, the definitional question is, again, less interesting than trying to understand how and to what degree the various features commonly associated with moral judgment support each other in a functional unity. In what ways, if any, do the typical emotional correlates of moral judgment (guilt, anger, pride, gratitude) help those judgments fulfill their functions, thereby contributing to an explanation of why people go on making them? If we can show that they do contribute, it would lend plausibility to the idea that they play a role in the Normal operation of paradigmatic moral judg-

[101] Indeed, an entire society may be stably coordinated around punishing a certain type of behavior even though none, or very few, of its members actually mind the behavior very much. It is sufficient, for this state of affairs to arise, that they are ignorant of each other's lack of personal commitment to the sanctioning regime and fear the consequences of unilateral defection. This type of "pluralistic ignorance" is discussed in (Bicchieri 2006, chap. 5).

ment. Correspondingly, a mental episode that lacked these characteristic emotional correlates (like the would-be-moral judgments of a sociopath) would be shown to be deficient for the purpose of fulfilling the function of moral judgment. It could yet have a different function, of course.[102] It could also be a non-optimal means whereby paradigmatic moral judgments sometimes perform their function.[103]

Among typical emotional correlates of moral judgment, we can distinguish between *self-directed* emotions (guilt, shame, pride) and *other*-directed emotions (anger, gratitude). The former are self-directed in that they derive from an evaluation of one's own conduct, the latter are other-directed in that they derive from an evaluation of others' conduct. These emotions are all

[102] The sociopath, or amoralist, is a figure dear to the meta-ethicist's heart. He usually appears in debates about motivational internalism, the view that moral judgment *essentially* involves a characteristic motivational component. Those who reject motivational internalism are forced to accept that the sociopath can make moral judgments, while those who defend it must insist that he cannot.

It is assumed in this debate that the sociopath has the ability to keep track of which actions are generally considered "right" and "wrong" and so at least to use these terms in what Hare described as an "ossified" or "inverted commas" sense (Hare 1952, 164). The question is whether this inverted commas use counts as making moral judgments. Now, if I am right that the sociopath's judgments are ill-suited to serving the function of paradigm moral judgments, we may ask what function they could have instead. One obvious hypothesis is that it is useful for the sociopath to be able to keep track of what people consider right and wrong, in order to be able to avoid their sanctions and hide his own sociopathy (cf. Mumm 2015).

If this is correct, it means that the sociopath's "moral" judgment are *parasitic* upon the existence of paradigmatic moral judgment in the community, in the sense that without the latter, there would be no function for the former to fulfill (cf. n. 43, p. 62). This, I take it, is a good way to substantiate the claim that they are less paradigmatic (see also Blackburn 1998b, 61).

Likewise, in the "pluralistic ignorance" scenario discussed in the previous footnote, it is dubious whether we are dealing with *moral* attitudes at all. The people in the pluralistic ignorance scenario behave with respect to the unpopular norm quite like the amoralist: while capable of keeping track of what is conventionally considered "right" and "wrong," they lack the characteristic emotional correlates of moral judgment. Of course, while both the amoralist and the pluralistic ignoramuses are attempting to track the *paradigmatic* moral judgments of others, what is curious about the pluralistic ignorance scenario is that it is characterized by a massive *failure* of tracking: everybody thinks that everybody else has sincere moral convictions and that they alone are insincerely going through the motions.

A state of pluralistic ignorance has to originate somehow, and perhaps sometimes it originates in the genuine moral judgments of some person or people. These original judgments have then fulfilled their function, but the resulting state will be inherently unstable and liable to break down through "information cascades" wherein people discover their mutual lack of commitment to the norm (Bicchieri 2006, 196 ff.). Moreover, the mechanism whereby they have succeeded would appear to be parasitic on the more typical case where people actually come to share genuine moral judgments, since pluralistic ignorance arises only because people believe each other to have genuine moral convictions. Hence, it seems unlikely that the pluralistic ignorance-mechanism on its own could have contributed to the persistence of moral judgment. It is an abNormal case.

[103] Another issue, which I won't treat here, concerns the metaphysical relationship between moral judgments and their emotional correlates, i.e., whether the emotions are Normally *caused by* the judgments, or somehow *part of* the judgments, or perhaps even the *causes of* the judgments. Discussion of the etiological function of moral judgment can proceed in abstraction from this issue.

recognizable as mechanisms for motivating the behaviors Normally associated with a moral judgment. The self-directed emotions, to which we may also refer collectively as *moral conscience*, directly motivate conduct in conformity with one's moral judgment. The other-directed emotions motivate sanctioning behavior: punishment in the case of anger, reward in the case of gratitude.

There are also good reasons to believe that our capacity to feel moral emotions have evolved both to sustain sanctioning regimes and to adapt us to them and allow us to function better within them, and that without these emotions, sanctioning regimes either could not arise or would function less efficiently. Richard Joyce, following Robert Frank (1988), suggests that moral conscience is an evolved adaptation whose function is to secure pro-social behavioral dispositions against shortsightedness and weakness of will, which might otherwise lead people to adopt (ultimately maladaptive) selfish strategies whenever the benefits of pro-sociality or the threat of sanctions are not sufficiently salient (Joyce 2007, chap. 4). Since moral considerations are often precisely such that they require us to offset short-term self-interest against more long-term benefits, and the animal brain notoriously gives a steep discount to future gains over present, it makes sense to think of moral conscience as a special motivational mechanism designed to counteract this tendency toward short-term thinking and thus enable the subject to reap the long-term benefits of being in harmony with her fellows.

Joyce is certainly right that moralized behaviors are often behaviors that involve a tradeoff between short-term and long-term gain. If nothing else, the moralization of the behavior by itself imposes such tradeoffs, since the whole point of the mechanism of coordinated sanctions is to impart relative (social) benefits to those who conform to moral norms, and these benefits are often delayed in time relative to whatever immediate payoff one stands to gain from violating them. A motivational mechanism that "binds" the subject to a certain course of action and "eliminates certain practical possibilities from the space of deliberative reasoning" (Joyce 2007, 111) would thus be a useful complement to external sanctions in enabling a moral judgment to perform the function assigned to it by the CoRH.[104]

Guilt is primarily directed toward past actions, but can also motivate one to behave more morally in the future, so as to avoid further guilt and/or to

[104] It is worth mentioning that in saying this much, I have not committed myself to Joyce's further view that moral conscience is a biological adaptation. My view is compatible with the possibility that the cultivation of moral emotions is at least in part a learned behavior.

In regard to this, it is interesting to note that the moral conceptions of many cultures centrally involve the employment of conscious techniques, often conceptualized in religious terms, for strengthening moral resolve (Zigon 2008). These techniques are products of culture rather than evolution. But of course, in order to be motivated to learn and employ these techniques a person probably needs something like an innate conscience in the first place. Conscience may thus be a matter of innate motivational dispositions and culturally acquired techniques working in concert.

reestablish one's credentials as a reliable group member, which may have been tarnished by the past transgression.

A long tradition, including Hume and Adam Smith (cf. Blackburn 1998b, 200-04), has viewed the self-directed moral emotions as a sort of internal reflection of the (expected) other-directed emotions that others direct toward oneself. If I believe that others will feel rage, I feel guilt; if I believe that others will feel esteem, I feel pride; and so on. Since guilt is itself psychologically punishing and aversive and pride is itself psychologically rewarding, these emotions also serve as internal reflections of overt socialized sanctions.[105] In a world where the alternative to the internal punishment of guilt is punishment of a concrete, external kind, the capacity to feel guilt will be a good investment. Though guilt feels bad, it does not impose substantial material costs on the subject (cf. Boehm 2008, 2014; Mameli 2013). Pride, too, as an internal reflection of external rewards, will tend to further encourage behavior that yields those rewards.

Gibbard, in addition, stresses the way guilt makes people seek forgiveness for their crime, in effect motivating them to take steps to be reintegrated in the moral community (Gibbard 1990, 67–68). This allows others to learn that the sanctions have had their desired effect of influencing behavior, and so spares them any future expenditure of sanctioning resources, which further strengthens the efficiency of the sanctioning regime. (A parallel account, *mutatis mutandis*, could be given for pride).

In summary, it seems plausible that, though moral conscience can certainly motivate a person to act morally entirely independently of any considerations of future sanctions, the *function* of conscience is to adapt the subject to the sanctioning regime of the surrounding society. In so doing, it benefits the individual as well as increasing the efficacy of the sanctioning regime itself.

As for the other-directed moral emotions—anger, gratitude, etc.—Robert Frank's (1988) account explains their role as well. Since sanctions can be costly and the sanctioned behavior has already occurred at the time of sanctioning, there is a temptation to abstain from sanctions. But if people regularly succumbed to that temptation, the efficacy of the sanctioning regime would be compromised: sanctions would present less of a credible threat/promise, and would be less efficient in shaping behavior. For the subject, then, other-directed moral emotions again serve as commitment devices. In this way, they benefit the subject in the long term by encouraging others to act in preferred ways, as well as helping the subject avoid higher-order sanctions, and it also strengthens the efficacy of the sanctioning regime.

I conclude that the emotional responses play an important part in the mechanism whereby moral judgments Normally perform their function,

[105] Nietzsche (2000, 520–21), somewhat curiously, takes guilt to be an internalization, not of *others'* punishments, but of one's own desire to punish others—a desire that, supposedly, finds no outlet in civilized society and so must be turned against the self.

alongside overt sanctions. They serve to motivate these sanctions themselves, as well as motivating conforming behavior in anticipation of future sanctions. Without them, the sanctioning regime would function less efficiently, if it would function at all.

In this section we have discussed the relation between emotional dispositions and overt behavior on the part of the individual subject of a moral judgment. But what is all this coordination for? What ultimate end does it serve? To ask this question is to ask for the *invariant* function of moral judgment.

I have claimed that the function of a moral judgment is to influence behavior, and that it is distinguished from other types of directive attitudes by the special mechanism (coordinated sanctions, moral emotions) whereby it Normally performs this function. This view, which I have christened the CoRH, is a view about the *adapted* functions of specific moral judgments. I have not yet said anything about the *ultimate* purpose of having a morality. Is there a single type of outcome towards which all moral judgments produced by the moral faculty are directed, like token waggle dances are all directed at the outcome of bringing nectar back to the hive? In other words, does the moral faculty have an invariant function, and if so, what is it? It stands to reason that moral judgments, and the faculty that produces them (however this faculty is to be understood, as an innate disposition or a learned behavior), have managed to persist not only because they have managed to influence the behavior of the group each in their specific direction, but because they have jointly managed to influence the group towards *certain* types of outcomes. As I stressed in the chapter introduction, identifying this invariant function is important for identifying the descriptive content of moral judgments. Hence, we now turn to this topic.

5.4. The Invariant Function of the Moral Faculty

In the literature on the evolution of morality, we find many proposals as to what the invariant function of the moral faculty could be. A recurring idea, and an eminently plausible one, is that morality evolved due to its ability to promote cooperation and harmonious social coexistence (Kitcher 2007; Prinz 2009; Smyth 2017). Neil Sinclair also defends a version of this view:

> What, then, is the evolutionary function of moral judgements? I propose it is interpersonal co-ordination. Roughly, moral judgements are the products of a mechanism that allows groups of interacting individuals to co-ordinate their actions and emotions for mutual benefit. The function of the moral habit is therefore to produce mutually beneficial co-operative patterns of action and emotion. (Sinclair 2012, 649)

According to Sinclair, the "moral habit"—his name for the moral judgment faculty—is supposed to track cooperative behavior options and motivate the agent to choose those options. The invariant function of moral judgments, which they derive from the function of the moral faculty, is thus to produce cooperation.

Now, cooperation is obviously important to morality in several ways. Many paradigmatic objects of moral approval and condemnation are instances of, respectively, cooperative and uncooperative behaviors. Moreover, the mechanism of coordinated sanctions is itself an exercise in cooperation. But that is not yet to say that the (exclusive) function of the moral faculty is to promote cooperation. We may call this latter, stronger thesis the "cooperation view."

The general idea behind the cooperation view is familiar. It is evident that human life involves constant and pervasive cooperation and that cooperation is the means whereby we attain much of what we need and value in life. But cooperation is vulnerable to cheating. Many situations will produce temptations for agents to partake of the benefits of cooperation while shouldering none of its costs. This can happen, for instance, if the benefits yielded by a cooperative endeavor are difficult to control and distribute discriminately, or if the occasion for the agent's reciprocation is delayed relative to his enjoyment of the benefits. Many of these situations can be modelled, at least if we abstract from the agents' pro-social motivations, as prisoners' dilemmas, where cooperation yields the highest total benefit but each agent stands to gain more for himself by cheating. But if agents succumb to the temptation to cheat sufficiently often, cooperation breaks down. Even if I am personally prepared to cooperate, I won't do so if I have reason to fear that others will cheat me. Cooperation requires commitment, and it requires trust that others are similarly committed.

It would therefore be good if there were some kind of constraint on our and others' actions that could minimize the risk that cheating will disrupt our mutually gainful cooperative endeavors. According to one common perspective, morality constitutes this constraint.

The contractualist (or contractarian) tradition, going back to Hobbes, has tried to understand the constraint, hence morality, in terms of a contract or agreement, entered into by agents mindful of the value such a constraint would have. David Gauthier expresses this idea particularly starkly:

> Given the ubiquity of such situations [of mutually gainful cooperation threatened by the temptation to cheat], each person can see the benefit, to herself, of participating with her fellows in practices requiring each to refrain from the direct endeavor to maximize her own utility, when such mutual restraint is mutually advantageous […] We may represent such a practice as capable of gaining unanimous agreement among rational persons who were choosing the terms on which they would interact with each other. And this agreement is the basis of morality. (Gauthier 2013, 575)

As Gauthier knows, however, an agreement is only as good as people's dispositions to actually abide by it. An actual agreement is as vulnerable to cheating as the cooperative endeavors it was meant to prop up. What is needed, before any actual agreement can be efficient, is a *psychological* constraint, something that people cannot—or not so easily—ignore. That psychological constraint could consist, for instance, in an emotional disposition to feel guilt at cheating and to punish cheaters so as to diminish the temptation that cheating presents. Thus emerges the view of morality as a psychological package designed to underpin cooperative endeavors and protect them against the temptation to cheat: the cooperation view.

There are a few reasons to doubt the cooperation view, however. One relates to the fact that, as any ethicist will tell you, not every act of moral approval or condemnation is concerned with cooperation or lack thereof. We can moralize about nature, wild animals, matters of personal hygiene or religious reverence, etc. Of course, it is possible that these are abNormal objects of moral judgment. The moral faculty could have evolved for one purpose and later come to be exploited for other ends. But there are alternatives to the cooperation view built on the assumption that some of these non-cooperative objects of moral judgment derive directly from its evolutionary function.

One such view is the approach called *Moral Foundations Theory* (MFT). MFT posits the existence of as much as five or six different "moral foundations" (Graham et al. 2013). Inspired by the sentimentalist moral psychology of Jonathan Haidt and Craig Joseph (Haidt and Joseph 2004; Haidt 1995), MFT defines a moral foundation as an innate affective disposition underlying moral judgment across cultures, and conjecture that each constitutes an evolutionary adaptation which has been selected as a response to stimuli from a particular domain. The preliminary list in (Graham et al. 2013) includes five such foundations: *Fairness*, *Care*, *Authority*, *Loyalty* and *Sanctity* (with *Liberty* as a possible sixth one). The moral foundation of *Sanctity*, for instance, is associated with the emotion of disgust, and is conjectured to have been selected for helping us avoid infections. In time, the moral foundations have come to be generalized to apply outside the domains for which they were adapted, which is why moral transgressions can trigger disgust-reactions even when there is no risk of infection involved. MFT explains differences between the moral views of people and cultures as differences in the degree to which these moral foundations influence moral judgment.

On at least one reading of MFT, then, cooperation is only one type of outcome (though an important one) for which the moral faculty has been selected. MFT buys this conclusion at the price of construing the moral faculty as a plurality of functionally distinct psychological faculties, foreclosing the possibility of a unified theory of the function of morality. We can call this type of view "variantism" about the function of morality ("pluralism" would be a better term, but I reserve that for a different, though related, view to be discussed in the next chapter).

A partisan of the cooperation view has several options for dealing with the sort of anthropological and social-psychological data advanced by the variantist. Simply because an emotional faculty has some evolutionary function, and this function is also (in some cultures) moralized, that function need not be an evolutionary function *of the moral faculty*. It is conceivable that the moral faculty has been selected for its capacity to promote cooperation, but that in so doing, it also sometimes has the added benefit of promoting the ends of other faculties.

