The Normative Will

Practical Judgment as Volitional Interpretation

Copyright © Sander Voerman, 2012

Cover design by Babana Media
Printed by Optima Grafische Communicatie

ISBN 978-94-6169-316-7

The Normative Will

Practical Judgment as Volitional Interpretation

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op maandag 10 december 2012 om 16:15 uur door

Sander Arthur Voerman
geboren op 13 maart 1979 te Rotterdam

Promotores

Prof. dr. A.P. Thomas
Prof. dr. G.C.G.J. van Roermund

Overige leden van de promotiecommissie

Dr. H. Lillehammer
Prof. dr. H.K. Lindahl
Prof. dr. P. Noordhof
Prof. dr. M.V.P. Slors

Work on this thesis was funded by the Netherlands Organization for Scientific Research (NWO).

Acknowledgements

My work on this thesis has benefited greatly from the helpful comments and advice of many teachers, colleagues, and friends. I would like to thank Alan Thomas and Bert van Roermund for their invaluable supervision. When Alan became involved about two years ago, the view I was going to defend had already been worked out in considerable detail—except that what I had in actual writing consisted mostly of sketches, concepts, and too many half-finished papers. Alan quickly familiarized himself with my arguments and taught me much about the intricacy of philosophical alternatives that I thought I could glance over with a few sweeping statements. Then he made me write over three hundred pages in little more than a year.

Bert has been a continuous source of reflection and practical support from the beginning. He helped me set up the project, write a grant proposal, and organize my thoughts. He may disagree with some of the views defended in this thesis, but in a procedural sense he basically masterminded the whole thing.

I also want to thank Herman de Regt. He was my supervisor during the first years of the project, but more importantly, he has been my mentor in philosophy ever since I first attended his lectures. It was in discussion with him that I first started developing the ideas on volitional interpretation that culminated in this thesis.

Special thanks to Marc Slors, who has *de facto* been yet another supervisor with whom I met regularly during the first years to discuss my writings. Special thanks also to Gary Watson for our weekly discussions during my three month visit to the University of California, Riverside. Thanks to Michael Smith for inviting me over to the Humboldt-Universität in Berlin to discuss my criticisms of his philosophy. I also want to thank Hallvard Lillehammer, Paul Noordhof, and Hans Lindahl, for agreeing to be on the committee, reading the thesis, and providing insightful comments.

The importance of discussion with friends and fellow students is hard to

overestimate. The ideas and arguments in this thesis have been scrutinized during many hours of pleasant conversation with Tjeerd van de Laar, Derek Strijbos, Tonnie Staring, Marijke Vonk, and Aukje van Rooden. They sharpened my views, provided useful examples, and always made me experience why I love philosophy so much. I thank Sean Gould for reading and commenting on every chapter I wrote. Thanks also to Katrien Schaubroeck for extensive comments on an earlier version of chapter 7. And thanks to my fellow PhD students at Tilburg University for providing me with feedback from so many different points of view.

Thanks also to the many philosophers in my field who have provided me with useful comments during seminars and personal conversations in conference rooms, restaurants, and the pub. Thanks in particular to Jan Bransen for many stimulating discussions and for letting me steal his ideas. Thanks to Maureen Sie and the participants in the "meta-ethics and methods" seminars for commenting on my chapters. Thanks to Bert Musschenga and the participants in the OZSE moral psychology course I followed. Thanks to Gerrit Glas and the participants in the expert meetings on philosophy and psychiatry. At Tilburg University, several of my chapters were discussed in the seminars of the ethics research group, the legal philosophy section, and the philosophy of mind section. Thanks also to Stephan Hartmann and participants in the philosophy of science seminar of TiLPS.

I am afraid I may not always recall which participants in these various seminars made the comments that led to the most significant improvements, so I would simply like to thank everyone who made the effort to reflect on my work and provide me with feedback. Finally, I thank my partner Nathalie. Her enthusiasm about my work has been, and continues to be, very inspiring.

Contents

Introduction

> Do What Thou Wilt,

wrote the 16th-century humanist François Rabelais,

> because men that are free, well-born, well-bred, and conversant in honest companies, have naturally an instinct and spur that prompteth them unto virtuous actions, and withdraws them from vice, which is called honour. (1542/1653, ch. 57)

He was criticizing an assumption, or way of thinking, that we still often encounter today: that what we want, or would really want, to do is one thing, and what we *should*, or morally *ought*, to do is another thing. The idea that "if everybody would just go around and do whatever they please, there would be anarchy and the world would be a mess," as some people might put it, perhaps with the added remark that in so far as the world already happens to be in such a state, this is exactly because people have been going around selfishly trying to get whatever they wanted to have. On this view, *morality*, by contrast, requires us to look beyond the immanent nature of our desires for something that transcends their contingent character. Proponents of such a view maintain that certain acts are obligatory no matter how much we may dislike them and other acts are wrong no matter how attractive they might seem. In the words of Thomas Nagel, morality does not allow you to "beg off" (1970, p. 4).

But those who hold this view have difficulties in answering certain questions about morality. How do we acquire knowledge about how we should live? What sorts of facts, or truths, might there be for our moral beliefs to get right? And why should we care about such a morality in the first place? How do we explain its rational authority for reasonable agents? In this thesis I will argue that we stand a better chance of answering these questions if, like Rabelais, we adopt instead the view that what we should do is going to be in accordance with what we really want. Indeed (and

perhaps unlike Rabelais) I shall be claiming that what we should and what we want—in a particular sense of willing—are really one and the same thing.

The Idea of a Normative Will

To be sure, some philosophers have also held that moral acts must be willed by the agent even though they did believe that moral truths transcend our contingent nature. Thus, St. Augustine argued that in addition to having knowledge of what is right, we must also have the will to do the right thing, in the absence of which people knowingly do the wrong thing instead. In other words, we have the capacity to *decide to act* either in accord with, or against, our own beliefs about what we should do. In this thesis I will call this capacity an "executive will." But, according to St. Augustine, those beliefs that we must want to act upon are still beliefs about a morality that transcends, and to some extent opposes, our contingent—and in his view sinful—desiderative nature.

Now I agree that our nature does involve many desires to do things that we should not do. Our world is, in certain respects, in a sorry state as a result of people's actions motivated by such desires. But the reasons we have to act differently, I submit, are themselves grounded in natural desires as well: desires that constitute a mode of wanting that is not executive, nor something that we actively decide or control, but rather one that is *opaque*. This is the sense in which we can want something without knowing it yet. It is what we try to figure out when we wonder "what we really want." I call this source of reasons the "normative will."

The main philosophical challenge for my proposal is to explain how certain desires can constitute such a normative mode of wanting if that would generate reasons to disapprove of other desires. How can we even make sense of the idea that some desires are better than others, so to speak, without once again presupposing a morality that would transcend their contingent nature?

Summarily, my solution is based on the idea that when we *govern* our own agency, we act on our understanding of how our different desires and affective experiences are interrelated, rather than merely being driven by separate desires in different situations. Our affective and emotional lives are structured in various ways and we often try to make sense of the motivations we experience in the light of what we experienced before.

Furthermore, we expect ourselves to have certain desires in the future and to have positive or negative responses to certain events should they occur. Consequently, if these expectations turn out to be false, we must adjust our ideas about ourselves, which is how we gain self-knowledge. My claim is that as we gain this knowledge of the ways in which our affective dispositions are structured, *patterns* will manifest themselves in our emotional lives that establish "what we really want" or "what we are really about." The desires that contribute to such a pattern constitute an agent's normative will. By contrast, the desires that the agent should not act upon will now appear as a kind of *noise* in that overall picture of what he, as a person, is about. Thus, some desires are better than others because of how meaningful they are to the person who has them.

Perhaps it may seem that, on this view, everything could be reasonable, or moral, as long as someone happens to exhibit a pattern of desires in its support. Furthermore, if different persons have different desires, their moral beliefs might conflict without either of them being at fault. But we have substantial intuitions about the content of morality, and we are used to moral conversations in which we treat these as having an intersubjective scope of validity. It is true that, conceptually, my analysis of normativity will not imply anything of moral substance. But I will argue that our substantial intuitions about morality do not need to cover any conceptually possible creature capable of deliberative action (for example, they need not apply to fictitious alien invaders from Mars). Instead, intuitions about morality's altruistic implications, for example, are to be understood as beliefs about how human nature has endowed different human beings with similar affective response patterns. Our moral differences may often be better explained by the hypothesis that some of us have poor self-knowledge, than by the hypothesis that our inner selves (so to speak) are radically different.

So rather than thinking of mankind as selfish and sinful by nature, my claim is that the values and virtues we intuitively associate with morality (insofar those intuitions are sound) are supported by the most accurate interpretation of what human nature is first and foremost really all about. Like Rabelais, I believe that under the appropriate conditions—conditions that are conducive to self-understanding and for which his term "freedom" is not at all a bad choice—we will find all our reasons for being social and altruistic within our own passionate and emotional characteristics.

A Debate Inspired by Hume

The picture I just sketched may appear attractive to some and unappealing to others, but the point is of course to argue about it systematically. Why is it so difficult for those, who defend a morality beyond the empirical nature of our desires, to explain how we acquire knowledge of this morality, what sort of facts or truths it involves, and why these facts or truths are normative in a way that non-moral facts or truths are not?

An influential way to frame these difficulties is to adopt a distinction from David Hume between what he called "reason" and the "passions" (1886/1964, p. 193). The former refers to our capacity to form beliefs which can be true or false, whereas the latter is his term for the desires that motivate us towards certain goals. The point of the distinction is that these two aspects seem conceptually independent: merely appreciating that something is true by itself never entails a motivation towards anything. Conversely, any motivation we actually have seems therefore to presuppose a desire towards some end that we might in principle have regardless of what we believe to be true.

From this analysis of the relation between truth and motivation, Hume drew a further conclusion about the status or purpose of reason itself: that it cannot establish any course of action as reasonable in its own right, but only determine what we should do, contingent upon the desires we happen to have. This claim was captured in his famous slogan that "reason is, and ought only to be the slave of the passions." Therefore, Hume also sought to justify moral ideas in terms of characteristics of our human nature, which he called "moral sentiments."

Many philosophers have questioned whether Hume's conclusion about reason follows from his theory of motivation. Michael Smith has construed the issue as involving a paradox: moral knowledge seems to require facts that true beliefs get right; moral authority seems to require rational, self-governing agents to be motivated by those beliefs; but the Humean theory of motivation denies that beliefs can have such motivational impact (Smith, 1994). He tries to resolve this paradox by pointing out that one is often motivated by several conflicting desires, while we should expect a fully rational agent to be entirely coherent in this respect. Hence, rationality seems to require that we reject some of our desires in order to meet this constraint of coherence. In order to resolve the paradox, we could therefore adopt Bernard Williams's "Internal Reasons View," according to

which our practical beliefs are true or false in virtue of what we would desire after having deliberated rationally upon our actual desires (Williams, 1980/1981a).

This notion of deliberation brings us back at the aforementioned question of how to judge that some desires are 'better' than others. I have explained that my proposal is an attempt to answer this question without once again presupposing truths that would go beyond the empirical nature of our desires. But those who take their cue from Immanuel Kant may wish to reason in the opposite direction: if, by reflecting upon the ways in which the desires of rational, self-governing agents must be constrained, we run into assumptions about a principle or principles of judgment beyond our empirical nature, then so much the better. That way we might legitimately arrive at something *a priori* by analyzing what is simply presupposed about rationality in the very idea of having desires and being an agent in the first place (Korsgaard, 1996).

If such a strategy could succeed, then we might nevertheless wonder whether the particular concepts of rationality, agency, or self-government, in terms of which such an *a priori* principle would have to be understood, could still be distinctively *human*. More precisely, we might wonder whether highly intelligent alien lifeforms could conceivably be at fault for not acting in accordance with that principle, provided that their concepts would be suitably different from ours or unintelligible from our point of view. On such a theory, even *a priori* moral judgments might still be local to the human perspective. By contrast, Michael Smith argues that in order for some moral judgment to be justified, every conceptually possible agent whose action or judgment would go against it would therefore have to be mistaken. In his view, an *a priori* judgment that a certain act is right under a certain type of circumstances is made true or false by some fact about those circumstances, and not by a perspective towards those circumstances that may be different depending on who is making the judgment (Smith, 1994, 2004a).

Bernard Williams, however, was not convinced that a transcendental analysis of whatever *a priori* presuppositions his Internal Reasons View might have would yield anything of moral substance: it might clarify certain formal principles of deliberation and planning, for example, but it would not justify altruistic behavior. Instead, he maintained that the altruistic motivations that we would have, after rational deliberation, are still related to the particular characteristics of the motivations from which

we start those deliberations. If our actual motivations had been different, then the conclusions of our deliberations might be different as well. Williams called this "relativism," but I will call it "relationalism," because "relativism" is commonly associated with certain further implications that this view need not have, as I will argue in this thesis. Different versions of relationalism have been defended by Harman (2000), Frankfurt (2004), and Street (2009, forthcoming), amongst others. Conversely, I will call views according to which morality does not depend on motivations that could have been different "nonrelationalist." This includes Smith's account of *a priori* moral facts about reasons that I shall discuss in detail.

The Significance of Disconfirmation

The debate between these relationalist and nonrelationalist views about reasons has reached a stalemate. Philosophers on each side start from their own intuitions about what it means for reasons to be normative and presuppose their own metaphysical assumptions about what sort of facts about reasons they claim can or cannot exist. Each side has their own problems to deal with, problems which I shall discuss at length in this thesis. Ultimately, I must conclude that neither side has managed to come up with a fully convincing answer to that fundamental question of how to explain the authority of one motivation over another when the two are in conflict.

To be sure, fully convincing answers are rare in philosophy, and I do not wish to suggest that my proposals will settle this issue decisively. But I will try to offer a way out of the deadlock, and one that I think might tip the scales in favor of relationalism. Now, it is common in meta-ethics to start out with questions or intuitions about moral properties, facts, truths, or truth-conditions: are they natural or not, how can they be part of the fabric of the world, could they be *a priori*, must they be the same for all agents, and so on—exactly the kind of questions I have been discussing above. But my own line of argument will be based on the idea that there is another question, one that might appear to be derived from these matters, but which I am going to treat as prior to them, and that is: how do we *disconfirm* our practical judgments in everyday life?

After all, one of the reasons why many of us think that morality must be a matter of truth and knowledge, perhaps even the most important reason, is that our practical judgments do not merely change over time in

the manner that some of our tastes or preferences are susceptible to change. We do not acquire a taste for free speech in the sense in which one might acquire a taste for drinking tea after having preferred coffee for many years. Nor do we think that over time, more and more people have come to reject racism or sexism in the same way that sweaters with shoulder pads have gone out of fashion. Instead, we take these changes in our practical views to have been *corrective*, we think our views have *improved*. Racism and sexism have been disconfirmed, and this seems to presuppose a background norm governing such a process of disconfirmation, namely, truth.

So if it is because of our experiences of disconfirmation in practice that we believe in truths, or facts, about reasons for action, then why don't we try to come up with an account of how disconfirmation of practical judgments actually works in real life in order to figure out what sort of facts about reasons there might be? And this is exactly what I will set out to do.

Philosophers have held views about this, of course, and the most widely held account seems to be that we can disconfirm practical beliefs by testing whether they violate certain principles of reason, such as Kant's categorical imperative. But I will argue that this does not explain how practical disconfirmation happens in the real world. Instead, I will propose the theory that we disconfirm our practical judgments in the light of unexpected affective responses: when we are surprised by our own emotional reactions to the consequences that we did intend our actions to have.

Not only does this rule out the nonrelationalist views of an *a priori* morality, which are committed to disconfirmation by rational principles, but it also shows us where certain relationalist approaches may have fallen short. It turns out that the results of deliberation are not just rational transformations of a conflicting desire set into a coherent one. Instead, we make empirical discoveries about our own affective dispositions, which means that our affective nature is significantly *opaque*. But traditionally, relationalist accounts, whether under the banner of "relativism" or "Humeanism," have been understood in terms of attitudes that are, at least for the most part, readily accessible to the agent as the premises of his deliberations. By contrast, on the theory of the normative will that I am proposing, we deliberate upon hypotheses about dispositions that we learn more about as we go along in order to figure out what patterns there are for us to act upon.

Structure of this Thesis

This thesis is divided into three parts. In part I, "Normativity and Motivation: Setting the Stage," I construct a conceptual framework in order to articulate more precisely how the aforementioned ideas or intuitions about practical normativity give rise to two philosophical problems: the "Facts Problem," which captures the issue concerning truth and truth-conditions, and the "Disconfirmation Problem," which focuses on the matter of disconfirming practical judgments. In part II, "Facts about Reasons: The Status Quo," I turn to the stalemate between relationalist and nonrelationalist solutions to the Facts Problem, discussing the difficulties into which each approach runs. Then, in part III, "Practical Disconfirmation: A New Perspective," we make the shift towards treating the Disconfirmation Problem as the prior issue. I will propose a solution on relatively independent grounds and then argue that this proposal subsequently leads us to a relationalist solution to the Facts Problem as well.

Part I consists of three chapters. Chapter 1, "Five Principles of Practical Normativity," introduces and motivates my own preferred formulations of the intuitions for which I want my theory to account. The "Facts Principle" and "Disconfirmation Principle" capture the truth-conditional and disconfirmational aspects of a cognitivist and non-skeptical approach to practical judgment. In addition, the "Intersubjectivity Principle" allows that certain reasons have an intersubjective scope of validity. The "Authority Principle" establishes a conceptual connection between practical judgment and motivation, which makes it a version of motivational internalism, but one that differs from standard accounts in that it defines the connection in terms of self-government rather than practical rationality. Finally, the "Distinctness Principle" contains my formulation of the Humean belief-desire distinction. The combination of these principles gives rise to the Facts and Disconfirmation Problems mentioned above. The purpose of the rest of the thesis is to solve these problems.

In chapter 2, "The Internal Reasons View," I examine the famous distinction from Williams between "internal" reason statements, which must be reachable from the agent's own motivations by a "sound deliberative route," and "external" reason statements that do not have this requirement. I discuss possible objections against his defense of the claim that only the former type of statements can be true. I explicate the premises that we need in order to block these objections and make Williams's argument

valid.

Then, in chapter 3, "Internal Reasons, Relationalism, and Motivation," I am going to relate his framework for talking about reasons to my own framework from chapter 1. I argue that my Principles imply a version of the Internal Reasons View, even though Williams himself might not have accepted the Distinctness Principle and may have favored a motivationally anti-Humean defense instead (his writings were rather ambiguous in this respect, as it turns out). I also explain how the Internal Reasons View may lead to skepticism about the plausibility of nonrelationalist convergence, especially if the view is defended using the Distinctness Principle.

Part II also contains three chapters. In chapter 4, "Outline of a Relationalist Solution," I explain how my defense of the Internal Reasons View provides us with a sketch of a dispositional solution to the Facts Problem. The idea is that we remove the mystery about why we would be motivated, under ideal conditions, in accordance with our normative reasons, by *analyzing* normative reasons in terms of the motivations that we would have under those conditions. The relationalist version of this approach, which I call "type-I dispositionalism," allows us to disambiguate the paradox resulting from the Facts, Authority, and Distinction Principles into two different and compatible implications. However, this is only an outline of a solution because it assumes that a deliberator can resolve conflicts between her own desires. The problem remains of explaining why that makes sense. Furthermore, nonrelationalists complain that no relationalist account can do justice to the intersubjective validity of moral considerations. Hence, the second problem for type-I dispositionalism is to explain the Intersubjectivity Principle.

In chapter 5, "The Nonrelationalist Alternative," I discuss Michael Smith's nonrelationalist solution, according to which all conceptually possible agents would desire the same states of affairs under ideal conditions of rational agency. I distinguish a "proceduralist" and a "nonproceduralist" interpretation of this view, which I call "type-II" and "type-III dispositionalism." I argue that critics of Smith such as David Sobel and David Enoch seem to have presupposed the type-II interpretation, according to which any incoherence in our prior attitudes should in principle be demonstrable to us from the perspective of those attitudes themselves: we cannot be "ineliminably unlucky," as I will call it, in the attitudes we start deliberating with. Sobel and Enoch make a strong case that the *a priori* convergence that Smith's realism requires is highly implausible on those assumptions. I

will discuss some strategies that Smith might wish to employ to counter this criticism, but my conclusion is that these strategies fail.