It is easy to see how this could happen, since cooperation is not, evolutionarily speaking, an end in itself but simply a means whereby organisms can pursue various further ends. Someone's failure to act in accordance with my preferences can be a failure to act cooperatively, even if my preferences are not themselves preferences *for cooperation*. So if I feel strongly about hygiene and my partner refuses to shower regularly, I may come to resent this behavior not only because I find it disgusting, but also because it evinces inadequate respect for my preferences: my partner is not cooperating in creating a domestic environment that is satisfying to us both. When a preference becomes moralized, i.e. when a coordinated sanctions regime is erected around it, failure to comply will easily be seen as a sign of inadequate respect for the whole group. Indeed, since behavior in defiance of the preference demands costly sanctions, it imposes a cost on the group and is, for that reason alone, easily perceived as uncooperative (cf. Tomasello 2015, 100).[106]

Here, however, the variantist could turn the tables on the proponent of the cooperation view, because the latter has just admitted that cooperation is no more than a *mechanism* whereby various further goods are attained. Shouldn't the attainment of these further goods, rather than the mechanism whereby they are attained, count as the *real* purpose of the moral faculty?[107]

As teleosemanticists we don't have to choose. Something can have more proximate and more ultimate functions. And indeed, by defending the CoRH I have committed myself to cooperation being among the more *proximate* outcomes that moral judgment are supposed to bring about, since it involves the idea that moral judgment perform their functions by coordinating *re-*

[106] Richard Joyce argues convincingly that pro-social emotions like altruism in themselves are insufficient for morality. Creatures filled with compassion or empathy would still lack the capacity to make moral judgments as long as they lacked the notion of a prohibition (Joyce 2007, 50–51). Generalizing Joyce's point, it seems as though no preferences are themselves sufficient for a creature to possess morality unless they are accompanied by the conception that conduct in defiance of those preferences is prohibited. If this is right, and if the notion of a prohibition, as seems plausible, requires the existence of socialized sanctions, then the production of a coordinated sanctions regime seems likely to be an essential component of how moral judgments perform their function, as I have proposed above.

[107] This type of response, however, is not available to the proponents of the actual moral foundations theory, who count "fairness/cheating," i.e., emotional faculties adapted for cooperation, among the moral foundations together with "care/harm" and "sanctity/degradation," i.e. moral foundations adapted to produce substantial outcomes (physical welfare and hygiene, respectively) (Graham et al. 2013, 12–14).

sponses, which is a kind of cooperation (cooperating to punish moral transgressions and reward moral conformity). As I mentioned on p. 45, the distinction between proper function and Normal mechanism is not absolute.[108]

I cannot here adjudicate the debate between the variantist and the proponent of the cooperation view. However, if we can't give a determinate answer to the question of what the invariant function of the moral faculty is, we will have to rely on terminological abstractions in order to discuss the issues downstream from this question. I will say that the invariant function of the moral faculty is to produce *successful social states*: successful, because they constitute teleological success for the faculty; social, because they (per the CoRH) constitute society-involving states of affairs. If the cooperation view is right, then a successful social state is, roughly, one where people cooperate. If the variantist is right, successful social states may include ones where people are cleanly. It is worth reminding ourselves, as well, that whichever of these is correct, the CoRH predicts that teleological success for moral judgment can come in degrees (cf. p. 149). Hence, social states can be more or less successful.

Even if the cooperation view is correct, it doesn't follow that there is a single type of state that every single moral faculty is supposed to produce and maintain, because there are many diverse types of situations that require cooperation, many ways to cooperate, and many ways to distribute the proceeds of cooperation. To see this, it is enough to recall that cooperation typically requires the solving of coordination problems and that these, by definition, admit of several different solutions. If the moral faculty is to have any chance of helping to produce cooperation, there must be some means for it to "choose" between these alternative solutions. More formally, there must be some principles that determine which, out of the various possible coordination schemes, an individual moral faculty is at any given time supposed to help bring about, as well as mechanisms that allow it to do so. Likewise, even if the variantist is right and the moral faculty is for producing a number of diverse states, insofar as these states or combinations of them constitute incompatible alternatives there must be some principles that determine which state or mutually compatible combination of states an *individual* moral faculty is at any given time supposed to bring about. The moral faculty, presumably, has not persisted because it has worked at cross-purposes with itself.

[108] The cooperatively maintained sanctions regime can, of course, in turn be directed towards maintaining various forms of first-order cooperation. This requires agreement among the would-be sanctioners on what first-order cooperation should look like. Issues they must resolve include the circumstances under which a person will and will not be sanctioned for failure to participate in a cooperative endeavor, how payoffs from cooperation should be distributed among the cooperators, and which benefits are to count as payoffs from cooperation in the first place. Differing solutions to these coordination problems correspond roughly to differing conceptions of duties to cooperate, fairness, and the extent to which certain goods should be considered (more-or-less) public versus (entirely) private property.

There are, as far as I can see, two possibilities. Either the moral faculty is hard-wired to promote *one* successful social state out of the alternatives, or it is developmentally flexible, so that which particular state an individual moral faculty is supposed to bring about depends on its particular history and circumstances.

At one extreme, we can imagine the moral faculty as genetically hard-wired, our moral judgment dispositions more or less fixed at birth. It would be a "mental module" in the style of (Fodor 1983), one that ties a specific range of stimuli (cooperative vs. uncooperative behavior, for example) to a specific range of responses. Such a faculty would be directed towards a single type of outcome, and coordination among people would be secured through phylogenetically pre-established harmony. In view of the extent to which reasoning, acculturation and circumstances can influence our moral convictions, this seems implausible to me. Nevertheless, it reflects a fairly common conception of morality, one that views it as intuitive, an object of private reflection or *a priori* insight, and absolute.

At the other extreme, we can imagine the innate component of the moral faculty as little more than a developmental toolbox from which a wide array of different moralities can be built, corresponding to and underpinning different successful social states. A moral faculty of this kind would be developmentally plastic, designed to develop in different ways under different circumstances—somewhat like how the language faculty is designed to produce competence in different languages depending on the subject's linguistic environment (cf. Joyce 2007, 10). A developmentally flexible moral faculty could be designed to be sensitive, in its development, to a number of factors (the local climate, perhaps, or the amount of resources available), but in order to secure coordination of responses with others, it would primarily have to be sensitive to the moral convictions of people around it. This possibility correspond to a broad picture of morality as socially mediated, culturally contingent, and (in some sense of that ambiguous word) "relative."

If something like this latter picture is correct, what does that entail for the invariant function of the moral faculty? Each *individual* moral faculty would have an invariant function that depended on its particular developmental and environmental circumstances. We can assume, furthermore, that the invariant function is typically something the moral faculty has some means of performing. If this is correct, and if a moral faculty Normally performs its function by coordinating responses with others, we might conclude that its invariant function is typically to fit together with and contribute to whatever sanctions regime it finds itself surrounded by. However, it can hardly be teleologically irrelevant *towards which ends* this sanctions regime is directed. Even if the moral faculty is highly developmentally flexible, there could be conditions that "subvert" its Normal developmental program so that it becomes directed towards ends that cannot be considered proper functions for it.

To lend credence to this latter point, consider that it is possible for agents to moralize ends that are detrimental to their own fitness and hence, one assumes, detrimental to the fitness of their moral faculties. For an example, imagine a religious sect that makes a moral virtue out of sexual abstention, not only for a select priestly class but for the whole population.[109] Moreover, some ends that could be moralized might more readily be brought about, and have been brought about historically, by other mechanisms. For example, we could imagine a society that moralized purely prudential, self-regarding behavior, even though such behavior is more effectively engendered by enlightened self-interest.

We can further observe that if an end is "inaccessible" to the moral faculty because it conflicts with other features of human psychology or is precluded by other types of broad limitations on human behavior, that end cannot be a proper function of the moral faculty and judgments in favor of it must be considered abNormal. For instance, it presumably cannot be the, or a, function of the moral faculty to help produce the *utilitarian's* ideal state of affairs, i.e. one in which everybody (or even some significant proportion of people) always chooses that course of action from among their alternatives that will maximize total happiness. This is because people's cognitive and epistemic limitations rule out the kind of detailed insight into the causal consequences of their actions that would be necessary for them to make happiness-maximizing choices with anything approaching consistency (as has been forcefully argued by Lenman 2000).[110]

---

[109] If morality, as seems plausible, is in part a learned behavior, it is also at least conceivable that even though a given morality (perhaps even our own) constitutes a subversion of the Normal developmental program, it persists and spreads in human populations for reasons orthogonal to the evolutionary function of the developmental factors that allow it to arise. It could constitute a "selfish meme," a parasitic entity that persists by subverting innate motivational mechanisms, just like many bad habits can persist and spread by hijacking innate reward systems and directing them towards unintended ends (for a straightforward comparison, consider the practice of using opiates, which survives and thrives in human populations despite its often detrimental effects on the fitness of its practitioners).

If this were the case, morality would have its own purposes independent of the purposes of the motivational systems it relies on in order to persist. Those purposes would not, however, be relevant for understanding the *intentional content* of moral judgments, since the latter, on the teleosemantic analysis, derive from the *cooperative* function of representations.

[110] Readers sympathetic to utilitarianism should note that the conclusion that judgments in favor of utilitarianism are abNormal presupposes the assumption, which I have made the basis of the CoRH, that the function of a moral judgment is to bring about a state of affairs in which everybody (or sufficiently many) act in accordance with it. This assumption could be challenged even within the overall teleological approach to ethics I am following here. We could speculate that judgments in favor of utilitarianism have contributed to the persistence of the moral faculty, not by actually bringing about a utilitaria state of affairs, but by motivating people to *attempt* to bring about such a state of affairs, thereby bringing about some *other* state of affairs capable of explaining the persistence of the moral faculty. This would mean, however, that the moral faculty engages in a strange sort of subversion of our overall motivational machinery, making us pursue a certain outcome in order to bring about another, different outcome.

If there are these kinds of abNormal ends, a moral faculty directed towards one of them would be somewhat akin to a malformed trait, one that is no longer capable of performing its proper function.[111] If a moral judge finds herself in circumstances where the moral faculties of the surrounding agents are malformed in this way, adapting to them would *not* be a way for it to perform its invariant function. In such a scenario, the agent had better be able to influence her group to change their moral convictions, or else change groups herself, if she wants her judgments to attain teleological success. We will return to these issues, and how they bear on the question of the descriptive content of moral judgments, in section 6.4.

Another issue worth addressing, in connection with the picture of the moral faculty as developmentally plastic, is whether and to what extent the judgment dispositions arising from it under varying circumstances always qualify as "moral." So far I have been using the term "moral faculty" in a loose way to denote whatever innate endowment underlies our ability to make moral judgments, but it is conceivable that this endowment could, under certain conditions, develop Normally in such a way as to yield something we would be hesitant to call "morality. Here, the question that we considered in 5.1, about how to delimit the category of moral judgment, returns in a new guise, and it brings with it another question: is morality a cultural universal?

The view that it is has been challenged, both by philosophers (like Elizabeth Anscombe (1958) and Bernard Williams (2011); see also (Gibbard 1990, 54)) and on empirical grounds (Machery and Mallon 2010). Of course, when these writers question the cultural universality of morality, they don't mean to suggest that some cultures lack social norms, or cooperation, or altruism, or methods for maintaining social cohesion. It is hard to see how something could even be a *culture* without those things. They just identify "morality" with a rather narrow syndrome of views, conceptions and dispositions, and deny that *this* exists cross-culturally. In this chapter, I have been focusing explicitly on *paradigmatic* moral judgment, which is also a rather narrow syndrome of psychological and behavioral dispositions. We therefore cannot dismiss the possibility that paradigmatic moral judgment is something that arises contingently in only some groups.

But even if a developmentally flexible moral faculty entails the *possibility* that morality is culturally contingent, we must be careful not to conclude that

[111] Here it may be noted that the moral faculty's degree of developmental plasticity and its degree of *functional* plasticity are, at least in principle, independent variables. We can imagine that the moral faculty is developmentally flexible, but not, as it were, by design: it just happens to lend itself easily to subversion by abNormal developmental conditions and to being directed towards ends that do not accord with its proper function. Such a faculty would be fragile, something akin to an organ that can develop Normally only under very specific conditions that seldom obtain in the organism's actual range of habitats. One would expect such an organ to eventually either be lost entirely, acquire more robustness against variations in external circumstances, or become exapted to new uses.

it can only be culturally *universal* if the moral faculty is phylogenetically hard-wired. Sinclair entertains an argument of the latter kind (2012, 648–49). But even a developmentally flexible moral faculty could produce similar outcomes across variations in external circumstances—similar enough, at least, to all qualify as "moralities." Morality could be what Daniel Dennett calls a "good trick" (1995, 485–87), a simple technique easily invented by each human culture in the face of universal human problems.

As I suggested above, the choice between a phylogenetically hard-wired and a developmentally plastic moral faculty is also a choice between two different views on how the moral faculty Normally performs its function, and with that comes two different visions about the role of morality in human life: as (more or less) intuitive and absolute, or as (more or less) contingent and a product of culture. These correspondences are only rough indications, meant to paint a picture and anchor the preceding abstract discussion in something that is hopefully more familiar to the reader. In the next chapter, I will attempt to get more precise. I will relate the CoRH to the question of whether morality is *objective*, and what further assumptions—of both a philosophical and an empirical kind—we would have to make in order to secure that conclusion.

5.5. Summary and Conclusion

In this chapter I have presented a theory about the function of moral judgment, the Coordination of Responses Hypothesis, or CoRH. This theory is a close descendant of views defended by, among others, Allan Gibbard and Neil Sinclair.

Is the CoRH true? I have not offered much in the way of empirical support, relying instead on the authority of my predecessors to lend it credence. An empirical test of the theory goes beyond the remit of the present thesis, and would moreover quickly involve us in conceptual difficulties (such as how to delimit the category of moral judgment) that would require a study of its own. I can only hope that the reader shares my impression that the theory is plausible or that, if she does not, she will at least allow me to rely on it as an assumption in what follows.

In the next chapter, I will turn to the question of moral objectivity: is there one true morality, and if not, how can we account for the many features of moral thought and discourse that seem to indicate that there is? I will attempt to draw out the consequences for efforts to answer these questions with the tools of teleosemantics if the CoRH, or something sufficiently like it, is true. This discussion will also require us to consider moral *discourse*, and to return to the discursive non-descriptivism developed in the previous chapter.

6. Moral Objectivity

A central concern for meta-ethics is whether morality is *objective*: whether there are objectively correct, universally valid answers to moral questions, and whether a moral claim is true or false absolutely, regardless of who makes it or the context in which it is made and regardless of the perspective from which it is assessed.

In the last chapter I declared my allegiance to an evolutionary, teleological, and naturalistic view of the moral faculty. In this chapter, I would like to explore the consequences of this view for the question of objectivity. Can an evolutionary view of moral judgment, together with teleosemantics, vindicate moral objectivity? If not, can it account for any of the features of moral thought and discourse that seem to speak in favor of objectivity?

Objectivity is tricky to define. I will attempt a more precise definition in 2.1 below. For now, let us somewhat impressionistically characterize *moral objectivism* as the view that moral judgments are truth-apt and that, necessarily, any given moral judgment has one and the same set of truth-conditions regardless of who makes it and regardless of the perspective from which it is assessed.[112] I will try to substantiate this somewhat by mentioning a few views that, intuitively, are *not* forms of objectivism.

First, let us consider moral *error theory*, roughly, the view that moral terms *purport to* but systematically *fail to* refer to substantial properties, or refer to properties that are uninstantiated, and hence that no moral statements (or at least no atomic moral statements) are true (cf. Mackie 1990; Olson 2014). This view is at least compatible with objectivism according to the above characterization, but it is not what I have in mind with objectivism, and I mention the view only to set it aside.

Objectivism is a weaker view than *realism*, as the latter term is commonly used. Realism is typically understood to involve some further commitments: that moral truths are *mind-independent*, that moral judgments are made true or false by *metaphysically substantial* and perhaps also *intrinsically normative* properties,[113] etc. In particular, objectivism includes, whereas realism is

[112] The reader will note that criterion (2) would seem to preclude the sort of *discursive non-descriptivism* I have developed for ought-judgments in chapter 4. As the reader may have suspected, I am ultimately interested in developing a similar kind of view here—but only after we have given objectivism a fair chance.

[113] For discussion of the last desideratum, see (FitzPatrick 2009, 749).

often taken to exclude, various forms of *contractarian* or *constructivist* views on the nature of morality (e.g. Rawls 1980; Gauthier 1986; Korsgaard 1996; Scanlon 2000; O'Neill 2015). If an evolutionary account of morality can vindicate some form of objectivism, it is likely to be more closely associated with constructivism than with realism, since many evolutionary accounts represent evolutionary recapitulations of constructivist procedures (cf. p. 156). Hence, I have chosen to focus on objectivism rather than realism here. For many of our purposes below, however, the distinction will not be terribly important, and I will sometimes be sloppy with the terminology.[114]

A direct alternative to objectivism are those forms of contextualism or relativism that take a moral judgment to be true or false relative to standards that are idiosyncratic to the judger or to her culture (e.g. Harman 1975; cf. the indexical contextualism discussed on p. 111).

What about non-cognitivism? The issue is complicated. Non-cognitivism is often characterized as a position that denies that moral judgments are truth-apt in the first place (van Roojen 2018), and such a view clearly does not qualify as "objectivism" as I have defined it. But most non-cognitivists allow that moral judgments *are* truth-apt, while defending one or another flavor of minimalist or deflationist theory of truth. Such views make "is true" into something like a discursive tool for reaffirming a sentence, which makes the objectivist claim, as I have formulated it above, hard to evaluate without a correspondingly deflationary proposal for the semantics of "truth-conditions," especially as embedded under a necessity operator. Let us agree that non-cognitivism is not in the spirit of moral objectivity (although a quasi-realist like Simon Blackburn would probably deny that "objectivity" marks a meaningful distinction in the first place (see e.g. Blackburn 1998a)).