This leaves the type-III solution, to which I turn in chapter 6, "Dispositionalism Without Proceduralism?" According to this account, it is possible for an agent to be ineliminably unlucky in his desires, such that no amount of computational power could make him understand why his desire set is unreasonable. Thus, certain desire configurations may seem *a priori* unreasonable to us, even though we may believe or even know that we cannot prove this to the skeptic who would uphold such desires, and the type-III solution allows us to postulate that what he desires is nevertheless undesirable as a conceptual fact. I offer several objections to this view: such postulates seem too ad-hoc in the field of moral philosophy to warrant their theoretical expense; the presupposed facts about desirability are inconsistent with the direction of fit that desires have; the view cannot offer a real advantage in explaining intersubjectivity over type-I dispositionalism, which avoids the expensive metaphysics; and finally, the view seems to allow the same type of proceduralist self-knowledge that type-I dispositionalism allows, which turns the nonprocedural residue into a form of speculation without practical import.

Part III contains the final four chapters of the thesis, in which I propose and defend my account of the normative will. In chapter 7, "The Affective Response View," I develop the idea that we disconfirm our practical judgments in the light of unexpected affective responses to the intended consequences of our actions, rather than by testing those judgments against *a priori* principles of reason. I argue that my view leads to an attractive picture of deliberation as the "volitional interpretation" of his affects and emotions by an agent in order to determine his normative will. I also argue that this view implies type-I dispositionalism. And finally, I explain how the Affective Response View solves the Disconfirmation Problem.

Chapter 8, "A Normative Reality Within Ourselves," is a critical evaluation of Harry Frankfurt's (2004; 2006) work on practical normativity. Frankfurt claims that there is a "reality within ourselves," consisting of empirical facts about what we love, that provides us with normative reasons about which we can sometimes be mistaken. He argues that love and caring are *volitional* attitudes which involve a more complex mode of wanting than the affective attitudes of mere desire, and he attempts to answer the question of how to authorize some desires over others with reference to how they relate in such complex volitional structures.

I agree with him on this general picture, but I will criticize the specific account of the nature of the volitional that Frankfurt presents. The problem is that even though he accounts for some types of mistakes, there are many cases of disconfirmation that his framework cannot accommodate, because he relies on special attitudes that grant us privileged access to our inner selves under the appropriate conditions. Furthermore, he maintains that these attitudes are neither reducible to beliefs or desires, which makes both their empirical status and their motivational role mysterious.

Based on this critique, I make recommendations for an alternative theory: it should distinguish between "cognitive" and "normative" volitional attitudes, such that the former can be analyzed as involving beliefs, while the latter may be identified as structures of affective dispositions. Finally, at the end of the chapter, I try to sort out my verbal disagreements with Frankfurt concerning some of the things he says about "morality" and "values" and which may seem to be at odds with what he says about reasons.

In chapter 9, "The Nature of the Normative Will," I propose an account of volitional reality that follows the aforementioned recommendations. According to the "Affective Pattern View" the relevant dispositional structures are *patterns* in our affective lives, which manifest themselves as we increase our self-understanding. I borrow the idea of an ontology of patterns from Daniel Dennett (1991b), but whereas he has used it to explain *desires* as *behavioral* patterns, I am explaining *volitional* attitudes as *affective* patterns, which requires me to 'customize' the notion extensively.

The result is an account that allows me to explain intuitions about both the determinacy and the indeterminacy of moral choice. On the one hand, I will argue that my view can account for the idea that most Nazi officers in the Holocaust (perhaps even all of them that weren't psychopaths) got their practical judgments wrong, without postulating nonrelationalist moral facts. On the other hand, my view will accommodate genuine moral dilemmas, which leave the right choice indeterminate even at the 'intrapersonal' level. Furthermore, my account also combines intuitions about deliberation and decision-making as 'self-finding' and as 'self-making.' I shall argue that while cognitive attitudes can get normative attitudes wrong in some cases, there are other cases in which the latter may be *shaped* by the former, such that our deliberations also play a constitutive role in the genesis of our normative reasons.

With this account in place, I will reflect further on some of Harry

Frankfurt's key claims —in particular, on his ideas about *wholeheartedness*. I distinguish between "inner wholeheartedness" and "epistemic resolvedness," arguing that neither are to be pursued too fervently. Instead, I will claim that allowing ambivalence in our hearts may actually be a form of authenticity, of being true to the divided nature of our selves. Furthermore, I will discuss the potential harm of eradicated doubt in the light of our often heavily biased emotional mechanisms. Finally, at the end of the chapter I summarize how the Affective Pattern View has allowed us to explain why some desires can have authority over others, which now provides us with a type-1 dispositional solution to the Facts Problem.

One difficulty remains, however, which is the intersubjective dimension that relationalist theories are allegedly poorly placed to explain. We turn to this matter in chapter 10, "Intersubjectivity and Moral Discourse." I will discuss two objections against relationalism, both derived from arguments by Michael Smith. The first is that moral discourse would be without purpose if relationalism were true. I will meet this argument by showing, first, that we have many good reasons to assume shared volitional attitudes on many, if not most, occasions. In particular, it is plausible that we share certain basic moral values as a species. Second, I argue that there are several further reasons for engaging in moral discourse even in those cases where our values may turn out to differ.

The second objection is that the words or concepts used by people in moral discourse, such as "right" and "wrong" or "good" and "bad," simply have nonrelationalist meanings, regardless of whether purposeful discourse about relationalist concepts would be possible or not. Rather than simply denying this outright, I will admit that some people may indeed mean their judgments in a nonrelationalist sense. However, other people do not: I propose a semantic pluralism about what people actually mean, and a conceptual revisionism about what people *should* mean when they make their practical judgments.

This concludes my main line of argument, since I will then have addressed both of the difficulties for type-1 dispositionalism which I had formulated at the end of chapter 4: the authority of some desires over others and the intersubjective relevance of practical judgments in moral discourse.

1 Normativity and Motivation : Setting the Stage

1 *Five Principles of Practical Normativity*

It is part of the human condition that life causes us to ask the normative question of how we should live it. We wonder whether we should support a particular charity, what we should do about an unwanted pregnancy, and whether that interesting job offer is worth moving to another country. Let us call such questions "practical questions," and the answers that we give to such questions "practical judgments." In this chapter I will discuss a number of intuitions about practical judgments and formulate principles that are meant to capture these intuitions. Although each principle may seem plausible when considered independently, certain problems arise when we try to combine them. The purpose of the rest of this thesis is to solve those problems.

Before we turn to the first principle, let me say a few preliminary things about what "practical" means in this context. Very roughly speaking, practical questions have two characteristic features. First of all, they are *normative*: they are not about what we *shall*, but about what we *should* do. And second, of course, they are primarily concerned with action: they are about what we should *do*. Thus, practical questions are to be distinguished from other normative questions, such as questions in epistemology about the rationality of belief, questions in statistics about how correctly to draw conclusions from experimental results in various fields of science, or questions of spelling and grammar about which words and sentences conform to the rules or standards of a given language.

Notice, however, that the reference to "action" is not entirely helpful in this respect. Applying statistical rules in order to derive correct conclusions is, after all, also a form of action. And so is the use of language. Furthermore, when I talk about practical judgments, my concern is not only with judgments about concrete actions that the agent in question is capable of performing. We also make practical judgments about political ideals or

states of affairs that we do not have the power immediately to bring about, or perhaps even to influence at all. Rather, the point of practical judgments is to give a kind of bottom-line evaluation of approval or disapproval, to express whether the agent is in favor of something or against it.

There are many normative terms that we might use to represent the content of practical judgments, such as "should," "ought," "must," "right" and "wrong," "good" and "bad," "approve" and "disapprove," and so on. It seems that of these terms, "should" and "approve" have the least philosophical commitments built into them, which is why I prefer to use them when my purpose is to make distinctions between different philosophical claims.

I shall often use "should" when I want to represent practical judgments as judgments about actions on the part of the judger, as in "A judges that she should ϕ." When I want to represent practical judgments as judgments about general ideals or states of affairs, I will often use "approve" or "disapprove," as in "A approves of P." However, as far as I am concerned, these are not essentially different types of judgments, and their terminology is interchangeable. Thus, if we want to represent a practical judgment of approval of a state of affairs P as a judgment on the part of the agent that he should ϕ, then we may think of ϕ as something like "supporting P" or "contributing to P." In cases where the agent can do nothing at all with respect to P we can simply equate ϕ with "approving of P." Conversely, if we want to represent the judgment of an agent A that he should ϕ as a judgment of approval of some state of affairs P, then we can think of P as the proposition "that A does ϕ." With these terminological conventions in place, let us now turn to the principles that I want to discuss.

1.1 The Facts Principle

An important intuition about practical judgments is that we are concerned to *get them right* (Smith, 1994, p. 5; Frankfurt, 2006, p. 2, 27). When we are undecided about a practical question, we cannot simply resolve what it is that we should do by fiat. Practical normativity is not "up to us" in this sense (Frankfurt, 2006, p. 34). We can accommodate this intuition if we subscribe to *cognitivism*, the view that practical judgments express or establish *beliefs*: if a person judges that he should ϕ, then he believes that he should ϕ. Let us call such beliefs "practical beliefs." Many cognitivists also accept the stronger claim that there are facts which make certain practical

beliefs *true*. We might think of this stronger claim as a form of *realism* about the content of practical judgments, but as we shall see in section 1.3.2, there are certain ambiguities in how moral philosophers have used the term "realism" which make it undesirable to use that label for the more limited claim I have just presented. In order to avoid any possible confusion, and because the claim is going to be one of the central principles that I want to account for in this thesis, I shall simply call it the "Facts Principle."

> Facts Principle. If a person judges that he should ϕ, then he believes that he should ϕ; and there are facts in virtue of which such beliefs may be true.

Cognitivism offers an explication of the intuition that we are concerned to get our practical judgments right: we are concerned to adopt true practical beliefs. The Facts Principle allows us to be non-skeptical about that concern: we *can* get our practical judgments right because there are facts about what we should do. However, it also raises a difficult question: what sort of facts are facts about what we should do? What sort of facts make it true that I should keep my promise, or that John should help the woman that just fell downstairs? Let us call this the "Facts Question."

I am aware that some philosophers have expressed unease about the idea that there are such things as facts which make statements true.[1] However, my use of phrases like "made true by facts" or "true in virtue of facts" is intended in a very noncommittal sense. A fact, in this sense, is simply *what a true belief gets right*. This sense of there being a fact is meant to convey the idea that some things are "matters of fact" whereas others are not, and that it is a meaningful philosophical question to ask which things belong to the former category and which to the latter. Thus,

[1] E.g. Davidson (1974/2001b, p. 194): "Nothing, however, no *thing*, makes sentences and theories true: not experience, not surface irritations, not the world, can make a sentence true. *That* experience takes a certain course, that our skin is warmed or punctured, that the universe is finite, these facts, if we like to talk that way, make sentences and theories true. But this point is put better without mention of facts. The sentence 'my skin is warm' is true if and only if my skin is warm. Here there is no reference to a fact, a world, an experience, or a piece of evidence." Davidson's unease with fact-talk stemmed from two related arguments. According to the infamous "slingshot argument," facts cannot be individuated into separate truth makers for different truths; hence, there is at best really just one "Great Fact" (Davidson, 1967/2001d, p. 19; 1969/2001c, p. 42; for discussion and response, see Neale, 2001). His second, more general line of argument involved the idea that our notion of facts, or of the Great Fact, does not *explain* the notion of truth and should instead be treated as explanatorily useless with respect to truth (Davidson, 1969/2001c, p. 43–44).

two philosophers might agree that there is a "fact of the matter" about whether Iron Maiden sold more albums than Judas Priest, while only one of them thinks that there is also a fact of the matter about which of the two is the greatest heavy metal band. The idea that there are facts, in the noncommittal sense that I have in mind, is the idea that it makes sense to wonder whether there is a truth about the greatest band that is ultimately determined in the same way as the truth about who sold the most albums. And when some people think that a certain matter is a matter of fact, while others think it is not, then it makes sense for the latter to raise a question for the former: to explain to them how it is that this matter belongs to the matters of fact. In order to answer this question, one must explain 'what there is to get right' in this matter, or, more conveniently formulated: *what sort of facts* we are talking about. Thus, the Facts Question is the question how the matter of whether I should keep my promise can be a matter of fact like the number of albums sold by Judas Priest, and unlike, perhaps, their being the greatest band or not. In the case of the number of albums sold, there seems to be less mystery about what there is to get right. But in the case of practical judgment, things are not so obvious.[2]

This idea—that we should be able to explain how certain controversial matters could be matters of fact like other not so controversial matters of fact—must be distinguished from the various metaphysical views that philosophers have proposed in order to *account* for this idea. For example, we can make a meta-ethical distinction between the view that ethical matters are matters of fact and the view that they aren't, while remaining neutral about the question whether matters of fact involve a *correspondence* between the structure of propositions in thought or language and some structure of facts in a stronger sense. Therefore, many 'pragmatist' or 'inferentialist' philosophers who reject the idea of facts in this stronger sense may still accept the noncommittal notion of facts alluded to in the

[2]Those who are so puzzled by institutional facts as to become anti-realist about the number of sold albums may substitute even less controversial matters of fact, such as whether my arm was cut off with an axe yesterday—it was not. In any case, there may be some indeterminacies about what counts as an 'album' that bear on the number of albums sold, and so on, but suppose I inadvertently mix up the number of albums sold by Maiden with that of Priest when I compile a chart of popular heavy metal bands, say, then it will be obvious to everyone what I got wrong. Instead, even though I think the Nazis got their practical judgments wrong, I find it a puzzling philosophical endeavor to articulate what it is that I think they were getting wrong.

Facts Principle.[3] For they might still want to be able to say that there is a fact of the matter about whether I should pay my taxes, while there is no fact of the matter about whether Maiden is greater than Priest.

Because many such philosophers have been afraid to talk of facts, I might have avoided the term myself in the interest of preventing unnecessary confusion. However, I consider myself to be a friend of facts: the word is part of ordinary language, and so are phrases like "facts of the matter." The notion of truth-making is, admittedly, a philosophical invention, but its function is merely to explicate the relation between the concept of truth and the common-sense intuition that some issues are matters of fact. In other words, it seems to me that the reason why we have the word "fact" in our language is precisely to be able to articulate the kind of intuition that motivates the Facts Principle. Therefore, it would be counter-intuitive not to make use of it. So to those who worry about inflated metaphysical claims I say: please allow me to use the term as my appeal to facts is as noncommittal as possible.

There is only one exception—one philosophical view on the basis of which even my noncommittal notion of facts would have to be rejected. Some philosophers take the 'deflationist' approach towards the truth predicate to the extreme and reject all philosophical questions about what matters of fact there could be as not well posed. This is the view known as *quietism*. If quietists are right, then the Facts Principle would have to be rejected, and the problems that I shall be trying to solve in this thesis would simply dissolve immediately. Hence, another way to understand my use of the notion of facts is as involving the view that philosophical problems will not go away in the way that quietists think that they will.

Nevertheless, there are other alternatives to the Facts Principle, of course. *Error theorists* agree that practical judgments express beliefs, and they also agree that true beliefs are made true by facts (at least in the minimal sense I have explicated), but they claim that all practical beliefs

[3]This also applies to Davidson's "Great Fact" (see footnote 1 above): even though I speak of facts in the plural, and of classes of facts such as "institutional facts" or "moral facts," my appeal to fact-talk is merely a way of demarcating the matters of fact, and to say that institutional matters can be matters of fact may be understood by proponents of the slingshot argument as a way of saying that such matters are included in the Great Fact. Furthermore, this way of speaking does not require our understanding of something as a matter of fact to be prior to our notion of truth in any substantial or explanatory sense, *as long as* we do not trivialize the notion of truth itself in the quietist manner, explicated below, that would prevent us from wondering about which matters belong to the matters of fact.

are false because the facts that would make them true do not exist. In contrast, *non-cognitivists* simply deny that practical judgments express beliefs in the first place. Although the Facts Principle seems to give the most straightforward justification of our concern to get practical judgments right, I shall not be arguing that quietist, error theoretic, or noncognitivist alternatives cannot accommodate this concern. I will simply treat the Facts Principle as a premise for my arguments in this thesis.

1.2 The Disconfirmation Principle

There is another principle that we might want to accept if we are to account for the concern to get practical judgments right. Note that it would not be a matter of great concern if we did not worry that we might also get them *wrong*. What we should do is not transparent to us, which makes us fallible in our attempts to adopt true practical beliefs. This idea also reflects another intuition about practical judgment: that some of the changes in our practical views are *corrections*. Consider the Montgomery bus drivers who, before the Boycott of 1955, forced African-American passengers to give up their seats on the bus for white passengers. Suppose that one bus driver, who used to judge that it was right for him to enforce this policy, extensively revised his views later on in his life, to the point where he would forcefully advocate racial equality and the abolition of any such policies. It seems plausible that such a revision involves more than a change in preference. Instead, we might want to be able to say that the bus driver *discovered* that his practical beliefs were *false*. Which yields the following principle:

> Disconfirmation Principle. Practical beliefs can be false. If someone falsely believes that she should ϕ, then under appropriate circumstances, she may discover that her approval of ϕ is unjustified and be rationally required to reject it.

But how did the bus driver discover his mistake? That is the question which follows from this second principle. How do we *disconfirm* practical beliefs? What sort of consideration makes it rational for an agent to reject her belief that she should ϕ? I shall call this the "Disconfirmation Question."

1.2.1 *Proceduralism vs. Nonproceduralism*

A satisfactory theory of practical normativity must be able to answer this question for every type of mistake in our practical beliefs that the theory allows. However, we can distinguish between a weaker and a stronger version of this requirement, depending on how we understand the phrase "appropriate circumstances" as it is used in the Disconfirmation Principle. According to a strong version, these circumstances are exhausted by concerns that lie in principle within the reach of inquiry. Roughly, that means (1) that the agent is reasoning correctly and (2) that he has access to the relevant empirical information. Of course, when we consider all the practical beliefs of an agent together, no real agent may ever meet these requirements. In contrast, if we would focus on a particular practical belief that an agent might have, then perhaps it may be possible for him to have access to all the empirical information relevant to that specific belief, and to have reasoned flawlessly with respect to that particular belief, even if he is just an imperfect fallible human being, living in the actual world. To be sure, disconfirmation is often possible even when we do not meet these requirements. The strong understanding merely claims that these requirements would be sufficient in order to disconfirm a false practical belief, not that they are necessary, so it can remain neutral about whether these requirements are even possible to fulfill in actual cases. In other words, the point of the strong interpretation is not that we may sometimes be fully informed or fully coherent, but rather that there is *nothing else* that might prevent us from discovering our mistakes other than empirical ignorance or flaws in our reasoning.

In contrast, according to the weaker understanding, the appropriate circumstances for disconfirmation may require a further ingredient: that the agent started out with the right prior beliefs. Of course, unless some of his prior beliefs were false he would not be able to disconfirm any of them in the first place, but the idea is that there are two types of false beliefs: those that could be corrected by the process of inquiry, and those that the agent could maintain coherently regardless of any empirical information that he might receive. Therefore, the appropriate circumstances for disconfirmation might require that the agent did not start out with false beliefs of the second type.

Let us call answers to the Disconfirmation Question that support the strong understanding "procedural," and answers that only support the weaker understanding "nonprocedural." Thus, according to procedural

views, all mistakes could in principle be uncovered by the process of inquiry. However, a few disclaimers about this terminology must be made. First of all, being procedural in this sense does not involve the claim that we shall actually be able to uncover all our mistakes in this way. In fact, it does not even require that it would be physically possible to uncover our mistakes in any finite amount of time. Second, it does not mean that rational inquiry can be reduced to the following of "procedures" in a fully formalized, methodological sense. It does not mean, in other words, that we can discover our mistakes in a countable number of formal steps. Third, the distinction between procedural and nonprocedural views of practical disconfirmation should not be conflated with the distinction between procedural justice (in the Rawlsian sense) and substantive justice. The former distinction is in meta-ethics, the latter in normative political theory, and they are orthogonal to each other.