Both objectivism and its competitors have some intuitive appeal, I think. Intuitions hostile to objectivism derive primarily from the fact that we are often unwilling to judge people by our own moral standards when they inhabit cultures or circumstances very different from our owb (see e.g. Harman 1975, 5), a theme I broached briefly in relation to ought-judgments in section 4.4. To this intuitive source of appeal we can add a more theoretical point. Critics of objectivism have emphasized the existence of widespread and seemingly intractable moral disagreement between cultures and among groups within the same culture (e.g. Mackie 1990, 36–38; Harman and Thomson 1996, 8–10; for an extensive discussion, see Tersman 2006). The existence of disagreement of this kind poses a problem for objectivism, since

[114] Peter Railton (1986, 164–65) offers a useful list of no less than thirteen factors that could be at issue in discussions over "moral realism." Of these, the two that are most directly relevant to making the distinction I want to make are "(12) Relativism—Does the truth or warrant of moral judgments depend directly upon individually- or socially-adopted norms or practices? (13) Pluralism—Is there a uniquely good form of life or a uniquely right moral code, or could different forms of life or moral codes be appropriate in different circumstances?" (Railton 1986, 165).

it would seem to require the objectivist to countenance the existence of massive and widespread mistakes in moral reasoning, ones that moreover seem resistant to rectification by rational argument. One can certainly doubt the existence and extent of such disagreement on both empirical and philosophical grounds (cf. Moody-Adams 2001; Tersman 2006, chap. 2). For the moment, let us leave disagreement on the table as a source of intuitions against objectivism, no matter how reliable.

In favor of objectivism, we can advance considerations pertaining to the *importance* of morality.

> Many try to live by their moral views. Some even die (or kill) for them. Behind the interest in questions about the objectivity of ethics lies the nagging suspicion that unless there is room for some objectivity, the role of moral thinking in our lives is somehow inapt. (Tersman 2006, 1)

The flipside of this "nagging suspicion" is that, on the assumption that the role of moral thinking in our lives is *not* inapt, this role in many respects points towards objectivism.

We can also advance considerations regarding disagreement *in favor of* objectivism. The very fact that we describe the purported situation of widespread moral disagreement as a case of *disagreement* would seem to suggest that the disagreeing judgments are governed by the *same* standard. Intuitively, out of two disagreeing parties, at most one can be right. If both can be right, or if there is nothing to be right about in the first place, it is unclear in what sense we are dealing with an actual disagreement. And the disagreements that morality involves us in are more than idle chit-chat. We often take quite seriously the task of converting those we disagree with or, if that seems infeasible, rebuking them. This all seems to imply the existence of a shared standard and the possibility of a shared consensus at the end of inquiry, two hallmarks of objectivity.

Moral debates, moreover, often take forms that are characteristic of arguments over objective matters. We give reasons for and against our positions, appeal to basic principles and try to reveal inconsistencies in our opponent's position, while searching for common ground.

In this chapter, I will try to evaluate the prospects for a form of moral objectivism on evolutionary, teleosemantic grounds, i.e., a form of objectivism that takes the moral faculty to be an evolved adaptation and that uses the teleosemantic principles of content-determination to assign objective (judge-independent, assessor-independent) descriptive content to moral judgments. Throughout, I will assume that the coordination of responses hypothesis that I described in chapter 5, or something close to it, is the correct account of the function of moral judgments.

I will begin this project in section 6.1 by discussing exactly what sort of view about the meaning of moral terms and the nature of moral judgments is

entailed by objectivism. In section 6.2 I briefly review prior debates concerning the compatibility of an evolutionary view of the moral faculty with objectivism. In section 6.3 I try to show what kinds of assumptions we need to make about the evolutionary history of the moral faculty in order to be able to vindicate objectivism, and attempt to make an initial assessment of the plausibility of these assumptions. Finally, in section 4, I try to show how we could account for objectivist intuitions, on the same evolutionary/teleosemantic grounds, if the strategies outlined in section 3 should fail.

6.1. Attitude Individuation

Objectivism, I have said, requires that a given moral judgment necessarily has one and the same set of truth-conditions regardless of who makes it and regardless of who evaluates it. Now, since two moral judges cannot both make the same judgment *token*, "given moral judgment" must be understood as "given moral judgment *type*." And so my definition yields us several different types of objectivism according to how we choose to type-individuate moral judgments.

If we interpret "truth-conditions" as "descriptive content" and "moral judgment type" as "attributive type," objectivism amounts to the view that attributive moral judgment types are descriptively individuated (cf section 4.2). On this interpretation, however, the indexical contextualist from p. 111, who believes that a predicate like "morally right" indexes different properties depending on the context in which it is uttered, would turn out to be an objectivist. According to the indexical contextualist, when you and I utter "φ-ing is morally right" in two different contexts we thereby attribute different properties to φ-ing. It follows that we in fact express *different* attributive moral judgment types, and so the fact that our respective judgments have different truth-conditions doesn't contradict objectivism as I have defined it.

The brunt of dialectical pressure against a view like contextualism comes from the observation that when you and I both say "φ-ing is morally right," we don't *seem* to be saying different things or expressing different judgments. We seem to be agreeing, to be saying the same thing, i.e., to be expressing the same attributive type. We can perhaps understand the objectivist, then, as someone who maintains that moral judgment types are descriptively individuated but takes our sameness-of-type intuitions at face value.

But this type of objectivist is liable to be hoist with her own petard. The intuitions about sameness of type that she presses on the indexicalist can be pressed on her in turn by a non-cognitivist, with examples pertaining not to different uses of the same word in the same language, but to different, seemingly synonymous words in different languages. The non-cognitivist can construct scenarios where, if two such words possess any descriptive content at all, they possess different descriptive contents, even though these words

are seemingly used to say the same thing. These scenarios purport to illustrate that we type and identify moral judgments based not on their descriptive content, but on what we may call (following Matti Eklund (2017, 10, 38)) their *normative role*: roughly, their action-guiding, deliberative, and/or motivational role. Let me briefly review some of these scenarios and the connected arguments.

R. M. Hare, the influential early non-cognitivist, discusses a case of a certain missionary who arrives on a cannibal island. He learns the word for "(morally) good" in the cannibals' language, which also happens to be "good," but he systematically applies it to other people than the cannibals themselves do: "people who are meek and gentle and do not collect large quantities of scalps; whereas [the cannibals apply it to] people who are bold and burly and collect more scalps than the average" (Hare 1952, 148). If the cannibals' word "good" were a descriptive (referential) predicate like "red," Hare maintains, the missionary's semantic competence with the term would manifest itself in his ability to apply it to the same things as the cannibals. But though this is not the case, the missionary and the cannibals can understand each other perfectly well, and there's no reason to think that the missionary is mistaken about the word's meaning. Hence, Hare concludes, the cannibals' "good," and by extension our own, are not descriptive predicates. Rather, their meaning is determined by their function of "commendation" (Hare 1952, 148–49).

A similar argument has been advanced by Terry Horgan and Mark Timmons (2007) against a version of moral realism, namely the *synthetic naturalist realism* or "Cornell realism" defended by Richard Boyd (1988) and David Brink (1989). The latter view is premised on the idea that moral predicates (like "good" and "right") refer to natural properties, specifically those properties that "causally regulate" our use of the predicates and thereby explain why the things we say using those predicates tend to be true. A moral judgment such as "murder is wrong" attributes one of these natural properties (the property of wrongness, in this case) to the act of murder.

Horgan and Timmons invite us to imagine another planet, Moral Twin Earth, which is almost identical to ours with one exception. They have a set of predicates that are phonologically and orthographically identical to our moral predicates: "good," "right," "wrong," etc. They call these predicates their "moral" vocabulary, and they use them in very much the same way as we do our moral vocabulary, e.g. in practical deliberation, decision-making, and so on. In other words, these predicates have the same *normative role* as our moral ones. But they are causally regulated by different natural properties. For instance, if ours are causally regulated by *consequentialist* properties, theirs might be causally regulated by *deontological* properties. The synthetic naturalist realist is forced to say that the Twin Earthlings' "moral" predicates mean different things than our corresponding moral ones. When they say "abortion is wrong," they don't mean that it's *wrong*, and if I were

to believe that abortion is morally permissible, I wouldn't be disagreeing with them. But this, Horgan and Timmons submit, is implausible. "Reflection on the scenario just does not generate hermeneutical pressure to construe Moral Twin Earthling uses of 'good' and 'right' as not translatable by our orthographically identical terms" (Horgan and Timmons 2007, 460).

Together with Hare's cannibal island, the Moral Twin Earth thought experiment generates a set of intuitions about the conditions under which words in an alien language can be said to say or mean the same thing as our moral predicates. We can account for these intuitions using the apparatus of discursive non-descriptivism. Such an account would look as follows. Moral predicates do not in fact refer to any determinate property at all. Their stabilizing function is to produce moral judgments in hearers, which Normally track different properties depending on the hearer's background and circumstances. The sameness of "meaning" between our moral predicates and theirs revealed by the thought experiments is sameness of stabilizing function.[115] Since the speech-acts made with these predicates produce attitudes with different descriptive contents in different hearers, the attributive types into which these attitudes fall are non-descriptively individuated. If, as I have suggested, objectivism entails that attributive moral attitude types are descriptively individuated, then this diagnosis of the two thought-experiments entails that objectivism is false.

It should be noted that if the diagnosis is correct, the assumption in which the two thought experiments are couched—that the aliens' moral *terms* have different descriptive content than our own—is actually incoherent. It can't be *both* that moral *terms* have the stabilizing function of attributing non-descriptively individuated attitudes *and* that they have a determinate descriptive content of *any* kind (this is the line of thought that led us to discursive non-descriptivism in section 4.2). Rather, what these thought experiments yield, if our diagnosis is correct, is a dilemma: either 1) we individuate moral attitudes descriptively, in which case we have to conclude that despite appearances our "good" means something different from the corresponding term in the cannibals' and the Twin Earthlings' lexicon; or 2) we individuate them non-descriptively and respect the meaning-intuitions pushed by the thought experiments, in which case we have to be non-descriptivists about moral terms.

How can the objectivist respond to the two thought experiments and the arguments made on their basis? She can, of course, deny the intuitions: despite initial appearances, the cannibals' and the Twin Earthlings' terms do not mean the same as ours. This is a risky move, however. In thus rejecting

[115] According to Millikan, the sameness-relation that obtains between two words (in two different languages, or in the same language) when we correctly claim that they "mean the same" is, typically, sameness of stabilizing function (Millikan 1984, 78). This view is specifically intended to provide a sameness-of-meaning relation defined over *non-referring* terms such as, for instance, logical constants.

the evidential force of sameness-of-meaning intuitions, the objectivist leaves her flank open to the indexicalist, who can likewise reject those intuitions that speak against his claim that moral terms in *our* language are indexicals.

Moreover, it behooves the objectivist to account for, or to explain away, the relevant intuitions *somehow*. A natural move at this point may be to insist that whereas the sameness-of-meaning intuitions pushed by the thought experiments might track *something* the respective predicate families have in common—perhaps something like *force* or *practical import* or (to use Eklund's term) *normative role*—whether or not they count as *moral* predicates still depends on their referents. But the result would be a rather unsatisfying kind of objectivism, one not very dissimilar to indexicalism in spirit. Though it would secure the result that morality is objective, it would do so at the price of admitting a plethora of morality-analogs, none of which has any obvious claim to being better, or more important, than any of the others.[116]

A different strategy for the objectivist is to try to account for the intuitions within her own framework. This is easier in Hare's cannibal case, since the case is under-described. The objectivist can insist that there must be some single property to which both the missionary and the cannibals purport to refer. She can account for the fact that their uses diverge in a way they wouldn't do for "red" by insisting that the property in question is epistemically less accessible than redness. Perhaps it is highly relational in a way that requires knowledge of non-obvious features of the environment to apply it correctly. This is not at all an implausible line.

The Twin Earth case is more difficult, since here it is *stipulated* that if the Twin Earthlings' moral terms have any descriptive content, it is different from that of ours. The stipulation is made assuming Boyd's and Brink's causal regulation-theory of reference. Perhaps this assumption can be jettisoned, but Horgan and Timmons have constructed an analogous argument for rival theory of reference analytical functionalism (Horgan and Timmons 2009). There is no reason to believe that further analogs could not be constructed for yet other theories of reference.

The objectivist's best bet might be to argue that Twin Earth scenarios are, without exception, metaphysically impossible. What would such an argument look like? Recall that one crucial assumption in the Twin Earth scenario is that the Twin Earthlings' terms have (what we called) the same "normative role" as ours. It is therefore open to the objectivist to insist, following a proposal by Eklund (2017, 10–11), that terms that have the same normative role *necessarily* also have the same descriptive content.

[116] This point roughly recapitulates the argument in (Eklund 2017, 4 ff.) Note, however, that I am only talking about morality, and this leaves open for the objectivist to appeal to a more general kind of normativity to make the requisite distinctions between morality-analogs (betterness, importance, etc.) Eklund's argument concerns normativity *simpliciter*, which blocks this move. I claim no conclusiveness for my point, but since I favor a different strategy for the objectivist, I will leave it here.

This, in essence, is the strategy I will explore on behalf of the objectivist in the coming two sections. I will do so within the evolutionary-teleosemantic framework I have been developing on the assumption that the CoRH, or something sufficiently like it, is an accurate theory of moral judgment. Within this framework, I posit, Eklund's "normative role" is best understood in terms of the *adapted proper function* of a judgment and the mechanism whereby that function is Normally performed. In other words, if the CoRH is correct, the normative role of the judgment that φ-ing is wrong will be equivalent to the function of contributing to making it the case that everyone or sufficiently many in the group abstain from φ-ing, Normally performed by means of coordinating punishment of φ-ers and rewards to non-φ-ers. The same will hold, *mutatis mutandis*, for the normative roles of other kinds of moral judgments, following the function-specifications on p. 147 above.

The strategy then becomes to show that any judgment faculty recognizable as a moral faculty, because it produces judgments with the same or a sufficiently similar kind of function as the judgments produced by my moral faculty or yours, would also necessarily produce judgments whose descriptive content are the same as those of the corresponding judgments from my faculty or yours. I will explore the constraints on this strategy, and the options available for bringing it to completion, in section 6.3. We may observe straight away that it relies on the assumption, which I have defended and employed in chapters 3 and 4, that even motivational states (a category which, if the CoRH is correct, includes moral judgments) have descriptive content.

To sum up the above discussion, and to give us something concrete to work with in what follows, let me offer the following definition of moral objectivism:

> MORAL OBJECTIVISM $=_{def}$ the view that 1) moral judgments have descriptive content, 2) they are descriptively type-individuated, 3) they are sometimes descriptively correct, and 4) our sameness-of-type intuitions regarding them should be taken at face value.

By negating different conjuncts out of the four that make up the definition, we get different rivals of objectivism. By negating (1), we get a kind of non-cognitivism that is inconsistent with my assumption that the universal hybridity thesis is true. By negating (3), we get the error theory. By negating (4), we get indexical contextualism or the indexicalism-adjacent bastard form of objectivism described above. By negating (2), finally, we get the view that I am most interested in pitching against objectivism, a version of the discursive non-descriptivism defended in chapter 4. I will call this view "moral pluralism." When I have exhausted my tentative case for objectivism in section 6.3, I will turn to make an equally tentative case for pluralism.

6.1.1. Individuation Problems for Non-Cognitivists

If moral pluralism is true, i.e., if moral judgment types are not descriptively individuated, they must be individuated in some other way. I have proposed, in accordance with the CoRH, that they are individuated by their *adapted proper function* and by the motivational mechanisms whereby that function is Normally performed. This is a form of non-cognitivism, broadly speaking. However, David Merli (2007) has argued that the same kind of Twin-Earth style arguments that the non-cognitivist directs against the realist can also be turned around on the non-cognitivist herself. The non-cognitivist aspires to account for moral agreement and disagreement in terms of agreement and disagreement in motivational attitude and for the sameness of meaning of moral sentences in terms of sameness of motivational attitude expressed/conveyed by them. For this to be more than a loose suggestion, the non-cognitivist must characterize the relevant type of motivational attitude and, in so doing, he must ensure that the attitude he characterizes is distinct from other, non-moral motivational attitudes (say, mere desires or attitudes of aesthetic disapproval) (see also Harman and Thomson 1996, 110; Miller 2013, 46). The problem, according to Merli, is that it is possible to construct scenarios analogous to Hare's and Horgan and Timmon's for plausible candidate characterizations. If we tie moral attitudes to the emotion of guilt, for instance, we can imagine a community that lacks this particular emotion but whose members still seem capable of making moral claims and of understanding our moral claims, disagreeing with us, etc. (Merli 2007, 39–43).

The issues here are complex, and my own intuitions vacillate. I have expressed tentative sympathy (on p. 161) for the view that morality in the narrow sense could be culturally and historically contingent and left room for this possibility in my formulation of the CoRH. If there are (possible) communities that lack moral judgments as the CoRH defines them but which we would nevertheless be able to treat as making moral claims, this would spell trouble for the CoRH. But it is worth noting that those philosophers who consider morality to be culturally and historically contingent typically complement this claim precisely with a denial of cross-cultural and cross-historical moral understanding.[117]

Let us leave these issues to the side and suppose that Merli is right about our sameness-of-meaning intuitions. Following (Björnsson and McPherson 2014), I deliberately construed the CoRH as a theory of *paradigmatic* moral

[117] Some examples: "In the actual world which we inhabit the language of morality is in [… a] state of grave disorder. What we possess […] are the fragments of a conceptual scheme, parts which now lack the contexts from which their significance derived […] we have – very largely, if not entirely – lost our comprehension, both theoretical and practical, of morality" (MacIntyre 1984, 2); "the concepts which are prominent among the moderns seem to be lacking, or at any rate buried or far in the background, in Aristotle. Most noticeably, the term 'moral' itself […] just doesn't seem to fit, in its modern sense, into an account of Aristotelian ethics" (Anscombe 1958, 1).

judgments, and I believe it is well suited to form the basis of a response to Merli that likewise takes its lead from these two authors.