Which of the two approaches is most plausible? On the one hand, we might wonder whether nonprocedural views really account for the intuition behind the Disconfirmation Principle. To be sure, they firmly uphold the idea that practical beliefs may be false, but this idea is no longer supported by the concept of disconfirmation in all cases of falsehood. Thus, if I think that your practical belief that you should kill the traitor is false, and you ask me why, then on the nonprocedural view, even if I am right then I might not be able to explain to you, in principle, why you shouldn't kill the traitor. I might end up having to simply claim, dogmatically, that you just started out with the wrong prior beliefs. The question is whether that really gives an account of our idea that we can get our practical judgments wrong.

On the other hand, the challenge for procedural views is whether they can come up with an answer to the Disconfirmation Question from which it would follow that people are getting it wrong in all the cases in which it seems intuitively correct to say that they are getting it wrong. Think of the hard-headed religious fundamentalist who simply sticks to his "moral" principles and concludes that women who had pre-marital sex should be put to death. Most of us would want to be able to say that he is getting it wrong. But can the proceduralist explain this solely in terms of logical coherence and empirical fact? I will return to these questions in parts II and III of this thesis.

1.3 The Intersubjectivity Principle

Practical questions are the subject of debate between different persons. When I wonder whether I should ϕ, I may want to discuss the reasons for and against ϕ with others. In such a situation, it is common to ask something like "what would you do?" And when another person has a convincing argument why he should ϕ under those circumstances, then this may lead me to believe that I, too, should ϕ. Furthermore, once I do believe that I should ϕ, I may feel criticized when yet another person explains why she disapproves of ϕ, and perhaps I will then be forced to reconsider my reasons for and against ϕ.

This also applies to our discussions of moral principles, political ideals, and questions about whether to approve or disapprove of certain states of affairs. In general, when A approves of P while B disapproves of P, then arguments that count in favor of A's practical judgment may also count against the practical judgment of B. On the assumption that practical judgments are subject to justification and disconfirmation, we can formulate this idea as follows:

> Intersubjectivity Principle. The same considerations which justify A's approval of P may, under the appropriate conditions, disconfirm B's disapproval of P and require B to judge in approval of P instead.

1.3.1 Relationalism vs. Nonrelationalism

The most straightforward way to account for this principle is to adopt the view that if A and B both approve of P, their practical judgments have the *same content*. On this view, the content of both judgments is something like "it should be the case that P." This may be contrasted with the view that the content of each judgment contains a reference to the agent making the judgment. According to this second view, the content of A's judgment might be represented as "according to what is normative for me, agent A, it should be the case that P," or "it is normative for A that P" or perhaps simply "it should$_A$ be the case that P." Let us call the latter view "relationalism" and the former "nonrelationalism." Thus, according to relationalism, for A to make a practical judgment about P is for A to make a judgment about a *relation* between A and P, whereas according to nonrelationalism, it is about P itself and need not involve any relation

to A (it *might* still involve various relations to A in certain cases, but only insofar P is itself related to A, for example when P is the proposition that A is selling his house).

I am aware that this terminology is unusual. Some readers may feel that "relativism" and "nonrelativism" (or perhaps "absolutism") would be the more common terms to use for these views. It is true that "relativism" is an often used term in moral philosophy, whereas "relationalism" is not.[4] And there are prominent philosophers who have used the term "relativism" in more or less the same way that I am using "relationalism" here.[5] However, there are also moral philosophers who have used the term "relativism" in a different sense. In particular, many philosophers associate "relativism" with the view that different moralities are normative for different human beings, which does not follow from relationalism as I have defined it above. In order to be able to distinguish between these two views, I shall make use of the term "relationalism."

With this terminology in place, it should be clear that the relationalist will need to come up with some story about why arguments that count in favor of "it should$_A$ be the case that P" would also count against "it shouldn't$_B$ be the case that P," since the two judgments seem logically independent. Instead, on the nonrelationalist view, B's judgment in disapproval of P would simply be the logical negation of A's judgment in approval of P.

1.3.2 *Relationalist Cognitivism vs. Nonrelationalist Realism*

If we combine nonrelationalism with the Facts Principle, we get the view that A and B are expressing a belief in the same proposition when both judge in approval of P, and that there may be a fact (or collection of facts) that makes this proposition true. On this view, practical beliefs are no different from any other beliefs about matters of external fact: my belief that innocent animals should not be tortured is made true by the same fact (or collection of facts) as yours. In contrast, if we combine relationalism with the Fact Principle, we must conclude that if A and B are both judging in approval of P, they are expressing beliefs in different propositions—that P should$_A$ be the case and that it should$_B$ be the case.

[4]Wim de Muijnck has used the term "relationism" for the general metaphysical view that relational properties are more fundamental than intrinsic properties (2003, pp. 12–13). That claim is logically independent from what I am calling "relationalism" here.

[5]Michael Smith, for example (1994, p. 164; 2000/2004e, p. 204).

In order to be able to represent practical beliefs in a manner that is neutral with respect to the disagreement between relationalism and nonrelationalism, I will sometimes use the format "practical belief in approval of *P*." When two agents have practical beliefs in approval of the same state of affairs, I will say that their practical beliefs are "similar." Thus, on the nonrelationalist reading of the Fact Principle, similar practical beliefs are the same beliefs, whereas on the relationalist reading, they are not. Furthermore, when *A* holds a practical belief in approval of *P*, while *B* holds a practical belief in disapproval of *P*, I shall say that their practical beliefs are "dissimilar." According to the nonrelationalist reading, dissimilar practical beliefs are logically contradictory beliefs, whereas on the relationalist reading, they are not.

A popular doctrine holds that the nonrelationalist reading of the Facts Principle must at least be true with respect to *moral* questions. Or in other words, that there are *moral facts* on the basis of which all similar moral beliefs have the same truth values, regardless of the agents holding those beliefs. Many philosophers know this view as "moral realism," but because some might also think of the Facts Principle itself as a form of realism regardless of whether it should be understood in the relationalist or nonrelationalist sense,[6] I will call this view "nonrelationalist moral realism."

In contrast, I shall call the view that accepts the Facts Principle on the relationalist reading in all cases, including the moral cases, "relationalist cognitivism." Calling the latter view "cognitivism" instead of "realism" introduces an asymmetry in our terminology for the two readings of the Facts Principle, which may not be very elegant, but it has the advantage that it avoids the aforementioned association that many philosophers make between "realism" and nonrelationalism. Furthermore, even though "cognitivism" is logically weaker than the Facts Principle, in practice the only views which accept cognitivism while rejecting the Facts Principle (i.e., quietism and error theory) are always motivated by the wish to accommodate a nonrelationalist understanding of practical belief. Therefore, in the relationalist camp, proponents of the Facts Principle can be called "cognitivists" for practical purposes.

To summarize, we have now formulated two views, *nonrelationalist moral realism* and *relationalist cognitivism*, which both claim that there are facts which make certain practical beliefs true, but while the former view

[6]See for example Smith (2000/2004e, pp. 204–206).

holds that these have to be the same facts for all agents in moral cases, the latter allows that these may be different facts for different agents because the truth conditions of all practical beliefs involve a relation between the object of the judgment and the agent making the judgment.

It may not be entirely clear which practical questions qualify as moral and which do not, but let us say that typically, when persons have dissimilar practical beliefs about some *moral* question, this tends to lead to *conflicts* in practice.[7] For example, if an environmentalist believes that he should preserve the forest and wildlife in some area, and a project developer believes that she should start a project to build a shopping mall in that same area, then the environmentalist and the project developer have a conflict, which makes the practical question of whether to build the shopping mall a *moral* question. Nonrelationalist moral realists draw our attention to the fact that in such cases of conflict, we do not merely discuss what sort of reasons we might have individually in order to justify our beliefs. Instead, we tend to speak in more general terms, and debate what is *right*, or *good*, and what is *wrong*, or *bad*, or *evil*. Thus, the environmentalist does not merely want to defend why *he* should oppose the shopping mall. A nonrelationalist moral realist would say that the environmentalist wants to argue that it is *right* to preserve the forest and that it would be *wrong* to build the shopping mall and that therefore, the project developer should not approve of the project *either* and must be mistaken in her practical beliefs.

Instead, if you think you should paint your living room green, while I think you should paint it white, then intuitively, there seems to be less of a conflict, because it's your house and in the end you can paint it pink for all I care. It would be odd to exclaim that painting your room green is wrong or evil (even though some of us may have had the experience that certain people decorate their homes in ways that ought to be forbidden).

We may even wonder whether the issue between relationalism and nonrelationalism can really arise in nonmoral cases. Take the last example. There actually is a subtle ambiguity in the statement that I think you should paint your living room white. I can make two different judgments. I can judge, first of all, that *I* would paint it white if it were *my* house. But I can also judge that *you* should paint it white because I know that *you* will like it better, once you're finished, than if you would paint it green. The

[7] Here I am using the term "moral" in a broad sense that includes *political* questions as well as moral questions that arise in private life.

second judgment does not require the first at all: in fact, I might judge that you should paint it white even though I know that I myself would paint it *black* if it were my house. But given this distinction, there is nothing for the relationalist and the nonrelationalist to disagree about anymore. In the case of the second judgment, the relationalist can happily concede that the judgments do contradict each other: they are simply different judgments about what color you would like best. Thus, they are really just instrumental judgments about how you could reach a goal that you and I are not in disagreement about (i.e. that you should pick the color you would like best once it's on your walls). On the other hand, in the case of the first judgment (that I should paint the room white if it were mine) the nonrelationalist can happily concede to the relationalist that it does not actually contradict your judgment (that you should paint it green). After all, your judgment is about what you would like, whereas my first judgment is about what I would like, so the judgments are not about the same proposition in the first place: they are not expressing dissimilar practical beliefs.

To develop this point further: let us suppose that I actually believe that I should paint the room green if it were my house, because I happen to believe that the green color that you have picked matches my taste. In such a case, I will say that our practical beliefs are "isomorphic." In general, if A believes that he (A) should ϕ, and B believes that *he* (B) should ϕ, then A and B have isomorphic practical beliefs. Even nonrelationalist moral realists who subscribe to the universalizability principle (that is, that the same agents should act the same under the same circumstances) can agree that isomorphic practical beliefs can have different truth conditions, by considering the tastes of the respective agents as part of the circumstances. Intuitively, differences in taste seem to make circumstances relevantly different in non-moral cases, but leave circumstances relevantly similar in moral cases.

With these remarks in mind, it seems to me that in order to give a nonrelationalist account of the Intersubjectivity Principle, one really only needs to cover the moral cases. Of course, that does not mean that there is no intersubjectivity in nonmoral cases. It just means that intersubjectivity in the nonmoral cases can be explained in a fairly trivial way with reference to similarities and differences in our tastes.[8] Instead of "nonrelationalist

[8]This does not mean that we are not going to be interested in nonmoral cases, because accounting for the Facts and Disconfirmation principles in nonmoral cases is a lot less trivial!

moral realism" I will therefore often speak simply of "nonrelationalist realism."

If nonrelationalist realism is true, then we must be able to answer the Fact Question in such a way that for any person A, the facts which make her practical beliefs about moral questions true or false do not involve any particular features of A that persons in general need not necessarily possess. Furthermore, we would have to be able to answer the Disconfirmation Question in such a way that anything that would disconfirm such a belief of A would in principle also disconfirm a similar belief of any other agent.

The question for nonrelationalist realism is whether it is possible to come up with answers that meet these requirements. And the question for relationalist cognitivists is how answers to the Facts and Disconfirmation questions that do not meet these requirements could possibly take into account the Intersubjectivity Principle. I will return to these questions in the next chapters.

1.4 The Authority Principle

According to yet another intuition about practical normativity, when we make practical judgments we exercise a kind of *authority* over ourselves. We must subscribe, as agents, to our own practical judgments. This intuition is often expressed in terms of *reasons*: there seems to be something very odd about a person who would resolve his doubt about whether to ϕ by making the practical judgment that he should ϕ, but who subsequently would not consider himself to have any reason to act accordingly. Smith gives us the following example:

> Suppose we are sitting together one Sunday afternoon. World Vision is out collecting money for famine relief, so we are waiting to hear a knock on the door. I am wondering whether I should give to this particular appeal. We debate the pros and cons of contributing and, let's suppose, after some discussion, you convince me that I should contribute. There is a knock on the door. What would you expect? I take it that you would expect me to answer the door and give the collector my donation. But suppose I say instead "But wait! I know I *should* give to famine relief. But what I haven't been convinced of is that I *have any reason* to do so!" And let's suppose that I therefore refuse to donate. What would your reaction be? (1994, p. 6)

Smith thinks our reaction would be one of "extreme puzzlement." Judging that you should donate simply *means* that you think you have a reason to donate. Or does it? What does it mean, exactly, to think that you have a reason to do something? It does not imply, to be sure, that one will always be *motivated* to act accordingly. Suppose that instead of being able to wait for the collector to knock on his door, the person in the example would have had to make a booking himself, using internet banking, say. Still a small effort to make, but nevertheless the kind of thing people often fail to 'get around to.' If the person would say "I know I have a reason to do it, but I am a bad and lazy person and I didn't get around to it yet" then we might agree that he is lazy, and perhaps also that that is a bad thing, but we would not be puzzled about what he is *saying*. It is, after all, an unfortunate characteristic of our human nature that we are susceptible to weaknesses that may prevent us from being motivated to act upon our own reasons. The resulting disparity between our judgments and our motivational tendencies can be understood as a limitation on the *freedom* of our agency:

> [H]uman beings are only more or less free agents, typically less. They are free agents only in some respects. With regard to the appetites and passions, it is plain that in some situations the motivational systems of human beings exhibit an independence from their values which is inconsistent with free agency; that is to say, people are sometimes moved by their appetites and passions in conflict with their practical judgments. (Watson, 1975/2004b, pp. 31–32)

In a later article, Gary Watson identifies this mode of freedom as the "power of self-government" (1996/2004c, pp. 260–261).[9] Following up on this terminology, let us say that a person acts upon her "self-adopted reasons" when she is self-governing in her agency. The intuition that we

[9]See also Dewey (1891/1957, pp. 160–161). Interestingly, Watson has also revised his original statement about the conceptual relation between practical judgment and freedom in this sense of self-government. For he argued that there could be "perverse cases" in which people endorse, in the self-government sense, a course of action that they do not judge to be the best (1987/2004a, p. 169). Thus, Watson would no longer accept the Authority Principle as I define it in this section. I return to this matter in section 10.5.5. Furthermore, note also that the sense of "freedom" alluded to in this context is not the kind that includes weakness of will and may be used to hold people responsible for their lackings in self-government. I distinguish these and other concepts of free agency in section 8.4.2.

exercise authority over ourselves when we make practical judgments may now be formulated as follows:

> Authority Principle. When someone makes the practical judgment that she should ϕ, then it follows with conceptual necessity that she has a *self-adopted reason* to ϕ: she is either sufficiently motivated to ϕ (she has a "motivating reason" to ϕ), or insofar she lacks that motivation, this is due to an impairment in her self-government, such as a compulsive disorder or weakness of will.

The Authority Principle explains the normative character, the "demandingness" of practical judgments. For what could this normative character possibly consist in, we may ask, if the person who *makes* the judgment would herself not feel required to live up to it? That is exactly what puzzles us in the case of Smith's example. The person in the example does not seem to be acknowledging any lack of freedom or self-government, because he claims to be acting in according to the reasons he has. But if his practical judgment does not provide him with reasons, it becomes unclear what that judgment really means. Instead, if we adopt the Authority Principle, then we can explain our puzzlement by claiming that what this person is saying is simply incoherent.

Note that the concept of "being *sufficiently motivated* to ϕ" allows that one also has a desire *not* to ϕ, or that one experiences other negative feelings about ϕ-ing, as long as the resultant force, so to speak, of the totality of one's affective attitudes towards ϕ, is positive. Let us call such a positive resultant attitude a "resultant desire" to ϕ, or a desire to ϕ in "the resultant sense." Thus, the resultant desire may incorporate a multiplicity of desires, but also sensation responses such as pain and pleasure, and emotions such as regret or jealousy: all states that contribute motivating impetus. Where ϕ is a concrete action that the agent is capable of performing, having a resultant desire to ϕ means that she *will* ϕ. In the case of a practical judgment about a political ideal or states of affairs P that the agent does not have the power immediately to bring about, we can represent the resultant desire as a propositional attitude. In such cases, according to the Authority Principle, if A judges in approval of P, then it follows with conceptual necessity that either A has a resultant desire *that* P, or A is impaired in her self-government. Depending on the circumstances, having a resultant desire that P may involve actions like "supporting P" or "contributing

to *P*" that the agent *will* perform in some way, when presented with the opportunity. Furthermore, it may involve various intentions, plans or strategies that the agent is pursuing in order to create such opportunities.[10]

1.4.1 Internalism vs. Externalism

A more conventional term for the Authority Principle in meta-ethics is "internalism," and the view that rejects it is known as "externalism" (Brink, 1986). However, like "realism" and "relativism," these terms have lead to a considerable amount of terminological controversy. First of all, sometimes "internalism" refers to those views who claim that *judgments* carry motivational implications, of which the Authority Principle is an example, whereas on other occasions, "internalism" may refer to views according to which *reasons* have motivational implications, of which Williams's claim that all reasons are "internal reasons" is an example. I will discuss the distinction between internal and external reasons in chapter 2.

The second thing we should be aware of is that externalism is usually defended as a view about *moral* judgments, whereas internalism is usually a view about *practical* judgments. This is not a problem if all participants in the discussion agree that all moral judgments are practical judgments, but the problem is that some authors have used the term "moral" differently in this context. For example, Harry Frankfurt maintains that "Morality is most particularly concerned with how our attitudes and our actions should take into account the needs, the desires, and the entitlements of other people" (2004, p. 7). As Frankfurt sees it, such concerns may be outweighed by others. Thus, in his usage of the term "moral," moral judgments are not 'all things considered' judgments, whereas practical judgments are. This is important because, as we shall see below, Frankfurt is an internalist about practical judgments in the 'all things considered' sense. In other words, he does accept the Authority Principle. But at the same time, his way of using the term "morality" leads to a kind of

[10]The more abstract or 'remote' the proposition *P* becomes, the more difficult it gets to specify necessary and sufficient behavior dispositions in order to capture the idea that the agent possesses a resultant desire that *P*. At some point, this conceptual question seems to turn into a question of practical reason for the agent himself: "Given that I approve of *P*, what concrete steps should I take in order to promote it?" In such cases, we might perhaps answer the conceptual question as follows: the agent has a resultant desire that *P* if he has a resultant desire to turn whatever answer he himself gives to that practical question into action.

externalism about moral judgments.[11]

Thirdly, internalist theses about practical judgment are usually defined as claims about *rationality* in the literature, not as claims about authority or self-government. For example, Smith defines his internalist thesis about practical judgments as follows: "If an agent judges that it is right for her to ϕ in circumstances C, then either she is motivated to ϕ or she is practically irrational" (1994, p. 61). In order to distinguish it from other varieties and definitions of internalism, he calls this the "Practicality Requirement." The purpose of this reference to rationality is similar to that of my reference to self-government: namely, to account for the well-known weaknesses that we have briefly discussed. Both the reference to rationality in the Practicality Requirement and the reference to self-government in the Authority Principle need to be further explicated, of course, in order to argue convincingly for or from these principles. However, at a first glance, "self-government" seems to me to express better the connection between judgment and motivation that we are after, and the one that externalists mean to deny, for two reasons.

The first reason is as follows. Let us consider Smith's example of the person who refuses to donate once more. Externalism is basically the view that what this person claims is *coherent* and that there needs to be no weakness in his motivational apparatus. But some externalists might be willing to say that coherence, even though necessary, is not sufficient for rationality, and that such a person may still be called *irrational*. Such an externalist could therefore happily accept the Practicality Requirement.[12] But he could not accept the Authority Principle, because he is committed to the idea that the person in the example is, motivationally speaking, fully-functional and in charge of his own agency.