The core of Björnsson's and McPherson's response is their view, which I have also subscribed to, that paradigmatic moral judgments are *function-bearing* and that their function crucially involves coordinating the actions of individuals (their account is specifically concerned with wrongness-judgments, but the point generalizes) (2014, 17). Björnsson and McPherson aim to address Merli by providing an explanation for *why* we should have the intuition that even people whose judgment faculties differ significantly from ours count as moral judges. They suggest that a person capable of paradigmatic moral judgment will have "accommodating" classificatory intuitions, being prepared to count as moral judgments psychological states that diverge in various respects from the paradigm. That is because, in their words, "classificatory intuitions should themselves be understood as elements" of the function of the paradigmatic moral faculty (2014, 21). Accommodating classificatory intuitions is itself a means of promoting the function of a paradigmatic moral faculty because, by treating even non-paradigmatic subjects as moral judges, engaging them in moral discourse, sanctions, and so on, the paradigmatic judge will be more likely to influence their actions towards coordination (2014, 22).

If this view is right, what is required of a non-paradigmatic judgment for us paradigmatic moral judges to accommodate it and treat it as moral? The account predicts that we should be accommodating towards a moral judgment insofar as such accommodation actually supports the function of our own judgment, i.e., insofar as the interlocutor's judgment is produced by a faculty that can be exploited so as to contribute to bringing about coordination. An alien lacking a paradigmatic moral faculty may still evince a sensitivity to social pressure, supported by emotional mechanisms encouraging long-term strategies, etc.; and if we can influence these psychological faculties in the right way, we can make the alien coordinate with us. Such an alien should be treated as a competent moral judge.

If this is right, it should also mean that if, as a response to some moral claim of ours, an alien interlocutor forms a non-paradigmatic judgment, we should treat this as an episode of successful communication to the extent that this judgment is of a kind that is capable, in accordance with its function, of helping the judgment that we *expressed* by our claim perform *its* function. As teleosemanticists, we could ask whether this episode is *actually* an instance of successful performance of the statement's stabilizing function, rather than just being treated as such by us for the sake of accommodation. And conversely, if *we* responded by tokening paradigmatic moral judgments to one of the *alien's* would-be moral claims, would that claim thereby have performed *its* stabilizing function? If it is really the case that we can translate the alien's moral vocabulary to our own, and that the alien's judgments can

belong to the same attributive types as ours, then the answer to both questions must be "yes."

I believe there's no clear-cut answer to these questions. There can be marginal cases of successful function performance that bleed gradually into cases of non-performance with no sharp demarcating line. If, as the title of Millikan's main work insists, language is a "biological category," we can expect linguistic meaning to be as fuzzy as other biological categories. But there is no particular reason to be stingy with what we consider sameness of meaning here. Compare regular descriptive discourse: perhaps there are some aliens who do not possess a paradigmatic human *belief* faculty but instead represent the world in some subtly different way. Yet their "belief"-faculty is sufficiently similar to ours that they possess a language with a descriptive vocabulary that we can learn, just as they can learn our language. It would be strange to insist that there could be no translations between our respective languages.

With this matter settled, I will go on to consider the prospects for objectivism. I won't be the first to use teleosemantic tools to try to defend some version of moral objectivism in the face of skeptical assaults. Before making my own positive case, I would like to take a look at my predecessors and the strategies they have employed.

6.2. Teleosemantics and Evolutionary Debunking

A number of authors have argued that if the faculty of moral judgment is an evolutionary adaptation, moral realism is likely to be false (Ruse and Wilson 1986; Street 2006; Joyce 2007, chap. 6). The interest in a teleosemantically informed understanding of moral judgment lies at least partly in its ability to blunt the force of these arguments (Harms 2000; Artiga 2015). I would now like to take a look at one version of this dialectic, in order to illustrate how the problem of objectivity arises from it.

(Street 2006) is probably the most influential instance of evolutionarily grounded skepticism against moral realism. In this paper, Sharon Street purports to develop a *reductio* argument against moral—and, more generally, normative—realism based on the assumption that our moral judgment propensities are the products of evolution through natural selection. The argument takes the form of a dilemma. Assuming there to be mind-independent moral truths, and assuming that our dispositions to make some moral judgments rather than others are the result of evolution, the realist has two options: either deny that this evolutionary process has been influenced by the moral truths, or affirm that it has. The former option leads to the unpalatable conclusion that our moral judgments are generally unjustified and, barring a crazy coincidence, most of them are likely false (Street 2006, 121–25).

The second horn of the dilemma is more interesting for us. Street concedes that it is an option for the moral realist to posit that our moral judgments have evolved to *track* the mind-independent moral truths. This *tracking account* states that "the [widespread presence of some evaluative judgements rather than others in the human population] is explained by the fact that these judgements are true, and that the capacity to discern such truths proved advantageous for the purposes of survival and reproduction" (2006, 126). But, Street argues, there is a better explanation for the data than the tracking account. This is the *adaptive link account*, according to which "tendencies to make certain kinds of evaluative judgements rather than others contributed to our ancestors' reproductive success not because they constituted perceptions of independent evaluative truths, but rather because they forged adaptive links between our ancestor's circumstances and their responses to those circumstances, getting them to act, feel, and believe in ways that turned out to be reproductively advantageous" (2006, 127). If the adaptive link account is correct, the moral truths again have no influence over our moral judgments, and the realist is back on the first horn of the dilemma.

Marc Artiga objects that the tracking account and the adaptive link account need not constitute *alternative* accounts. If teleosemantics is correct, the evolutionary success of a representational system is explained *both* by its ability to produce adaptive behavior *and* by its ability to (often enough) represent correctly (Artiga 2015, 3370).[118] As Artiga points out, by itself, the ability of a system to track some environmental factors or world affairs could hardly explain its evolutionary success. A tracking system of this kind must also have an effect on the organism's behavior that *adapts it to* the world affairs it tracks. So if the tracking account were a genuine rival to the adaptive link account, Street's skeptical conclusions would wildly overgeneralize: tracking could not explain the success of *any* representational system (2015, 3371). It's a good thing, it seems, that they're not genuine rivals.[119]

There is some dialectical distance between saying that moral judgments track *something* and saying that they track mind-independent moral truths of the kind the realist would prefer existed. Artiga acknowledges as much: "…to show that [the tracking account] is true one would need to show that

[118] This goes at least for descriptively representational systems. But if the conclusions of chapter 3 (also due to Artiga) are correct, all representational systems are descriptive (as well as directive).

[119] In fact, I believe Artiga undersells the dialectical force of teleosemantics against Street's argument. If teleosemantics is true, not only are the tracking account and the adaptive link account not genuine rivals, but the notion that the moral faculty could have evolved to adapt us to something *other than* what it represents is not even coherent. If a system's evolutionary success is explained by its ability to adapt the organism to the circumstances it tracks, then, insofar as it represents anything, it represents *those very circumstances*. So the idea that the moral faculty could have evolved because it adapted us to some circumstances *other than* those it purports to represent, hence yielding the kind of massive errors that are supposed to spell the moral realist's doom, is simply not coherent.

there are moral facts" (2015, 3370). Here, a possible further move on the realist's part is to deny that we have any independent grasp on the moral facts besides that they are what, *ex hypothesi*, are represented by our moral judgments. If so, and if our moral judgments represent something, then whatever that something is, it must *ipso facto* be the moral facts. And if our moral judgments track what they represent, they therefore track the moral truths.[120]

But this move would overlook the possibility of pluralism: that the moral judgments of different moral agents, in different circumstances or with different etiologies, Normally track different classes of facts, none of which would have a stronger claim to the title of "the moral facts" than any of the others. This is a possibility that we have yet to exclude but which follows naturally upon some of the theses I have advanced, together with auxiliary assumptions. First, consider the upshot of my discussion of the type-individuation of moral judgments in section 6.1 above. There, I said that consideration of Hare's and Horgan & Timmons's thought experiments suggests that moral judgments are type-individuated by what Eklund calls their normative role. I also proposed that we should identify the normative role of a moral judgment with its adapted proper function, as given by the CoRH. Second, on p. 140 I claimed that Normal conditions for a moral judgment, and therefore its descriptive content, depends on its *invariant* function. Finally, in section 5.4 I discussed the possibility that the moral faculty is developmentally plastic (p. 159), which could mean that the invariant function of a moral judgment varies with environmental and developmental conditions. If this is all correct, it would apparently follow that type-identical moral judgments could vary in descriptive content.

Whether or not objectivism can be vindicated on an evolutionary/teleosemantic basis would thus seem to depend, at least if the first and second of the above claims are granted, on the moral faculty's degree of developmental plasticity. In the next section, I try to spell out in more detail what assumptions we need to make about the teleology of the moral faculty in order to vindicate objectivism. Then, I attempt an initial assessment of the plausibility of these assumptions.

6.3. Prospects for Objectivism

I have suggested that the truth of objectivism—the view that moral judgments are descriptively type-individuated—depends on the degree to which

[120] The "tracking" vocabulary is not endogenous to teleosemantics, at least not in Millikan's version. I take it that the term typically denotes some rough, *ceteris paribus* counterfactual dependence. As we've seen (p. 49), it is a consequence of Millikanian teleosemantics that there will be some dependence of this kind between representations and their *representanda*, at least under Normal conditions.

the moral faculty's invariant function is subject to developmental plasticity. To make this claim more concrete, and evaluate the prospects for a moral objectivism based on teleosemantics and the CoRH, we first need to understand how the invariant function of the moral faculty helps determine a moral judgment's descriptive content.

As a point of departure, let us consider what Neil Sinclair has to say about the matter. As the reader may recall, Sinclair (2012) defends a teleosemantic theory of moral judgment similar to my CoRH. According to Sinclair the function of moral judgment is to enable us to solve *bargaining problems*, which constitute a subclass of coordination problems (cf. p. 144). Coordination problems require agents to coordinate their actions in order to yield mutually beneficial outcomes. In many cases, however, the different available coordination options lead to different distributions of costs and benefits among the cooperators. This means that the agents are incentivized to champion different coordination solutions, and thus, cooperation is liable to break down (cf. Schelling 1960). In that case, the cooperators need methods for reaching agreement on a specific solution, and the motivational dispositions to abide by these agreements. That is what the function of moral judgment consists in, according to Sinclair:

> The function of a judgement of the form 'Φ is wrong' for example, is to produce a stable pattern of co-ordination where no one Φ's, and relatedly, where people disapprove of Φing and disapprove of those who fail to disapprove of Φing. (Sinclair 2012, 653)

Though he doesn't use that terminology, producing a stable pattern of coordination is what Sinclair identifies as the *invariant* function of moral judgment. A stable pattern of coordination is thus what substitutes, in Sinclair's theory, for what I have called a successful social state (p. 158).

The primary purpose of Sinclair's paper, however, is not to defend this view of the function of moral judgment. Instead, like the present thesis, the paper is an attempt to intervene in some traditional meta-ethical debates using the tools of teleosemantics. To that end, Sinclair applies the teleosemantic principles of content determination to his function-hypothesis in order to derive the descriptive content of moral judgments:

> Under what past conditions did [the consumers of moral judgments], as so guided by judgements of wrongness, fulfill their function? These mechanisms would, by these actions, have produced beneficial patterns of co-operation only so long as judgements of wrongness pushed collective negotiation *away* from actions that are part of disadvantageous patterns of co-operation. Hence the Normal condition for the fulfillment of the consumer's function in this case would have been an appropriate type of mapping between judgments of wrongness and facts concerning which actions are part of mutually disadvantageous patterns of cooperation. Hence output-based teleosemantics assigns judgements of wrongness the descrip-

tive content: *Φ is part of a pattern of action that is mutually disadvantageous*. (Sinclair 2012, 656)

Sinclair also proposes a parallel analysis of moral *rightness* judgments, suggesting that the judgment that Φ is right has the descriptive content *Φ is part of a pattern of action that is mutually advantageous* (2012, 656).

In this quote, we see that Sinclair attributes descriptive content to moral judgments based on their invariant function. A right-judgment is descriptively correct if its target action is part of a pattern of action that it is an invariant function of the moral faculty to bring about, i.e., a successful social state (which, per Sinclair's hypothesis, consists in a mutually advantageous pattern of cooperation). And this seems like the correct verdict. Normally, a moral rightness-judgment produced a successful social state only if the action it encouraged were actually capable of contributing to or forming part of a successful social state.

We can generalize this to a principle for assigning descriptive content to moral judgments, assuming the CoRH or a CoRH-adjacent theory to be correct. Suppose a moral judgment has the adapted function to motivate the subject to engage in a certain behavior φ and to encourage her to make others do so as well. Suppose further that this judgment has the invariant function to bring about a certain type of successful social state *S*. If so, Normal conditions for the judgment's success will require that φ bears the right relation to *S,* for instance, that the state of affairs where everybody or sufficiently many φ itself constitutes a feature of *S*. That also seems like a good candidate for the judgment's descriptive content*:* that φ bears this relation to *S*. More precisely, the judgment's descriptive content is that the state of affairs which it is the *adapted* function of the judgment to bring about—*ex hypothesi*, the state of affairs where everybody φs (for right-judgments) or that nobody φs (for wrong-judgments)—also constitutes a feature of *S*. Let's make that indented:

(MORAL CONTENT) The descriptive content of a moral judgment is that the state of affairs, the bringing about of which is the adapted function of the judgment, is a feature of a successful social state, the bringing about of which it is the invariant function of the judgment.

The problem for Sinclair's analysis is that *prima facie* there are a number of *different* possible patterns of action that could all qualify as stable, mutually advantageous patterns of cooperation. As I have already stressed (p. 158), there are many different ways to cooperate and to distribute the proceeds of cooperation. So which of these candidate patterns of action should Φ be part of for the judgment that *Φ is right* to be descriptively correct?

One possible answer is: *some* such pattern, it doesn't matter which. Sinclair's own formulation suggests this answer, since he uses the indefinite "a

pattern." Read this way, Sinclair's proposal assigns uniform truth-conditions to rightness- and wrongness-judgments of the same type regardless of the context in which they are tokened. It would therefore be a form of objectivism. But it would not be a kind of objectivism one would want to defend. For one, it potentially allows seemingly contradictory moral judgments to be descriptively correct at the same time for the same judge. For suppose there is indeed some cooperative pattern, *A*, such that φ-ing is a feature of *A*. Suppose also that there is some other cooperative pattern *B* such that it is a feature of *B* that everybody *abstains* from φ-ing. Then, given the current proposal, my judgment that φ-ing is right and my judgment that it is right to abstain from φ-ing would both be correct. But these judgments are presumably contradictory. We were prepared for the pluralist possibility that mutually inconsistent judgments tokened in *different* contexts could both be correct, but an *objectivist* view that allows this same consequence is surely too much.

The proposal also severs the connection, so central to teleosemantics, between descriptive correctness and explanation of success. Suppose my tribe customarily engages in *A*, to great mutual benefit. Everybody φs along happily. Now, I get it into my silly little mind that the right thing to do is to abstain from φ-ing. On the proposed theory, this judgment is descriptively correct. But it seems very unlikely in my present social circumstances that my judgment would bring me much success of any kind. If I am really lucky, I might be able to convert my tribe to my new way of thinking and bring us all over to *B*. Most likely, however, I will make a fool of myself, suffer the disapproval of my tribesmen, and disrupt established cooperation arrangements.

The aforementioned difficulties illustrate that even if there are multiple ways that moral faculties have historically paid their evolutionary dues, an *individual* moral faculty cannot simultaneously be supposed to bring about several, incompatible outcomes. As I have said, we cannot suppose that the moral faculty has persisted by working at cross-purposes with itself. Rather, there should be principles that determine the invariant function of individual moral faculties as a function of its developmental and environmental context. If so, and if the invariant function of an individual moral faculty gives the descriptive content of the judgments it produces according to MORAL CONTENT, then the result, it seems, is pluralism. Conversely, in order to vindicate objectivism, we must be able to show that there is a single outcome or mutually compatible set of outcomes towards which all moral faculties are directed, perhaps one that is phylogenetically hard-wired.

In other words, suppose there were only one way to arrange social affairs, one successful social state, that explained the past persistence of the moral faculty via the latter's capacity to produce instances of this arrangement. Call that arrangement *L*. Then we could say, following MORAL CONTENT, that the content of the judgment that φ-ing is right is something like: the state of affairs where everbody φs constitutes a feature of *L*.

A view of this kind must contend with the fact of widespread institutional and normative diversity found among human cultures, a fact I have already gestured at in the introduction. The objectivist can insist that a majority of these are abNormal forms, misshapen products of malformed moral faculties incapable of performing their proper function (cf. p. 159). But this seems implausible. A better strategy would be to argue that the diversity revealed by the anthropological record constitutes mere variations on an underlying theme. The objectivist could claim that the pattern of action which it is the invariant function of the moral faculty to help produce and maintain is sufficiently abstract and relational that it assumes different proximal guises under different historical, cultural and environmental circumstances (exactly how different is a matter of debate). Nevertheless, so the argument would go, in all these cases we are dealing with a *single* pattern of action that the moral faculty is devoted to maintaining.