The second reason is that it seems that a person may be fully rational in his dealings with his motivational weaknesses. For example, an addict may be well aware of the motivational problems that his addiction gives rise to, and make the most rational plans in order to deal with those problems. In fact, certain motivational disorders seem hardly to have anything to do with rationality at all, as they are not disorders of rational faculties. In clinical terms, there is a distinction between *anxiety disorders*, such as obsessive-compulsive disorder (OCD), and *impulse-control disorders*, such as trichotillomania, kleptomania or pyromania (urges to pull one's

[11] I shall discuss Frankfurt's take on moral normativity in further detail in section 8.5.

[12] See Schaubroeck (2008, p. 11) for a brief discussion of this objection.

own hair, to steal, or to set things on fire, respectively, that the patient structurally fails to resist).[13] What these disorders have in common is that they typically involve behavior against the patient's own better judgment. However, whereas anxiety disorders are typically explained and treated with reference to irrational thoughts, impulse-control disorders typically aren't.[14]

Thus, in the case of obsessive-compulsive disorder, there may be all sorts of irrational thoughts that the patient has (the obsessions) and which, even though he does not endorse them (they are not his *beliefs*), drive his pathological behavior (the compulsions). But in the case of kleptomania, the drive to steal does not have such a cognitive origin. It is first and foremost an urge, a temptation, triggered by an opportunity which presents itself, that the kleptomaniac fails to resist. All the rational faculties of the kleptomaniac may therefore be fully-functional, yet his motivating system is not. If we understand the difference between rationality and irrationality along these lines, then the Practicality Requirement would account for anxiety disorders, but not for impulse-control disorders. Of course, Smith's notion of "practical irrationality" is meant in a different sense, which does cover such cases. But that only goes to show that his requirement offers little plausibility in its own right, and that it depends heavily on whether Smith can unpack "practical rationality" in an instructive manner. And if in the end Smith will be forced to simply *stipulate* that certain conative contents 'just are' irrational, then we may wonder whether "rationality" was really the appropriate term to begin with. I will return to this matter later on. For now, let me just conclude that the Authority Principle has

[13]Note that addiction, the example I just gave, belongs to neither categories. Instead, addictions are classified as *substance-related disorders*, which include varying sorts of motivational problems. Some addicts may be fully rational in the sense that they have no irrational thoughts and just have to overcome the need for the drugs. Others may have irrational thoughts, but know that these thoughts are false. And yet others may be completely unaware of their addiction or its pathological extent, which means that they are no longer acting against their better judgments—instead, they are simply getting their judgments wrong.

[14]The distinction is not so clear-cut, of course. Some anxiety disorders may not involve *thoughts* so prominently: a patient suffering from a phobia for spiders does not have the thought that spiders are dangerous or that something bad will happen when they are exposed to them. The fear for spiders is perhaps better explained as a form of disgust. Conversely, in some cases, treating someone who has been diagnosed with an impulse-control disorder may involve the correction of irrational thoughts. But these nuances do not affect my argument. The point is that in so far as disorders involve actions against the patient's better judgment, some of these actions are explained with reference to irrational thoughts, whereas others are not.

no such problems. After all, if the kleptomaniac judges that he shouldn't steal but simply fails to resist the temptation, then he is clearly not fully self-governing. The failure of self-government is exactly that which anxiety disorders and impulse-control disorders have in common.

1.4.2 *Normative Reasons for Action*

With these remarks in mind, we may wonder what sort of implications the Authority Principle has. If cognitivism is true, then the Authority Principle implies that practical judgments have a kind of dual nature: they express or establish both practical beliefs and self-adopted reasons for action. This implication reflects what Smith has called the "Janus-faced character" of practical judgments, which in his view drives the debate between cognitivists and noncognitivists (2002/2004b, p. 343). Those who want to accept both cognitivism and the Authority Principle are committed not just to the view that practical beliefs and self-adopted reasons often go hand in hand, but rather to the view that it is conceptually impossible for them to ever come apart. This means that in the case of a fully self-governing agent, his practical beliefs would be beliefs that determine his motivations.

Furthermore, what would be the implication for the Facts Principle? Suppose that some fact makes it true that A should ϕ. Then if the Authority Principle is correct, A could not come to believe this truth without adopting it as a reason for himself to ϕ. "Normative truths," as Harry Frankfurt puts it, "require that we submit to them" (2006, p. 34). Following Smith, let us call such truths "normative reasons for action" (1994, p. 94). Thus, a normative reason for action is a truth that determines a self-adopted reason for action once it becomes known to the agent whose action it is a reason for. The conjunction of the Facts and Authority Principles may now be summarized as the view *that there are facts about normative reasons for action*. As we shall see in the next section, this view is deeply problematic.

1.5 The Distinctness Principle

The idea that there would have to be a conceptual connection between what we *believe* and how we are *motivated* has troubled many philosophers. Their worry stems from a widely accepted theory of motivation, which is often credited to David Hume. In his *Treatise of Human Nature*, Hume

made the following two claims:

> I shall endeavour to prove *first*, that reason alone can never be a motive to any action of the will; and *secondly*, that it can never oppose passion in the direction of the will. (1886/1964, p. 193)

Nowadays, we would say that no set of *beliefs* could by itself motivate us to action. Instead, we are motivated by attitudes of a different kind, which may include *desires*, *passions*, *appetites*, and *emotions*—all the *affective* attitudes that contribute motivational force to what I have called our desires in the "resultant sense" in the previous section. Perhaps some affective attitudes, such as emotions, are better thought of as complex attitudes which involve both belief and affect. The point of the theory is not to rule out such possibilities, but rather to affirm that such attitudes would indeed have to be *complex*, in the sense that they can be analyzed as structures of more elementary belief-like attitudes whose content is not motivational, and more elementary affective attitudes whose content is not belief-like. For the sake of simplicity, I will sometimes represent all affective attitudes of this more elementary kind as *desires*. In this broad sense, desires include both the motivating content of complex attitudes like jealousy or anger as well as that of affective sensations like pain or nausea. This convention is consistent with our notion of desires in the resultant sense: it follows that an agent's resultant desire about something is the result of all his desires about it. In terms of this terminology, what the theory says is that beliefs cannot play the role of desires (which roughly corresponds to Hume's first claim) and furthermore, that beliefs cannot by themselves *contradict* any desires (roughly the second claim).

Of course, citing a belief may *explain* why someone is motivated to act the way he does. My belief that it is raining explains why I bring my umbrella with me. Furthermore, citing a belief revision may explain a motivational change. If I were to discover that it is neither raining nor likely that it will rain for the rest of the day, then I will no longer be motivated to carry my umbrella with me. Let us say that the belief that it is raining *generates* my desire to take the umbrella, and that the rejection of that belief *terminates* the desire.

The idea behind the theory is that citing the relevant beliefs in this example only explains my motivation on the *assumption* that I already had the desire to prevent myself from getting wet (and, perhaps, the desire to carry as little stuff with me as possible). In general, if an agent already

desires that P, then her belief that ϕ would realize P may generate a desire to ϕ. In that case she has a "primary reason" (Davidson, 1963/2001a, p. 4) or "motivating reason" (Smith, 1994, p. 92) to ϕ. But this is a reason to ϕ only because she already had the desire that P. Of course, her desire that P may also be motivated in the light of what she believes. But then her relevant belief would be that P would realize some further end Q, which means that she would have to already have the desire that Q. We may summarize this view into the following principle:

> Distinctness Principle. In order for an agent to be motivated to act in a certain way, she must desire that she acts in that way. Desires can be generated or terminated on the basis of what the agent believes, but only under the following condition: if a desire that P is generated by a belief that Z, or if it is terminated by a rejection of a belief that Z, then there is a Q such that (i) the agent desires that Q; (ii) the desire that Q is not generated by the belief that Z; (iii) Z implies that P would realize Q (or contribute to the realization thereof).

This principle implies that ultimately, motivated agents must have so-called "intrinsic desires," which are not generated by any beliefs at all.[15] Furthermore, the principle implies that every desire that is not intrinsic depends on one that is. An intrinsic desire is conceptually *distinct* from what the agent believes (hence the name of the principle). It is an "original existence" (Hume, 1886/1964, p. 195): it is its own reason for having the content it has, as it were; it does not have to represent what is true in the actual world. Moreover, even desires that are generated by beliefs do not have to represent the actual world. Instead, they represent how the agent wants the actual world to change given his beliefs about its current state. Unlike beliefs, therefore, desires are *non-cognitive* attitudes: they have no truth-conditions.

[15] I take this to be a material implication. In order to turn it into a logical entailment we'd presumably have to explicate certain plausible assumptions against deriving desires in cycles that would undermine their role in explaining behavior. Of course, the interrelated desires that we talk about when we explain our behavior often have dependencies in mutual directions, but if a desire that P were *fully* derived from a desire that Q then the desire that Q could not ultimately be fully derived from the desire that P. It might be, due to the holistic nature of language or consciousness perhaps, that no particular desire *attribution*, or even no particular desire *experience*, is ever fully intrinsic, but then a certain form of "original" desire which for conceptual reasons always escapes full articulation must still be presupposed. The content of desire has to come from somewhere, it cannot just be running in circles.

1.5.1 The Facts Problem

We can now see how taken together, the cognitivist analysis of practical judgments, the Authority Principle and the Distinctness Principle yield a puzzling conceptual connection between what self-governing agents believe and how they are motivated. For if believing something cannot by itself, without the help of any intrinsic desires, terminate an existing desire nor generate a new one, then how can it be a requirement of self-governing agency that we must be motivated in certain ways if we have certain beliefs?