The reader could be excused for wondering whether there is any real difference between this picture and the picture of a developmentally plastic moral faculty that is assigned different invariant functions in different circumstances. In both cases, we are dealing with a moral faculty that, under different circumstances, contributes in different ways to maintaining successful social states. And even on the developmental plasticity picture, there would presumably have to be some abstract, relational principles that determine what individual moral faculties are supposed to do in their respective contexts.

There is a difference, and it is a matter of scope. For the objectivist to be right, the function of a moral faculty would have to be to contribute to maintaining a general pattern of conduct that has different concrete manifestations under different circumstances. The developmental plasticity picture, on the other hand, only entails that what the moral faculty is supposed to do depends on the context. In the former case, function takes scope over context; in the latter, context takes scope over function. In other words, the objectivist must argue that the moral faculty's function is to maintain the abstract pattern itself *across* variations in local circumstances. If this pattern is indeed one of which the full range of human institutional and normative arrangements are mere manifestations, then the objectivist is committed to the view that the function of the moral faculty is to help maintain a pattern of conduct across space and time, one that encompasses all of humanity.

How plausible is this? On one hand, it might seem as though morality is often much more about local accommodations that work until they no longer do than about maintaining some abstract pattern of conduct across time and space. This also seems to cohere better with an evolutionary understanding of morality: satisficing, not optimizing, is the name of the game in evolution as well as in individual learning. If our moral faculties can maintain cooperation (or other successful social states) on the local level and in the short term, why shouldn't that be sufficient to explain their persistence?

On the other hand, we must guard against a conception of human social life as restricted to clearly delineated, hermetically sealed groups. Humans have always interacted across tribal, cultural and other social boundaries, and it is not implausible that the moral faculty plays an important role in making these interactions proceed smoothly (when they do). The moral faculty could secure this boon by producing local accommodations anew in each novel encounter, but it could also do it by helping to maintain a single pattern of conduct across time and space, giving people a universal "language" of social interaction in which to conduct their business. We don't have to suppose that this function is something that a token judgment could accomplish all on its own, as it were. It would only have to be able to contribute to it. Given the value that such a universal pattern of conduct would have for enabling interaction across social boundaries, contributing to its maintenance may very well have been how the moral faculty pulled its evolutionary weight.

These alternatives also make a difference for how to cash out a promissory note left over from my statement of the CoRH on p. 149. As the reader may recall, the CoRH states that the adapted function of a moral judgment is to produce a pattern of conduct that holds "universally, or at least broadly, within some group of which the judger is a member," but I postponed discussion of exactly which group this referred to. Now we may note that if the invariant function of the moral faculty is indeed to maintain a pattern of conduct across humanity, the group must be identified with nothing less than humanity itself. If, on the other hand, the invariant function of the moral faculty is to produce more local accommodations, the group could be more circumscribed.

Suppose that we can accept the hypothesis that the moral faculty is supposed to contribute to maintaining this kind of universal pattern of conduct. The next step is to attempt to identify the pattern in question. Assuming that it is the, or a, job of morality to allow us to maintain our diverse institutional and customary arrangements, at least insofar as these are independently viable and conducive to long-term social benefit, the pattern would have to be one that entails conformity with such arrangements in the contexts where they hold sway. Now, if morality were involved in *producing* such arrangements, it would be hard to square the diversity of those arrangements with the hypothesis that morality is supposed to maintain a *single* pattern of conduct. But perhaps we could instead chalk the production of new institutional arrangements down to other psychological faculties—leave it open for now which ones—and attribute to the moral faculty the function of simply maintaining whatever (viable) institutional arrangement the subject happens to find herself involved in, by motivating her to participate in them, punish those who defect from them, and so on. In that case, we would have a unified explanation of the persistence of the moral faculty that appealed to a

single type of outcome: the maintenance of viable local institutional arrangements. Call this the "institutional maintenance hypothesis."

This is certainly a difficult hypothesis to evaluate, but it does seem to generate one empirical prediction. Suppose φ-ing is an essential part of my local institutional arrangement, whereas abstention from φ-ing is an essential part of yours, and compare two moral judgments I might make:

1. The judgment that *not φ-ing is wrong*; and
2. The judgment that *not* φ-*ing if you're a participant in my local institutional arrangements is wrong*.

If the current hypothesis is to generate moral objectivism, then clearly it can't allow (1) to be descriptively correct. If objectivism is true, (1) must be correct regardless of whose head it is tokened in. But why would (1) as tokened in *your* head be correct? That token judgment would tend to disrupt, rather than maintain, local institutional arrangement, and if it were to maintain them, it would be by a fluke rather than through its Normal operation. And indeed, it is hard to see how (1) as tokened in my head could contribute Normally to maintaining the general pattern of respect for institutions, rather than simply maintaining my own local institutional arrangements at the expense of other such arrangements.

Let an "accommodating" moral judgment be one like (2) above, one that explicitly conditionalizes the action judged wrong (obligatory) on participation in the relevant institutional arrangements. And let a "categorical" moral judgment be one like (1), which categorically judges wrong (obligatory) a type of action, with no conditionalization on membership in specific institutional arrangements. We would expect, if the institutional maintenance hypothesis were true, that whenever there is known variation among institutional arrangements with respect to some type of conduct, people would also be disposed to make accommodating judgments with respect to that type of conduct. Failure to find such dispositions would be evidence either of widespread moral incompetence and abNormality, or of the falsehood of the institutional maintenance hypothesis.

Note that we need not expect people to make conditionalized judgment with respect to *every* kind of conduct. Plausibly, there are certain broad patterns of conduct that are either part of every viable institutional arrangement or of none. One example might be fairness, in some abstract sense; or predictability; or the absence of wanton killing. But with respect to all dimensions of conduct along which viable institutional arrangements *do* vary, we would expect to find accommodating moral judgments—at least insofar as the judges are aware of this variation (if they are not, they might mistake a contingent feature for a necessary one without moral incompetence).

Does this prediction bear out? There certainly seems to be evidence pointing in this direction. Well-known results from developmental psychology

indicate that children from a young age differentiate between conventional and moral transgressions and judge the latter to be both more serious and more generalizable (conventional transgressions are wrong only where the convention holds sway, whereas moral transgressions are universally wrong) (Turiel, Killen, and Helwig 1987).

These results are typically interpreted by the researchers themselves, as well as by philosophical commentators (e.g. Nichols 2004), as evidence that people distinguish two *different kinds* of wrongness-judgments: moral and conventional. But we could also interpret the results as showing that people make *accommodating* moral judgments with respect to behaviors that are merely conventional in the sense that there are variations in them among viable institutional arrangements. The wrongness-judgments themselves would still be moral judgments, only of the accommodating kind. This would predict, no less than the standard interpretation, that children will judge conventional transgressions as only contingently wrong, i.e., contingent on the authority of the convention.

It will not predict, perhaps, the further observation that children judge conventional transgressions to be less *serious* than categorical ones, although I wonder whether this is generally true. There are certainly merely conventional rules, violations of which I personally am inclined to judge very seriously. Traffic rule violations are a good example. To drive deliberately on the wrong side of the road would be, in my eyes, a very serious violation. And there is an obvious reason why I feel that way: the roadside convention is crucial to prevent unnecessary deaths. This doesn't mean, of course, that I expect people to follow *my* local conventions wherever they are. They should follow the conventions that are local to them. The point is that a violation can be *serious* even if it is *conventional*. Many conventions regulate matters of life and death. As for Turiel et al.'s data, it might be that it is not so much the *conventionality as such* of the violations that make children judge them to be less serious, as the specific nature of the conventions involved.

So there seems to be an empirical case for the institutional maintenance hypothesis. Not everyone is convinced. Mackie writes:

> [T]hese [accommodating moral convictions] are very far from constituting the whole of what is actually affirmed as basic in ordinary moral thought. Much of this is concerned rather with what Hare calls 'ideals' or, less kindly, 'fanaticism'. That is, people judge that some things are good or right, and others are bad or wrong, not because – or, at any rate, not only because – they exemplify some general principle for which widespread implicit acceptance could be claimed, but because something about those things arouses certain responses immediately in them, though they would arouse radically and irresolvably different responses in others. (Mackie 1990, 37–38)

And it seems that categorical, "fanatical" support for local institutional arrangements could indeed sometimes, even often, be an adaptive move, even for agents who are aware of the conventional nature of those arrangements. Such categorical partisanship could be a way to signal respect for the group and its conventions and reinforce others' commitment to them. The objectivist will have to show that such deployments of the moral faculty are marginal phenomena, constitute secondary and parasitic uses, or both.

The empirical issue is beyond the scope of this inquiry, so let us turn to *a priori* considerations. I can see at least three *a priori* problems for the institutional maintenance hypothesis. One concerns the *origins* of institutional arrangements. If they do not originate in the moral faculty, then where do they originate? The cooperation thesis discussed in section 5.4 kills two birds with one explanatory stone, by accounting for the emergence of human cooperative arrangements in terms of our moral faculty *and* for the emergence of the moral faculty in terms of its ability to produce and maintain cooperative arrangements. The institutional maintenance thesis leaves us without an obvious explanation for the first of these.

A second, related problem derives from the fact that real institutions are not perfectly coherent, exhaustive bodies of rules, but hodgepodges of historical accommodations and compromises that cover certain paradigmatic situations but are completely silent on others. The world will constantly throw up novel situations that demand new forms of cooperation not covered by existing institutional norms, or covered only via creative casuistry. In these situations, too, we need an explanation for how humans manage to find cooperative solutions, and here, too, the institutional maintenance hypothesis divests us of the appeal to morality.

Third, for objectivism to be true it must be the case that a given moral judgment has the same descriptive content as tokened not just in the head of any *human*, but in the head of any moral judge whatsoever. Even if we can establish the truth of the institutional maintenance hypothesis for the human case, can we exclude the possibility that moral faculty analogs among sociable extra-terrestrials do not have different evolutionary histories that confer different invariant functions on their moral judgments?

Our case for a teleosemantically supported, evolutionary moral objectivism has been inconclusive at best. Much more could be said. Hopefully, I have at least been able to point out some directions for further investigation. Next, to finish this chapter and thereby this dissertation, I would like to make a tentative case that, whatever the verdict on objectivism ultimately turns out to be, the alternative view that I have called *pluralism*, which denies that moral judgments are descriptively type-individuated, can still account for many of our objectivist intuitions with the help of the discursive non-descriptivism outlined in chapter 4.

6.4. Making Room for Pluralism

Pluralism, we said at the end of section 6.1 above, is the view that moral judgments are non-descriptively type-individuated. This means that tokens of the same moral judgment type, like the judgment that *killing is wrong* or that *abortion is permitted*, can have different descriptive content in the heads of different people, and hence be descriptively correct for one subject and incorrect for another. We have seen some potential sources of such subject-relativity above. The descriptive content of a moral judgment could depend, for instance, on the group the subject belongs to and the institutional arrangements prevalent in that group. Here, I will attempt to determine what kind of moral pluralism might reasonably be true; what the descriptive content of a token moral judgment would be, according to it; and whether we can employ the apparatus of discursive non-descriptivism to account for some of the objectivist features of moral discourse and disagreement. Throughout the discussion I will continue to assume, as I have done above, the teleosemantic content-determination principle and the CoRH, including the principle of attitude individuation entailed by the latter.

6.4.1. Local Group Pluralism

Several forms of what has traditionally been called relativism entail that the contents of moral judgments depend somehow on the customs and values of some group of which the subject is a member. For instance, Harman (1975) holds that moral judgments are made true or false by an agreement, explicit or tacit, to which the subject is a party, and Ruth Benedict asserts that "morality […] is a convenient term for socially approved habits" (Benedict 2001, 87), meaning thereby "approved by the speaker's own society."

At first sight, this idea seems like a natural starting-point for assigning descriptive content to moral judgments on the basis of the CoRH. A moral judgment, according to the CoRH, is supposed to help standardize conduct *within some group of which the judger is a member*. In the last section, I discussed the possibility that the group in question could be all of humanity, but I also expressed skepticism towards this possibility. It might be more plausible that the moral faculty has pulled its evolutionary weight by standardizing conduct within much more circumscribed groups, such as tribes, cultures, sects, and so on. We can call such groups "local groups" and frame the hypothesis that the adapted function of a moral judgment is to standardize conduct within the subject's local group. Call this view "local group pluralism."

Now let us attempt to assign descriptive content to moral judgments on the basis of local group pluralism. To do this, we need to determine what the view entails about the Normal conditions for moral judgments. If we aspire to follow in the footsteps of Harman and Benedict, we might be tempted to

conclude that a moral judgment is descriptively correct as long as it accords with the moral convictions of other members of the local group. And indeed, a judgment with this quality seems well-disposed to perform its function successfully, as it will come into this world already blessed with many "allies," i.e., many tokens of its own type sitting in the heads of other people, waiting to assist it in the performance of its function of standardizing conduct within the group.

If this were all that were required for Normal conditions for a moral judgment to obtain, we might get something of the following sort as a specification of the descriptive content of *S*'s judgment that φ-ing is right:

(LGP) φ-ing is widely judged right within *S*'s local group.

(We could formulate a corresponding principle, *mutatis mutandis*, for wrong-judgments, but I will concentrate on right-judgments here). LGP straightforwardly entails that moral judgment types are non-descriptively individuated, since it entails that the same judgment tokened by people from different local groups have different descriptive contents. This result, moreover, need not immediately vitiate those of our intuitions about agreement and disagreement that comprise part of the case for objectivism. We could utilize the resources of discursive non-descriptivism developed in chapter 4 to account for some of them. As the reader may recall, DND (following Björnsson 2015) defines relations of agreement and disagreement across attributive judgment *types*, according to the inferential relations between *tokens* of those types as instantiated in the head of the *same* person. So even if local group pluralism allows that the descriptive content of *my* token of the judgment that φ-ing is wrong is consistent with the descriptive content of *your* token of the judgment that φ-ing is right, DND could still account for the intuition that the two judgments are in disagreement. You couldn't convince me to share a token of the *type* your judgment belongs to without forcing me to revise my judgment, on pain of inconsistency.

However, LGP is unlikely to be the correct specification of the content of moral judgments, even if local group pluralism is correct, since it makes no mention of the invariant function of the moral faculty. We could perhaps maintain, with Gibbard, that "[the] biological function [of normative judgments] is […] to coordinate what is in one person's head with what is in another's" (Gibbard 1990, 110) and that this is *all* it is supposed to do. More likely, however, some possible ways of coordinating what is in one person's head with what is in another's will constitute malformed versions of the moral faculties involved, namely, if they are not conducive to producing successful social states (p. 159). In other words, Normal conditions for a moral judgment are likely to include both 1) that it matches the moral judgments of others in the community and 2) that the state of affairs this judgment is supposed to bring about is a feature of a successful social state. Tak-

ing this further requirement into account, we could try to give the following specification of the descriptive content of *S*'s judgment that φ-ing is right:

> (LGP 2) φ-ing is widely judged right within *S*'s local group, and the state of affairs where everybody φs is a feature of a successful social state.

But problems remain. *First*, note that the second conjunct in LGP 2 uses the same indefinite form ("a successful social state") as Sinclair's proposal discussed on p. 177 above, leading to similar problems. Suppose that φ-ing is indeed judged right within *S*'s group. Suppose further that φ-ing is also a feature of a successful social state, but *not* one that is actually instantiated by *S*'s local group. In the latter, φ-ing (though widely judged right) serves only to disrupt social relations, since it is in conflict with the group's other moral standards. These circumstances are not conducive to the success of *S*'s judgment that φ-ing is right, but LGP 2 nevertheless predicts that it is descriptively correct. This is symptomatic of a larger problem, namely, the fact that real social groups don't necessarily adhere to perfectly consistent moral doctrines. Conformity to one of the groups' moral precepts can contribute to good outcomes in some circumstances, detract from them in others.

Second, suppose that in *S*'s group φ-ing is widely judged right, i.e., the first conjunct is true, but the state of affairs where everybody φs is in fact not a feature of any successful social state, i.e., the second conjunct is false. It would then be descriptively incorrect for *S* to judge that φ-ing is right. But it would also be descriptively incorrect for her to judge that it is right to *abstain* from φ-ing (or to do something else that constitutes an alternative to φ-ing). Now it looks as though *S* is out of options. Whatever judgment she makes apropos of φ-ing, that judgment would be incorrect.

How can we address these issues? I suggested above that a judgment that matches already prevalent moral judgments in the group is well-disposed to successfully perform its function. But that is not the only way it could do so. It could also ensure its success by *changing* the moral convictions of the group to match it. There are various kinds of circumstances under which we might expect a moral judge to be receptive to new moral ideas, such as if the prevailing morality produces less than optimal outcomes for her personally or for the group as a whole, or if new circumstances arise on which the extant morality is ill-equipped to pass univocal judgment. Perhaps, then, the moral faculty is designed to adapt the subject to already prevailing views *or* to the potential for changing those views, and when the prevailing views are deficient in one of the ways described above, the moral faculty will Normally be sensitive to the potential for change towards a successful social state. With this in mind, let us try this third specification of the descriptive content of subject *S*'s judgment that φ-ing is right:

(LGP 3) The state of affairs where everybody φs is a feature of the successful social state that is most accessible to *S*'s local group.

Let me unpack this and explain what I mean by "most accessible." If *S*'s group is already in a successful social state—if it happily cooperates to everyone's benefit, let's say—then that state will typically be the one that is most accessible to the group (they're already in it, after all). Then a pattern of action which constitutes a feature of this state will be correctly judged right, according to LGP 3. Typically, we can assume, such an action will also be one that is widely judged right within the group, thus capturing the spirit of LGP. But it can also conceivably be the case that an action is widely judged right which is *not* a feature of the successful social state the group is actually in, and such an action will *not* be correctly judged right, according to LGP 3. Hence, LGP 3 fixes the first problem with LGP 2.

If the group is *not* in a successful social state—if it is riven by conflicts, or stuck in a suboptimal equilibrium—the *most accessible* such state will be something like the state that the group could most easily attain, were they to pursue such a state in a Normal way. This is mighty abstract, and I do not have any account of the concrete factors that inform the relevant accessibility metric up my sleeve. If an accessible state is one that could be attained by changing people's moral views, we should expect accessibility to be sensitive to such factors as the extent of the revisions of extant moral convictions required to reach the state and how easy it is to convince people to make those revisions. However, regardless of how accessibility is to be cashed out, it suffices for dealing with the second problem of LGP 2 that there *is* an accessibility metric that figures in the explanation of the moral faculty's past persistence. Even if *S*'s group is not currently in a successful social state, she can still correctly judge some actions right as long as they are features of the social state that is most accessible to the group.

It is difficult to evaluate LGP 3 due to its abstract character, but it strikes me as a plausible account of the descriptive content of moral judgments consistent with local group pluralism. Some problems remain, however. For one, we would perhaps like to believe that there is such a thing as righteous rebellion, a morally justified crusade against the moral precepts of one's local group. Can LGP 3 vindicate this hope? If the would-be moral hero's group is not in a successful social state, LGP 3 allows her rebellious attitudes toward conventional morality to be correct. But there are social arrangements that seem to work fairly well for the groups that condone them, while we would nevertheless consider a rebellion against the morality that upholds them justified. Take slavery. Slavery seems to be an institution that can often work fine for the groups that maintain it, at least if we don't count the slaves themselves as members of the group. I can see no immediate theoretical reason why, if the moral faculty is designed to produce merely local accommodations, slaveholding institutions could not qualify as one of the possible

successful social states it is in the business of producing and maintaining. But if this is correct, it seems that a moral hero who seeks to reform her group's slaveholding morality could never be anything but an uncooperative speaker and an incompetent moral judge.

LGP 3 also makes a mystery out of certain features of the phenomenology of cross-group moral discourse. I argued on p. 125 that speaker cooperativeness consists primarily in her efforts to make claims, acceptance of which would produce descriptively correct attitudes in her audience. Now if LGP 3 were correct, this would mean that a cooperative speaker would aspire never to try to convince an addressee to abandon the moral precepts of his group, as long as these contribute to a successful social state for that group. Any such attempt would be uncooperative and would essentially rely on manipulating the addressee's moral psychology to produce an abNormal result. Yet we can easily imagine circumstances where it would seem perfectly fine, even admirable, to argue against the moral precepts of an addressee's group, for instance, if the addressee is one of the slaveholders described above.

Indeed, once we begin to reflect on the reality of cross-group moral discourse, local group pluralism as a whole begins to look like a bad idea. The view, recall, is premised on the idea that the adapted function of a moral judgment is to standardize conduct within the subject's local group. At the same time, I have maintained that moral judgments are type-individuated by their adapted proper function, as per the CoRH. But if the adapted function of my judgment that φ is right is to produce widespread φ-ing within my group, whereas the adapted function of your token of the same judgment is to produce widespread φ-ing within *your* group, we can well wonder why our discursive practices should treat them as tokens of the same attributive judgment type in the first place. I have claimed that an attributive type is what a type of statement is designed to make people share (p. 117), and the type-individuation principle implicit in local group pluralism makes it somewhat mysterious why we would have statements designed for making us share judgments of those types across group lines.

Underlying these issues, I believe, is a deeper one. Plainly speaking, the idea of a local group as a clearly delineated social unit is an analytical fiction. In real life, the borders between communities are fuzzy and fluid, and it is not always clear where one ends and the other begins. People constantly enter into new constellations, creating new demands for coordination in attitude and action. Theories of the evolution of morality that trace its origins to the Pleistocene have the luxury of assuming that there was *one* group especially relevant to understanding the evolutionary development of morality, i.e., the local band, the primary organizational unit of cooperative foraging. But if our Pleistocene ancestors resembled modern hunter-gatherers, they would also have been members of other, larger groups (tribes) within which practical coordination would have been beneficial. And if we are going to be able to construe the moral faculty as continuously functional up until modern

times, with its fluid social arrangements, we have no similar recourse. If we are to make room for pluralism, we need a form of pluralism that accommodates these facts about social life and makes sense of our discursive practices. In the next subsection I try to provide just this.

6.4.2. Flexible Pluralism

The CoRH states, roughly, that moral judgments have the function of standardizing conduct within groups and that it Normally does so by coordinating responses. Now, standardized conduct—such as conformity to the same cooperative norms—can yield benefits regardless of with whom one interacts: a member of a different tribe or culture, or even an extra-terrestrial being. For a moral judgment to yield beneficial interactions with outsiders like these, it obviously doesn't matter whether the judgment matches those of one's local group. What matters is that it matches that of the prospective interaction partner herself. So let us suppose that this is the function of moral judgments: to bring about standardized conduct whenever such standardization is beneficial (in the etiological sense in which I have used this word before), regardless of whom the counterpart is. What should we conclude, in that case, about the descriptive content of moral judgments?

Here, things start to get difficult. As long as we could indulge in the fiction that there was a determinate, discrete group whose conduct the moral faculty was supposed to standardize, it was also fairly easy to assign Normal conditions to moral judgments. To contribute to bringing about successful social states in a given group, a moral judgment should coordinate with the judgments of other members of that group and contribute to a successful social state within it. But if the function of the faculty is to enable standardization of conduct with whoever comes along, there is no determinate group for its judgments to be coordinated with.

On one hand, certain groups will often be more important to coordinate with than certain other groups. The people with whom you interact regularly, who wield influence over many aspects of your everyday life, who can grant or withhold resources or other benefits of a material or emotional kind, who can be expected to stick around—with them, it is more important to achieve cooperation and avoid conflicts than it is with random strangers on the street. But not even these people will always form a coherent group. Moreover, the importance of coordination with different others will be context-sensitive and vary across time.

Perhaps we could say that Normal conditions for a moral judgment are those conditions that enable the judgment to help you achieve a beneficial standardization of conduct with *whoever* it is, at any given time, most important for you to do this. On the basis of this suggestion, we could construct a kind of "flexible" pluralism, one that ties Normal success-conditions for moral judgments, and hence their descriptive content, *not* to the moral opin-

ions of the local group, but to whichever group is at any given time the most important. This would give us something like the following specification of the descriptive content of *S*'s judgment that φ-ing is right:

> (FLEXIBLE PLURALISM) The state of affairs where everybody φs is a feature of the successful social state that is most accessible to the group that is most important for *S*.

This is the last attempt I will make at specifying the descriptive content of a moral judgment. I do not claim that it is the correct one, even allowing for the various abstractions required to formulate it. There are most certainly further complications that I have failed to take into account. But I think that we can at least use FLEXIBLE PLURALISM to explain a number of features of moral discourse that resisted explanation by local group pluralism. For this reason, it is worth closer examination.

What makes a group the most important one for a subject? "Important" should not be understood in its everyday sense (which is normative), but as denoting a complex property that can be roughly defined as follows: a group is more important than another for a subject if there is more potential (etiological) benefit to be had for the subject from coordinating her moral judgments with it, and more potential loss to be suffered from failure to do so. Many factors enter into this: the resources the group controls, the status of the subject within the group, the urgency of the situations that require coordination of judgment, the prospect of future cooperative exchanges, and so on. In general, if a sensitivity to group importance is to be part of an explanation of the moral faculty's evolutionary success, group importance should track the group's ability to yield the subject historically fitness-enhancing outcomes, i.e. "benefits" in chapter 4's sense. This also means that one group could qualify as more important than another because its most accessible successful social state is disposed to benefit the subject more than that of the other.[121]

With these ideas in hand, let us return to the case where a speaker is trying to get an addressee to abandon the moral convictions of his local group. FLEXIBLE PLURALISM allows that such an attempt could be a cooperative use of moral discourse. On FLEXIBLE PLURALISM, the addressee's moral faculty is not designed for coordinating his judgments with his local group specifically (insofar as he belongs to a clearly defined local group to begin with),

[121] Relatedly, it may be that the importance of a group is determined by *how* accessible its most accessible social state is for it. In one sense, perhaps, the group with whom an agent would benefit most from coordinating is all of humanity. But collectively, humanity might seem to be very far from any type of stable cooperation regime or anything else that might plausibly qualify as a successful social state. If this is the case, humanity as a whole could be less important for agents than more immediate, tangible groups, since the promise of such an outcome is too distant to merit the effort that would be necessary to reach it.

but simply for coordinating them with whatever group is currently most important for him. And it could certainly be the case that the speaker is speaking on behalf of a group that is more important for the addressee than his local one. That could be the speaker's local group or some other group to which the speaker belongs. At the extreme, it could be only the speaker herself, who proposes to coordinate with the hearer and thereby form a new group of two, set against the world like Bonnie and Clyde. At least under extreme circumstances and at least in the short run, such a minimal group could very well be the most important one for the addressee.

In those cases where the speaker is attempting to talk the addressee out of his customary moral opinions, then, she is not engaging in a mere fruitless appeal to moral principles that have no authority over him, but neither is she necessarily, as the objectivist would have it, appealing to objectively valid moral principles. At least in Normal cases, where the speaker intends the stabilizing outcome of her speech-act, she is attempting to *forge* a community where none existed before or to *extend* a community to include the addressee as well. She is inviting him into a community of judgment and thereby, potentially, into a community of action as well, one that ideally is capable of benefiting both parties.

We have not yet said anything about *how* a speaker can convince an addressee to revise her moral judgments. As in the case of ought-judgments discussed in chapter 4, if I am to be a cooperative speaker, I should strive to ensure that the judgment-tokens I aspire to put in the addressee's head will be descriptively correct. Conversely, if the addressee has a Normal psychology, he will be sensitive to indications that this is so. Hence, we can expect that in typical cases of moral argument, speakers will advance considerations apt for showing that the token judgment that would constitute the addressee's acceptance of the claim would in fact be descriptively correct.[122] With this principle in mind, we can look at common patterns of moral argumentation and ask whether they are compatible with FLEXIBLE PLURALISM.

In the most common cases of moral dispute, the speaker and listener will both be safely ensconced within the same culture and already, to a large extent, share each other's opinions and the opinions of those around them. Moral disagreement will be over matters of detail, perhaps due to different interpretations of a situation, perhaps because the situation itself is novel and it is unclear how existing moral precepts apply to it. In these cases, there are familiar patterns that moral argumentation typically follows. The disputants try to demonstrate to each other that their respective opinions are consistent with moral views that their interlocutor, and the people around them, can be

[122] Of course, neither interlocutor need necessarily conceive of what they are doing in these theory-laden terms. All that is needed to account for the past persistence of moral discourse is that the dispositions of speakers to make claims and the dispositions of addressees to accept them Normally track, by some psychological mechanism, the descriptive correctness of the judgment constituting the addressee's acceptance.

expected already to hold and share. They might do this by advancing empirical information about the situation under dispute or by investigating the logical relations between the moral judgments themselves. Under FLEXIBLE PLURALISM, we can interpret these familiar argumentative moves as attempts by the speaker to show that the addressee would be correct in accepting her claim by showing that it aligns with the moral convictions of his most important group, which happens to be the culture to which they both belong. If the group is already in a successful social state, coordinating with the prevalent moral judgments that underpin this state will be a recipe for teleological success.

In other cases, perhaps especially the kind of which I spoke of above, when people are disputing across cultural or tribal lines, the speaker might appeal to more general moral principles, such as fairness, justice, humanity, or the common good, and try to show that the position she is defending is that which best conforms to these principles. Now, words like "fairness," "justice," and "the common good" do of course stand in need of philosophical analyses of their own. It might be that calling something "fair" is just another way of morally commending it, motivating an analysis of fairness-judgments parallel to that which I have given for rightness- and wrongness-judgments. But it could also be that fairness admits of a straightforward non-normative analysis, perhaps in terms of the distribution of surplus resources from cooperative endeavors, in which case a moral commitment to fairness would be more like a commitment to the wrongness of abortion, i.e., a moral stand in favor of certain descriptively specifiable patterns of behavior. But even so, a commitment to fairness may be expected to be one that is widely shared not only in the addressee's local group but among people in general. We could then understand a speaker's appeal to it as an attempt to demonstrate for the addressee that his commitment to a countervailing opinion, while it may align him with his local group, misaligns him with a larger fraction of humanity with whom it may be more important for him to coordinate.

But we can also see these appeals in another light. We do not have to suppose that the importance of a group depends only on factors external to the group's morality. The moral precepts of the group may themselves impact its importance for a given subject. If some social state favors one category of people over others, depending on, for instance, their gender, race, or innate abilities, then the group whose moral precepts are conducive to that state could for that reason be more important to a person who belongs to the favored category.

But if this is right, then by symmetry of reasoning, the group in question should be proportionally less important for those who do not belong to the favored category. The more a judgment favors one group over others, the more difficult it will be to ensure community of judgment with those others. If, furthermore, a large group will *ceteris paribus* be more important than a small one, because large groups control more resources and wield more

sanctioning power than small ones, then the descriptive correctness of a moral judgment for a person could also depend on the capacity of this judgment to attract a maximally large number of adherents. The term "fairness" might be our conventional means for indicating this quality in a judgment and the social state it underpins.

So these are some of the methods available to a speaker to convince the addressee to accept her claim. We can interpret them as attempts by the speaker to show that the token judgment that would constitute the addressee's acceptance of the claim would in fact be descriptively correct.

Interestingly, it seems as if the decision to adopt the moral conviction of one's interlocutor might contribute to *making* the interlocutor into a member of one's most important group. By abandoning the precepts of my family, tribe, clan, etc., I will by that very token come into conflict with them: attitudinal conflict, to be sure, but also practical conflict, insofar as I actually act on my new-found convictions in their vicinity. By the very act of changing opinions, I can foreclose access to that group, leaving me reliant on whatever community my interlocutor can provide me with. Through my decision, the latter's group *becomes* my most important group, if it weren't already. And so a moral claim can function as a kind of self-fulfilling prophecy. The judgment made in response to it becomes correct by the act of making it.

A similar dynamic obtains from the speaker's point of view. The capacity of the judgment she expresses to perform its function can only be reinforced by the hearer's acceptance. She is not only forging a community with the addressee, but also adapting the environment to enable her judgment's success. Her moral speech act is thus, in successful cases, doubly stabilized. The linguistic form used to convey the judgment is stabilized, by adapting the addressee's language faculty to the speaker's, and the judgment conveyed is also stabilized, by adapting the addressee's moral faculty to the speaker's. In contrast to descriptive beliefs, moral judgments can therefore contribute to making themselves correct. While some descriptive beliefs may be able to make themselves true under specific circumstances,[123] these are marginal and flukey phenomena. For moral judgment, in contrast, it is a direct consequence of their Normal way of functioning. Just like linguistic devices, they stabilize their environments, producing and maintaining Normal conditions for their own continued functioning.

This parallel between linguistic and moral stabilization is to be expected, given my assumption that moral judgments perform their function by coordinating responses. The parallels between morality, on my account, and language, on Millikan's, are fairly direct in this respect: both are ultimately for producing certain beneficial outcomes (successful social states in one case, successful communication in the other), and both produce this outcome part-

[123] For instance, my belief that I'm about to lose my job makes me depressed, which lowers my productivity, which makes me lose my job.

ly by influencing others so as to stabilize the environment and thus to enable the function's continued performance.

There are limits to these parallels, however. There is no analog on the side of language to the fact that moral convictions can clash. Different language devices are merely alternative tools for attaining the same communicative outcomes, whereas different moral convictions are often tools for attaining alternative, rival outcomes. There is no such thing as contradictory linguistic conventions (although people can harbor contradictory normative convictions about the propriety of certain linguistic conventions), but there are such things as contradictory moral judgments, and when they meet, tangible conflict can result.

The title of this section was "making room for pluralism." When formulating it, I had in mind making *theoretical* room for pluralism, i.e., accounting for the objectivist features of moral thought and discourse on pluralist assumptions. I have tried to indicate how that could be done using the apparatus of discursive non-descriptivism. There is another way the words of the title could be understood: in terms of making *practical* room for pluralism, i.e., accommodating, within one society, the existence of multiple, rival moral convictions and beliefs. But that might be precisely what we cannot do, if moral judgments perform their function by coordinating sanctions. If that is so, Normally functioning moral judgment will nigh-inevitably bring their subjects into conflict with those that harbor contradictory judgments.

The tendency of moral judgments to bring people into conflict also explains some of the objectivist features of moral discourse. If you and I harbor contradictory moral judgments and neither of us is willing to revise our judgments and enter into the others' community of judgment, then we remain at least potential enemies and potential victims of each other's sanctioning regimes. This fact lends emotional and practical urgency to moral disputes and motivates the search for common ground. Hence, the same feature of moral judgment that allows us to make *theoretical* room for pluralism is also one that precludes making *practical* room for pluralism.

6.5. Summary and Conclusion

In this chapter, I have discussed the prospects for a teleosemantically informed objectivism on the basis of the CoRH, and, with the help of discursive non-descriptivism, tried to show how the objectivist features of moral discourse could be accounted for given pluralism. To conclude, I would like to reflect briefly on some of the things I have *not* done.

In chapter 4 I discussed "ought" and the judgments conveyed thereby, and in this chapter I have discussed (the moral) "right" and "wrong" and the judgments conveyed by these terms. In each case, my analysis employed some of the same ideas and the same overall framework: teleosemantics,

discursive non-descriptivism, and a normative psychology influenced by the non-cognitivist tradition. What I have not done is to connect these two accounts to each other, to produce a unified account capable of predicting whether and when a person, for instance, *ought* to do what is *right*. At present, the two accounts hang somewhat disconnected from each other, united by common ideas but not by any serious attempt at theoretical unification. To accomplish the latter, I believe a first step would be to investigate the relation between the "benefit" appealed to in my account of "ought," and the "successful social states" and "most important groups" appealed to in my account of moral judgments. It is plain that being part of a well-organized group can be beneficial for a subject in many ways: the question is whether the relations between the successful social states of groups and the benefits for individuals are the *right ones* for explaining the relations that actually obtain between oughts and morality.

A second thing I have not done is to attempt a detailed analysis of the relationship between person-level norms, on the one hand, and conventions, on the other. As we saw in chapter 2, according to Millikan, the rules that govern linguistic conventions, such as semantic rules, are first and foremost teleological norms, i.e., Normal conditions for the successful functioning of language devices. Other kinds of (non-linguistic) conventional rules could also be understood as teleological norms or as part of conventional patterns of action that have themselves persisted due to their ability to incentivize people to go on engaging in them (Millikan 2005a).

Conventions, Millikan claims, can persist and proliferate without the support of sanctions (e.g. Millikan 2008, 89). This is plausible, especially if conventions include learned behaviors that do not yield any substantial benefits, but are merely reproduced due to people's tendency to imitate each other. Still, most conventions *are* supported by sanctions of various degrees of consistency and severity. It would be a terminological stretch to call every instance of an attitude that motivates sanctions of unconventional behavior a *moral* attitude, but the relatively slight annoyance we can feel at reading a misspelled word, or the mild shame we can feel at having ourselves produced it, are at least distant cousins of moral indignation and guilt. It is not implausible that these attitudes, just like full-fledged moral ones, have proliferated due to their ability to help standardize behavior in social groups. As I argued on p. 182 above, other conventional violations merit the full force of moral condemnation.

If morality is for producing cooperation, and cooperation frequently requires conformity to conventions, then it is to be expected that conventions should often be moralized, especially those that people cannot be expected to abide by merely out of self-interest (such as pure coordinating conventions that solve coordination problems with no element of bargaining) and that produce benefits that are of a more general social nature rather than only befalling the agent herself and her immediate partners. Moral judgments,

conversely, by motivating conformity with them and punishment of nonconformists, can influence which conventions persist. It is therefore plausible that conventional patterns of behavior and moral judgments have mutually reinforced one another and coevolved.

On the other hand, conventions can themselves be subject to moral criticism, and frequently, compliance with or violation of conventional norms will be a morally neutral matter. A further research question is therefore whether the account I have sketched can account for the conditions under which conventions are and are not correctly moralized. What does the account predict about the circumstances under which a moral judgment in favor of conformity with a given convention will constitute a Normal exercise of the moral faculty, and do those predictions accord with our intuitions about the moral value of conventions of various kinds?

A complicating factor is that the words "right," "wrong," and related vocabulary can be used in connection with merely conventional rules as well as with moral norms. Many such uses carry a weak normative force, if any. That is to say, they do not seem to be associated with any strong motivational component on the part of the speaker to see the prescribed actions performed or to punish the hearer if he violates the prescription. In contrast to the sociopath's motivationally disengaged moral judgments (cf. n. 102, p. 151), these uses don't seem deficient, either.

In terms of the imperative speech-act categories discussed on pp. 127 f., I posit, such uses can be compared to pieces of friendly advice. They are made for the addressee's benefit, should he ever have to rely on the convention in question, and their weak normative force is due to the speaker's intention that the hearer will comply with them out of regard for his own self-interest. Moral uses, on the other hand, are more like commands in that they are backed up by the force of sanctions. Of course, *unlike* commands, they aren't backed up by the speaker's *own* authority or power, but by the distributed sanctioning power of the whole community (or a significant portion of it). At the same time, they are somewhat like requests, in that they often appeal to the speaker's sense of responsibility towards the larger group. If terms like "right" and "wrong" have directive force, it is not so surprising, perhaps, that they should lend themselves to performing a range of speech-acts analogous to those performed by imperatives.

Epilogue

In the preceding pages, I have tried to show how the conceptual apparatus of Millikanian teleosemantics interacts with some traditional problems in meta-ethics. No doubt there is much more that could be said. As is usually the case in philosophy, every attempted solution seems to open up new problems, nested inside each other like Russian dolls. On many subjects I have merely indicated possible ways forward, and for large swathes of meta-ethical territory, such as the deep and fascinating issues concerning the relationship between morality, personhood, and self-identity (emphasized by Korsgaard 1996 among others) and related issues concerning the relation between morality, agency, and free will, I have stayed completely silent.

What is the upshot, then? I have named this work *Communities of Judgment*, reflecting my belief that this concept can serve as a focus for its core ideas. To recapitulate: in order for our normative judgments to perform their functions, they must secure the assistance of others. To accomplish this, the subject must ensure that her judgment is descriptively correct not only for herself, but also for sufficiently many, sufficiently important others. By adapting to these others, she can bring them into community of judgment with herself. Whereas normative claims strictly speaking lack descriptive content, their capacity, or lack thereof, to produce such communities of judgment provide an intersubjectively valid standard of assessment that approximates the theoretical role of truth-conditions.

The expressivist tradition has attempted to understand the meaning of moral sentences in terms of the attitudes we use them to *express*, but this, I believe, is starting in the wrong end. The take-home message of Millikan's theory of discourse is that to understand the intentional properties of a sentence, we must in the first instance look at what it is supposed to *accomplish*, i.e., how it is supposed to affect the audience. If the function of a claim is to bring the addressee into community of judgment with the speaker, then the standards by which that claim should be judged, including teleological standards as well as agent-level norms, derive primarily from the addressee. We might consider putting this in the form of a slogan: A claim, to be correct, has to be adapted to the addressee. But this should then be qualified by the observation, made at the end of chapter 6, that a claim can also be a way of adapting an addressee to it. In this push and pull between addressee and speaker, we glean something of the drama of human morality.

The importance of the consumer in the analysis of intentionality is one of the Millikanian themes that have played a significant role in my arguments. Another is the possibility, even ubiquity, of hybrid representations. Here there may be cause for worry, even if the reader grants me the conclusion of chapter 3, that universal hybridity is not as bad as it initially appears. Many of the attempts I have made to specify descriptive content for directive attitudes have resulted in vague formulations, laden with potential indeterminacy and terminological promissory notes (such as "important" and "accessible").

In a way, I think this reflects the nature of the subject-matter. What is right or what we ought to do is hardly epistemically transparent to (most of) us, and it would be surprising if the teleosemantic tools I have been employing had supplied clean, elegant correctness-conditions for normative judgments. At the same time, this makes the proposals hard to evaluate—doubly so, since I have repeatedly had to rely on empirical speculation.

This worry is aggravated by the fact, which I have repeatedly insisted upon, that a specification of an attitude's descriptive content does not transparently reveal its inferential role. Hence, the philosopher is divested of one of her main tools for evaluating semantic proposals: conceptual analysis, testing of semantic intuitions, and so on. I have claimed (p. 100) that these types of intuitions can be a *fallible* guide to the contents of attitudes and hence serve as a fallible test of candidate content specifications. This, the reader will note, is a method I myself have employed repeatedly above. However, I haven't provided any details about exactly under what conditions they can and cannot be expected to fail, which would be necessary to yield a principled methodology. To provide a more systematic account of the relation between content, inference, and disagreement is therefore high on the list of priorities for further development of my proposals.

I would not wish my proposals to appear designed to be impervious to criticism, so let me finish by outlining two other ways in which they could be criticized (even assuming the overall evolutionary/teleosemantic framework I have relied upon). For one, since much of what I have said above is intertwined with empirical assumptions (especially the coordination of responses hypothesis of chapter 4), it is vulnerable to empirical criticism. Secondly, even if a teleosemantic theory of moral thought and talk, by its nature, will be a theory not of how we *actually* use moral thought and talk but how we Normally use it, we must assume that there is at least some fallible inference to be drawn from Normal to actual use, or it would be difficult to explain why people go on making moral judgments and statements. This inference can be exploited to criticize my proposal, by drawing out its implications for how we should be expected to use moral thought and talk to see if those implications match reality.

Bibliography

Altham, J. E. J. 1987. “The Legacy of Emotivism.” In *Fact, Science and Morality: Essays on A. J. Ayer’s Language, Truth and Logic*, edited by Graham Macdonald, 275–88. Oxford: Blackwell.

Anscombe, G. E. M. 1958. “Modern Moral Philosophy.” *Philosophy* 33 (124): 1–19.

———. 1985. *Intention*. 2. ed., [Nachdr.]. Oxford: Blackwell.

Artiga, Marc. 2014. “Teleosemantics and Pushmi-Pullyu Representations.” *Erkenntnis* 79 (S3): 545–66.

———. 2015. “Rescuing Tracking Theories of Morality.” *Philosophical Studies* 172 (12): 3357–74.

Austin, John L. 2009. *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955*. 2. ed., [Repr.]. Cambridge, Mass: Harvard Univ. Press.

Ayer, Alfred Jules. 1952. *Language, Truth and Logic*. Unabridged and unaltered republ. of the 2. (1946) ed. New York, NY: Dover Publications.

Barker, Stephen J. 2000. “Is Value Content a Component of Conventional Implicature.” *Analysis* 60 (3): 268–79.

Baumard, Nicolas, Jean-Baptiste André, and Dan Sperber. 2013. “A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice.” *Behavioral and Brain Sciences* 36 (01): 59–78.

Benedict, Ruth. 2001. “Anthropology and the Abnormal.” In *Moral Relativism: A Reader*, edited by Paul K. Moser and Thomas L. Carson, 80–89. New York, NY: Oxford University Press.

Bicchieri, Cristina. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.

Björnsson, Gunnar. 2015. “Disagreement, Correctness and the Evidence for Metaethical Absolutism.” In *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau. Oxford: Oxford University Press.

———. 2018. “Strategic Content: Representations of Epistemic Modality in Biosemantics (and Success Semantics).” *Theoria* 84 (3): 259–77.

Björnsson, Gunnar, and Tristram McPherson. 2014. “Moral Attitudes for Non-Cognitivists: Solving the Specification Problem.” *Mind* 123 (489): 1–38.

Blackburn, Simon. 1993. *Essays in Quasi-Realism*. New York: Oxford University Press.

———. 1998a. “Moral Relativism and Moral Objectivity by Gilbert Harman; Judith Jarvis Thomson.” *Philosophy and Phenomenological Research* 58 (1): 195–98.

———. 1998b. *Ruling Passions: A Theory of Practical Reasoning*. Oxford : Oxford ; New York: Clarendon Press ; Oxford University Press.

Boehm, Christopher. 2008. “Purposive Social Selection and the Evolution of Human Altruism.” *Cross-Cultural Research* 42 (4): 319–52.

———. 2014. “The Moral Consequences of Social Selection.” *Behaviour* 151 (2–3): 167–83.
Boisvert, Daniel R. 2008. “Expressive-Assertivism.” *Pacific Philosophical Quarterly* 89: 169–203.
Boyd, Richard N. 1988. “How to Be a Moral Realist.” In *Essays on Moral Realism*, edited by G. Sayre-McCord, 181–228. Cornell University Press.
Boyd, Robert, and Peter J. Richerson. 1992. “Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizeable Groups.” *Ethology and Sociobiology* 13: 171–95.
Brandom, Robert. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, Mass: Harvard University Press.
Brink, David. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge Studies in Philosophy. Cambridge ; New York: Cambridge University Press.
Broome, John. 2013. *Rationality through Reasoning*. The Blackwell/Brown Lectures in Philosophy 4. Chichester, West Sussex ; Malden, MA: Wiley Blackwell.
Brunnander, Björn. 2011. “On the Theoretical Motivation for Positing Etiological Functions.” *Canadian Journal of Philosophy* 41 (3): 371–90.
Burge, Tyler. 1979. “Individualism and the Mental.” *Midwest Studies in Philosophy* 4 (1): 73–121.
Chomsky, Noam. 1995. “Language and Nature.” *Mind* 104 (413): 1–61.
Churchland, Paul M. 1981. “Eliminative Materialism and Propositional Attitudes:” *Journal of Philosophy* 78 (2): 67–90.
Clark, Andy, and David Chalmers. 1998. “The Extended Mind.” *Analysis* 58 (1): 7–19.
Copp, David. 2001. “Realist-Expressivism: A Neglected Option for Moral Realism.” *Social Philosophy and Policy* 18 (2): 1–43.
Cuneo, Terence. 2018. “Can Expressivism Have It All?” *Philosophical Studies*, November.
Davidson, Donald. 1987. “Knowing One’s Own Mind.” *Proceedings and Adresses of the American Philosophical Association* 60 (3): 441–58.
Davies, Paul Sheldon. 2000. “Malfunctions.” *Biology and Philosophy* 15 (1): 19–38.
Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University Press.
Dennett, Daniel. 1971. “Intentional Systems.” *The Journal of Philosophy* LXVIII (4): 87–106.
———. 1995. *Darwin’s Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon & Schuster.
Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT/Braford.
———. 1986. “Misrepresentation.” In *Belief: Form, Content, and Function*, edited by R. Bogdan, 17–36. Oxford: Oxford University Press.
———. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, Mass: MIT Press.
Dummett, Michael. 1973. *Frege: Philosophy of Language*. London: Duckworth.
Egan, Frances. 2014. “How to Think about Mental Content.” *Philosophical Studies* 170 (1): 115–35.
Eklund, Matti. 2017. *Choosing Normative Concepts*. First edition. Oxford: Oxford University Press.
Fessler, Daniel M. T., and Kevin J. Haley. 2003. “The Strategy of Affect: Emotions

in Human Cooperation." In *Genetic and Cultural Evolution of Cooperation*, edited by Peter Hammerstein, 7–36. Cambridge, MA: MIT Press.
Field, Hartry. 1978. "Mental Representation." *Erkenntnis* 9 (61): 9–61.
Finlay, Stephen. 2014. *Confusion of Tongues: A Theory of Normative Language*. Oxford Moral Theory. Oxford ; New York: Oxford University Press.
FitzPatrick, W. J. 2009. "Recent Work on Ethical Realism." *Analysis* 69 (4): 746–60.
Fodor, Jerry. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass: MIT Press.
———. 1984. "Semantics, Wisconsin Style." *Synthese* 59 (3): 231–50.
———. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Explorations in Cognitive Science 2. Cambridge, Mass: MIT Press.
———. 1990. *A Theory of Content and Other Essays*. Cambridge, Mass: MIT Press.
———. 2008. "Against Darwinism." *Mind & Language* 23 (1): 1–24.
Frank, Robert H. 1988. *Passions within Reason: The Strategic Role of the Emotions*. New York, NY: Norton.
Franzén, Nils. 2018. *Sense and Sensibility: Four Essays on Evaluative Discourse*. Uppsala: Uppsala Universitet.
Frege, Gottlob. 1948. "Sense and Reference." *The Philosophical Review* 57 (3): 209.
———. 1952. *Translations from the Philosophical Writings of Gottlob Frege*. Edited by P. T. Geach and Max Black. Oxford: Blackwell.
———. 1956. "The Thought: A Logical Inquiry." *Mind* 65 (1): 289–311.
Gauthier, David. 1986. *Morals by Agreement*. Oxford : New York: Clarendon Press ; Oxford University Press.
———. 2013. "Why Contractarianism?" In *Ethical Theory: An Anthology*, edited by Russ Shafer-Landau, 571–80. John Wiley & Sons.
Geach, P. T. 1960. "Ascriptivism." *Philosophical Review* 69: 221–25.
———. 1965. "Assertion." *The Philosophical Review* 74 (4): 449.
Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Oxford: Clarendon Press.
———. 2003. *Thinking How to Live*. Cambridge, Mass: Harvard University Press.
———. 2005. "Truth and Correct Belief." *Philosophical Issues* 15: 338–50.
Glüer, K., and A. Wikforss. 2009. "Against Content Normativity." *Mind* 118 (469): 31–70.
Godfrey-Smith, Peter. 1994. "A Continuum of Semantic Optimism." In *Mental Representation: A Reader*, edited by Stephen Stich and Ted A. Warfield, 259–577. Oxford: Blackwell.
———. 1996. *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.
Goodman, Nelson. 1983. *Fact, Fiction, and Forecast*. 4th ed. Cambridge, Mass: Harvard University Press.
Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. "Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism." In *Advances in Experimental Social Psychology*, 47:55–130. Elsevier.
Grice, H. P. 1957. "Meaning." *The Philosophical Review* 66 (3): 377.
———. 1989. "Utterer's Meaning, Sentence-Meaning, and Word-Meaning." In *Studies in the Ways of Words*, 117–37. Cambridge, MA: Harvard University Press.
Griffiths, Paul. 2019. "Beyond Concepts: Unicepts, Language, and Natural

Information." *Australasian Journal of Philosophy*, June, 1–4.
Haidt, Jonathan. 1995. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review*, no. 108: 814–34.
Haidt, Jonathan, and Craig Joseph. 2004. "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues." *Daedalus*, no. Fall: 55–66.
Hare, Richard Mervyn. 1952. *The Language of Morals*. Clarendon Paperbacks. Oxford: Clarendon Press.
Harman, Gilbert. 1975. "Moral Relativism Defended." *The Philosophical Review* 84 (1): 3–22.
Harman, Gilbert, and Judith Jarvis Thomson. 1996. *Moral Relativism and Moral Objectivity*. Great Debates in Philosophy. Cambridge, Mass., USA: Blackwell.
Harms, William. 2000. "Adaptation and Moral Realism." *Biology and Philosophy* 15: 699–712.
Hirsch, Eli. 1976. "Physical Identity." *Philosophical Review* 85 (3): 357–89.
Hirshleifer, David, and Eric Rasmusen. 1989. "Cooperation in a Repeated Prisoners' Dilemma with Ostracism." *Journal of Economic Behavior & Organization* 12 (1): 87–106.
Horgan, Terry, and Mark Timmons. 2007. "New Wave Moral Realism Meets Moral Twin Earth." In *Foundations of Ethics*, edited by Russ Shafer-Landau and Terrence Cuneo, 495–504. Malden, MA: Blackwell.
———. 2009. "Analytical Moral Functionalism Meets Moral Twin Earth." In *Minds, Ethics, and Conditionals: Themes from the Philosophy of Frank Jackson*, edited by Ian Ravenscroft, 221–37. Oxford : Oxford ; New York: Clarendon Press ; Oxford University Press.
Horwich, Paul. 1998. *Truth*. 2nd ed. Oxford : New York: Clarendon Press ; Oxford University Press.
Jackendoff, Ray, and Steven Pinker. 2005. "The Nature of the Language Faculty and Its Implications for Evolution of Language (Reply to Fitch, Hauser, and Chomsky)." *Cognition* 97 (2): 211–25.
Jacob, Pierre. 2019. "Intentionality." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2019 Edition. https://plato.stanford.edu/archives/spr2019/entries/intentionality/.
Joyce, Richard. 2007. *The Evolution of Morality*. 1. MIT Press paperback ed. Life and Mind. Cambridge, Mass.: The MIT Press.
Kitcher, Philip. 2007. *Biology and Ethics*. Oxford University Press.
Korsgaard, Christine M. 1996. *The Sources of Normativity*. Edited by Onora O'Neill. Cambridge ; New York: Cambridge University Press.
Kripke, Saul. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
———. 1984. *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Cambridge, Mass: Harvard Univ. Press.
Lenman, James. 2000. "Consequentialism and Cluelessness." *Philosophy & Public Affairs* 29 (4): 342–70.
Lewis, David. 1974. "Radical Interpretation." *Synthese* 27 (3–4): 331–44.
———. 1981. "Index, Context, and Content." In *Philosophy and Grammar: Papers on the Occasion of the Quincentennial of Uppsala University*, edited by Stig Kanger and Sven Ōhman, 79–100. Dordrecht: Springer Netherlands.

———. 1994. “The Reduction of Mind.” In *Companion to the Philosophy of Mind*, edited by Samuel Guttenplan, 412–31. Blackwell.
———. 2011. *Convention: A Philosophical Study*. Nachdr. Oxford: Blackwell.
MacFarlane, John. 2014. *Assessment Sensitivity: Relative Truth and Its Applications*. First edition. Context and Content. Oxford: Oxford University Press.
Machery, Edouard, and Ron Mallon. 2010. “Evolution of Moralitys.” In *The Moral Psychology Handbook*, edited by John M. Doris and The Moral Psychology Research Group, 3–45. Oxford: Oxford University Press.
MacIntyre, Alasdair C. 1984. *After Virtue: A Study in Moral Theory*. 2nd ed. Notre Dame, Ind: University of Notre Dame Press.
Mackie, John L. 1990. *Ethics: Inventing Right and Wrong*. Reprinted. Penguin Philosophy. London: Penguin Books.
Mameli, Matteo. 2013. “Meat Made Us Moral: A Hypothesis on the Nature and Evolution of Moral Judgment.” *Biology & Philosophy* 28 (6): 903–31.
Martin, Justin, Jillian Jordan, David G. Rand, and Fiery Cushman. 2017. “When Do We Punish People Who Don’t?” *SSRN Electronic Journal*.
Martínez, Manolo. 2013. “Teleosemantics and Indeterminacy.” *Dialectica* 67 (4): 427–53.
Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval. 2003. “Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism.” *The American Economic Review* 93 (1): 366–80.
Merli, David. 2007. “Expressivism and the Limits of Moral Disagreement.” *The Journal of Ethics* 12 (1): 25–55.
Miller, Alexander. 2013. *Contemporary Metaethics: An Introduction*. Second edtion. Cambridge, UK ; Malden, MA: Polity Press.
Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. A Bradford Book. Cambridge, Mass.: MIT Press.
———. 1986. “Thoughts without Laws; Cognitive Science without Content.” *The Philosophical Review* XCV (1): 47–80.
———. 1989a. “Biosemantics.” *The Journal of Philosophy* 86 (6): 281–97.
———. 1989b. “In Defense of Proper Functions.” *Philosophy of Science* 56 (2): 288–302.
———. 1990. “Truth Rules, Hoverflies, and the Kripke-Wittgenstein Paradox.” *The Philosophical Review* 99 (3): 323–53.
———. 1993a. “Speaking up for Darwin.” In *Meaning in Mind: Fodor and His Critics*, edited by Barry Loewer, First publ. in paperback, 151–64. Philosophers and Their Critics 3. Oxford: Blackwell.
———. 1993b. *White Queen Psychology and Other Essays for Alice*. Cambridge, Mass: MIT Press.
———. 1998. “A Common Structure for Concepts of Individuals, Stuffs, and Real Kinds: More Mama, More Milk, and More Mouse.” *Behavioral and Brain Sciences* 21 (1): 55–65.
———. 2000. *On Clear and Confused Ideas: An Essay about Substance Concepts*. Cambridge Studies in Philosophy. Cambridge [England] ; New York: Cambridge University Press.
———. 2004. *Varieties of Meaning: The 2002 Jean Nicod Lectures*. The Jean Nicod Lectures. Cambridge, Mass: MIT Press.
———. 2005a. “Language Conventions Made Simple.” In *Language: A Biological*

Model, 1–23. Oxford: Oxford University Press.
———. 2005b. "On Meaning, Meaning, and Meaning." In *Language: A Biological Model*, 53–76. Oxford: Oxford University Press.
———. 2005c. "Proper Function and Convention in Speech-Acts." In *Language: A Biological Model*, 139–65. Oxford: Oxford University Press.
———. 2005d. "Pushmi-Pullyu Representations." In *Language: A Biological Model*, 166–86. Oxford: Oxford University Press.
———. 2005e. "The Language-Thought Partnership: A Bird's Eye View." In *Language: A Biological Model*, 92–105. Oxford: Oxford University Press.
———. 2005f. "The Son and the Daughter: On Sellars, Brandom, and Millikan." In *Language: A Biological Model*, 77–91. Oxford: Oxford University Press.
———. 2007. "An Input Condition for Teleosemantics? Reply to Shea (and Godfrey-Smith)." *Philosophy and Phenomenological Research* 75 (2): 436–55.
———. 2008. "A Difference of Some Consequence Between Conventions and Rules." *Topoi* 27 (1–2): 87–99.
———. 2009. "Biosemantics." In *The Oxford Handbook of Philosophy of Mind*, edited by Brian P. McLaughlin, Ansgar Beckermann, and Sven Walter, 394–406. Oxford Handbooks in Philosophy. Oxford : New York: Claredon Press ; Oxford University Press.
———. 2010. "On Knowing the Meaning; With a Coda on Swampman." *Mind* 119 (473): 43–81.
———. 2017. *Beyond Concepts: Unicepts, Language, and Natural Information*. First edition. Oxford, United Kingdom: Oxford University Press.
———. 2018. "Biosemantics and Words That Don't Represent." *Theoria* 84 (3): 229–41.
Moody-Adams, Michele. 2001. "The Empirical Underdetermination of Descriptive Cultural Relativism." In *Moral Relativism: A Reader*, edited by Paul K. Moser and Thomas L. Carson, 93–106. New York, NY: Oxford University Press.
Moore, George Edward. 1903. *Principia Ehtica*. Cambridge: Cambridge University Press.
Mumm, John. 2015. "Two Functions of Moral Language." In *Motivational Internalism*, edited by Gunnar Björnsson, Caj Strandberg, Ragnar Francén Olinder, John Eriksson, and Fredrik Björklund. Oxford: Oxford University Press.
Nanay, Bence. 2014. "Teleosemantics without Etiology." *Philosophy of Science* 81 (5): 798–810.
Neander, Karen. 1995. "Misrepresenting & Malfunctioning." *Philosophical Studies* 79: 109–41.
———. 2013. "Towards an Informational Teleosemantics." In *Millikan and Her Criticis*, edited by Dan Ryder, Justine Kingsbury, and Kenneth Williford, 21–36. Malden, MA: John Wiley & Sons.
———. 2017. *A Mark of the Mental: In Defense of Informational Teleosemantics*. Life and Mind: Philosophical Issues in Biology and Psychology. Cambridge, Massachusetts London, England: The MIT Press.
Nichols, Shaun. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford ; New York: Oxford University Press.
Nietzsche, Friedrich. 2000. "On the Genealogy of Morals." In *Basic Writings of*

Nietzsche, edited by Walter Kauffman, 449–559. New York: Modern Library.
Nolan, Daniel. 2019. "Hyperintensional Metaphysics." *Philosophical Studies* 171 (1): 149–60.
Oliver, Pamela. 1980. "Rewards and Punishments as Selective Incentives for Collective Action: Theoretical Investigations." *American Journal of Sociology* 85 (6): 1356–75.
Olson, Jonas. 2014. *Moral Error Theory: History, Critique, Defence*. First edition. Oxford: Oxford University Press.
O'Neill, Onora. 2015. *Constructing Authorities: Reason, Politics, and Interpretation in Kant's Philosophy*. New York: Cambridge University Press.
Packalén, Sara. 2016. "Content and Composition: An Essay on Tense, Content and Semantic Value."
Pagin, Peter. 2016. "Problems with Norms of Assertion." *Philosophy and Phenomenological Research* 93 (1): 178–207.
Pagin, Peter, and Dag Westerståhl. 2010. "Compositionality I: Definitions and Variants." *Philosophy Compass* 5 (3): 250–64.
Papineau, David. 1984. "Representation and Explanation." *Philosophy of Science* 51 (4): 550–72.
———. 1993. *Philosophical Naturalism*. Oxford, UK ; Cambridge, Mass., USA: B. Blackwell.
———. 1998. "Teleosemantics and Indeterminacy." *Australasian Journal of Philosophy* 76 (1): 1–14.
———. 2001. "The Status of Teleosemantics, or How to Stop Worrying about Swampman." *Australasian Journal of Philosophy*, 279-89, 79 (February).
Price, Carolyn. 2001. *Functions in Mind: A Theory of Intentional Content*. Oxford : New York: Clarendon Press ; Oxford University Press.
Prinz, Jesse J. 2009. *The Emotional Construction of Morals*. Oxford: Oxford Univ. Press.
Putnam, Hilary. 1975. "The Meaning of 'Meaning.'" *Minnesota Studies in the Philosophy of Science* 7: 131–93.
Quine, W. V. 1956. "Quantifiers and Propositional Attitudes." *The Journal of Philosophy* 53 (5): 177.
Railton, Peter. 1986. "Moral Realism." *The Philosophical Review* 95 (2): 163–207.
Rawls, John. 1980. "Kantian Constructivism in Moral Theory." *The Journal of Philosophy* 77 (9): 515.
Recanati, François. 2012. *Mental Files*. 1st ed. Oxford: Oxford University Press.
Ridge, Michael. 2006. "Ecumenical Expressivism: Finessing Frege." *Ethics* 116 (2): 302–36.
Roberts, Debbie. 2013. "Thick Concepts." *Philosophy Compass* 8 (8): 677–88.
Roojen, Mark van. 2018. "Moral Cognitivism vs. Non-Cognitivism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018 Edition. https://plato.stanford.edu/archives/fall2018/entries/moral-cognitivism/.
Rosen, Gideon. 2010. "Metaphysical Dependence: Grounding and Reduction." In *Modality: Metaphysics, Logic, and Epistemology*, edited by Bob Hale and Aviv Hoffmann, 109–35. Oxford ; New York: Oxford University Press.
Rowlands, Mark. 2010. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. Cambridge, Mass: MIT Press.

Ruse, Michael, and Edward O. Wilson. 1986. “Moral Philosophy as Applied Science.” *Philosophy* 61 (236): 173–92.
Russell, Bertrand. 1903. *The Principles of Mathematics*. London: Norton.
Ryder, Dan. 2009. “Problems of Representation II: Naturalizing Content.” In *The Routledge Companion to the Philosophy of Psychology*, edited by Francisco Garzon and John Symons, 251–79. London: Routledge.
Scanlon, Thomas. 2000. *What We Owe to Each Other*. Nachdr. Cambridge, Mass.: Belknap Press of Harvard Univ. Press.
Schelling, Thomas C. 1960. *The Strategy of Conflict*. New York: Oxford University Press.
Schroeder, Mark Andrew. 2009. “Hybrid Expressivism: Virtues and Vices.” *Ethics* 119 (2): 257–309.
———. 2010. *Being for: Evaluating the Semantic Program of Expressivism*. Oxford ; New York: Clarendon Press.
Schulte, Peter. 2017. “Perceiving the World Outside: How to Solve the Distality Problem for Informational Teleosemantics.” *The Philosophical Quarterly* 68 (271).
———. 2019. “Naturalizing the Content of Desire.” *Philosophical Studies* 176 (1): 161–74.
Searle, John R. 1962. “Meaning and Speech Acts.” *The Philosophical Review* 71 (4): 423.
———. 2011. *Speech Acts: An Essay in the Philosophy of Language*. 34th. print. Cambridge: Univ. Press.
Sellars, Wilfrid. 1997. *Empiricism and the Philosophy of Mind*. Cambridge, Mass: Harvard University Press.
Shea, Nicholas. 2007. “Consumers Need Information: Supplementing Teleosemantics with an Input Condition.” *Philosophy and Phenomenological Research* LXXV (2): 404–35.
———. 2018. *Representation in Cognitive Science*. First edition. Oxford, United Kingdom: Oxford University Press.
Sinclair, Neil. 2012. “Metaethics, Teleosemantics and the Function of Moral Judgements.” *Biology & Philosophy* 27 (5): 639–62.
Smith, Michael. 1995. *The Moral Problem*. Philosophical Theory. Oxford, UK ; Cambridge, Mass., USA: Blackwell.
Smyth, Nicholas. 2017. “The Function of Morality.” *Philosophical Studies* 174 (5): 1127–44.
Stampe, Dennis W. 1977. “Toward a Causal Theory of Linguistic Representation.” *Midwest Studies in Philosophy* 2 (1): 42–63.
Stevenson, Charles L. 1944. *Ethics and Language*. New Haven: Yale University Press.
Stich, Stephen. 1983. *From Folk Psychology to Cognitive Science: The Case against Belief*. Cambridge, Mass: MIT Press.
———. 1993. *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*. 1. MIT Press paperback ed. A Bradford Book. Cambridge, Mass.: MIT Press.
Street, Sharon. 2006. “A Darwinian Dilemma for Realist Theories of Value.” *Philosophical Studies* 127 (1): 109–66.
Tersman, Folke. 2006. *Moral Disagreement*. Cambridge Studies in Philosophy. Cambridge ; New York: Cambridge University Press.

Tomasello, Michael. 2015. *A Natural History of Human Morality*. Cambridge, Massachusetts: Harvard University Press.
Turiel, Elliot, Melanie Killen, and Charles C. Helwig. 1987. "Morality: Its Structure, Functions, and Vagaries." In *The Emergence of Morality in Young Children*, edited by Jerome Kagan and Sharon Lamb, 155–244. Chicago: University of Chicago Press.
Williams, Bernard. 2011. *Ethics and the Limits of Philosophy*. 1. publ. in Routledge Classics. Routledge Classics. London New York: Routledge.
Williamson, Timothy. 2009. *Knowledge and Its Limits*. Repr. Oxford: Oxford Univ. Press.
Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Translated by G.E.M. Anscombe. Oxford: Blackwell.
Wright, Crispin. 1994. *Truth and Objectivity*. 2. print. Cambridge, Mass.: Harvard Univ. Press.
Zeman, Dan. 2017. "Contextualist Answers to the Challenge of Disagreement." *Phenomenology and Mind* 12: 62–73.
Zigon, Jarrett. 2008. *Morality: An Anthropological Perspective*. Oxford: Berg Publishers.