Things become even more puzzling if we also accept the Facts Principle. On the one hand, recall that together, the Facts and Authority Principles yield the view that there are facts about normative reasons for action, which are such that merely knowing them would make it conceptually impossible for self-governing agents not to be motivated in certain ways. On the other hand, according to the Distinctness Principle, for an agent to be motivated in certain ways, the agent must always possess certain intrinsic desires which do not depend on his beliefs. This seems to suggest that, whatever the agent believes, he could always have had different intrinsic desires, in which case his concrete motivations might also have been different. But now it becomes really difficult to answer the Facts Question. What sort of facts could make beliefs true in such a way that (a) self-governing agents who had those beliefs would have to be motivated in certain ways, if it is also true that (b) their motivation depends on intrinsic desires that could have been different regardless of their beliefs? Requirement (a), which follows from the Facts and Authority Principles, seems to contradict requirement (b), which follows from the Distinctness Principle. Let us call this the "Facts Problem."[16]

1.5.2 The Disconfirmation Problem

A similar problem presents itself when we try to answer the Disconfirmation Question. If the Authority Principle is correct, then it is conceptually necessary that a *change* in our practical views implies a *corresponding* change in our self-adopted reasons for action. If the Disconfirmation Principle is also correct, then this means that if an agent realizes that X disconfirms

[16]This is basically my version of Smith's "moral problem," which he constructs as a paradox resulting from three requirements that are similar to the three principles employed here (1994, p. 12).

her belief that she should ϕ, and requires her to adopt the belief that she should ψ instead, then in the light of X, she would not only be irrational if she failed to change her beliefs accordingly, but she would also be lacking in self-government if she would not lose her resultant desire to ϕ and gain a resultant desire to ψ. The question is, what sort of X might have this dual impact on her attitudes?

Cases of instrumental reasoning are easy, of course: suppose that X is evidence that ψ, rather than ϕ, would allow the agent to make it the case that P. If she had a derived desire to ϕ in order to fulfill her intrinsic desire that P, she may be expected, upon learning of X, to lose her desire to ϕ and acquire a desire to ψ instead. The problem is how to answer the Disconfirmation Question in non-instrumental cases. If, as the Distinctness Principle implies, intrinsic desires are entirely non-cognitive attitudes, which are not subject to matters of belief, then how could there be any X such that X would both disconfirm a belief and diminish an intrinsic desire of a self-governing agent? Let us call this the "Disconfirmation Problem."

We have now seen how the Facts, Disconfirmation, Authority and Distinctness Principles lead to two philosophical problems: the Facts Problem and the Disconfirmation Problem. I have not yet discussed how the Intersubjectivity Principle relates to these problems, and whether the difference between relationalism and nonrelationalism is relevant in this context. Furthermore, I have also not yet made a connection between the Disconfirmation Problem and the distinction between procedural and nonprocedural answers to the Disconfirmation Question.

I will turn to these matters in the remaining chapters of part I. In part II my strategy will be to discuss possible solutions to the two problems as I have constructed them in this chapter, and then to investigate whether those solutions become more or less plausible, or perhaps even impossible, depending on whether we adopt proceduralist, nonproceduralist, relationalist, or nonrelationalist interpretations of the Disconfirmation and Intersubjectivity Principles.

However, before we discuss any solutions to the problems as I have set them up, I will first, in chapters 2 and 3, discuss an important historical precursor to the solutions that I will later on propose. This precursor is a view defended by Bernard Williams, which bears important similarities to the view that I will develop in this thesis. Whereas this chapter has been a *thematic* introduction to the project of this thesis, in which I have set up the subject matter in my own preferred terms, the next chapter may be

thought of as a *historical* introduction. Its purpose is not only to give credit where it is due, but also to help those who are familiar with the debate locate my proposal within the literature. In chapter 3, Williams's ideas will be connected to some of the terminology from the present chapter, setting the stage for the further development of my own framework in chapter 4.

2 *The Internal Reasons View*

The proposal that I want to develop in this thesis bears certain similarities to the view that Bernard Williams defended in "Internal and External Reasons" (1980/1981a). Although I will often refer to some of the concepts and distinctions from the previous chapter in order to disambiguate Williams's terminology, my purpose in this chapter is to discuss his proposal on its own terms, while the next chapter will be devoted to questions about how the view may be related to the Principles from chapter 1.

In section 2.1 we take a look at the central claim that Williams puts forward, the "Internal Reasons View," and how it differs, exactly, from its opponent, the "External Reasons View." In section 2.2 we focus in more detail on the central distinguishing notion, that of the "sound deliberative route." Then, in section 2.3, I discuss Williams's argument in defense of the view, and in the final sections 2.4 and 2.5 we will examine two possible objections against his defense. In order to counter these objections, we must explicate certain premises that with help me to clarify some of my own purposes in the next chapters.

2.1 Internal and External Reasons

According to Williams, we should distinguish between an "internal" and an "external" interpretation of statements about reasons for action. On the internal interpretation, in order for such a statement to be true there must be elements in the "subjective motivational set" of the agent in question from which the agent could, by means of sound deliberation, acquire the motivation to do what the statement says he has reason to do. The subjective motivational set, abbreviated by Williams as "the *S*," contains the motivational characteristics of the agent, but construed in a very broad sense, so as to include:

> such things as dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may

> be abstractly called, embodying commitments of the agent. (p. 105)

In contrast, on the external interpretation, there can be a reason for an agent to ϕ even if no amount of deliberation on the basis of his S could provide him with the motivation to ϕ. Williams discusses the example of Owen Wingrave, who is urged by his family to join the army. Owen hates everything about the military. All his attitudes are against it. *Ex hypothesi*, Owen has no reason, on the internal interpretation, to join the military. But according to his family members, there is nevertheless a reason why he should join: all his male ancestors had done the same before him, and he would violate the family pride if he would not follow them in their footsteps. As Williams put it:

> Knowing that there was nothing in Owen's S which would lead, through deliberative reasoning, to his doing this would not make them withdraw the claim or admit that they made it under a misapprehension. They meant it in an external sense. (p. 106)

If a reason statement is true on the internal interpretation, then the agent has an "internal reason" in Williams's terminology; if it is true on the external interpretation, then he has an "external reason." Having established this distinction, Williams goes on to argue that there are no external reasons: all reason statements that are meant in the external sense, such as the statement made by Owen's family members, are "false, or incoherent, or really something else misleadingly expressed" (1980/1981a, p. 111). Conversely, any reason statement which is true must be understood in the internal sense. Let us call this view the "Internal Reasons View." In contrast, let us call the view that some reason statements are true in the external sense, and thus that there are external reasons, the "External Reasons View."

2.1.1 External Reasons and Externalism

Some authors have, instead, used the terms "internalism" and "externalism" to refer to these two views, including Williams himself (2001, p. 91; see also Skorupski, 2007, p. 93). However, we should distinguish the External Reasons View from the various other views that have been going under the name of "externalism" in practical philosophy which I discussed in

section 1.4.1. In particular, note that the view differs from the type of externalism according to which an agent might coherently judge that he should ϕ without having a self-adopted reason to ϕ. In other words, the External Reasons View differs from the sort of externalism that would reject my Authority Principle or Smith's practicality principle. Recall, once more, Smith's example of the person who admits that he should give to famine relief, but denies that he also has a reason to do so. This seems to be different from what Owen's family members are claiming about Owen: their claim is that Owen *does* have a reason to join the army. And the difference is not just a verbal matter of when to use the phrase "having a reason." The point of the famine relief example was to illustrate the idea that there might be nothing irrational or deficient about acknowledging, on the one hand, that one should do something, and deciding, on the other hand, nevertheless not to do it. On such a view, self-governing agency does not by itself demand submission to this sense of 'should.' In contrast, Williams is using the example of the Wingraves to articulate a more demanding way in which people mean their statements when they attribute reasons to others: a way that does involve the implication that the agent is flawed, deficient, or mistaken, in his agency if he is not responsive to the reason. Initially, Williams thought this would commit the external reasons theorist to the claim that such an agent would be irrational:

> There are of course many things that a speaker may say to one who is not disposed to ϕ when the speaker thinks that he should be, as that he is inconsiderate, or cruel, or selfish, or imprudent; or that things, and he, would be a lot nicer if he were so motivated. Any of these can be sensible to say. But one who makes a great deal out of putting the criticism in the form of an external reason statement seems concerned to say that what is particularly wrong with the agent is that he is *irrational*. (1980/1981a, p. 110)

Thus, if Owen would deliberate correctly upon his S and still lack any motivation to join the army, then to say that he nevertheless has an external reason to join the army is to say that he is irrational in not having any motivation to do so. Later on, Williams has retracted this explication of the meaning of external reason statements, saying that it was "too strong" (2001, p. 93). Nevertheless, what seems essential to the reasons attributed

by external reason statements is that the agent is held to be mistaken or deficient in *some* sense if he does not act upon those reasons.

If Owen were indeed mistaken in this sense, then it follows that, if he were to recognize and correct his mistake, he would also have to act accordingly. After all, if he would not act accordingly, then given the above analysis of his purported external reason, he would still be mistaken or deficient. Of course, it might be that in addition to his mistake, Owen was also suffering from a weakness of will that would still prevent him from joining the army even if he were to recognize the reason why he must join. But let us suppose that, instead, Owen was fully self-governing in his agency. And let us suppose furthermore that his external reason outweighs any reasons not to join the army. Then, given the idea that he is mistaken if he does not join the army, it follows that recognizing his mistake would make Owen join the army.

This implication reflects another general feature of every reason statement: that it must be a "possible explanation" of the agent's action (Williams, 1980/1981a, p. 106). As has often been remarked, reasons have a dual nature: they play both explanatory and justificatory roles. The justificatory role lies in the implication that Owen must somehow be mistaken if he does not act upon the reason he has. The explanatory role, conversely, lies in the implication that Owen will act upon the reason if he is not mistaken. So construed, the two roles are clearly two sides of the same coin.

2.1.2 *Internal and External Reasons Are Normative Reasons*

It may seem that this dual role is at odds with the distinction between normative reasons and motivating reasons for action. Indeed, Williams seems to be rejecting such a distinction:

> Some writers make a distinction between "normative" and "explanatory" reasons, but this does not seem to me to be helpful, because normative and explanatory considerations are closely involved with one another. (2001, p. 93)

However, I think the interrelatedness between the justificatory and explanatory dimensions of reasons do not preclude making such a distinction. The distinction may be unhelpful for those who would want to rid normative reasons of all explanatory relevance, but we need not deny the explanatory

relevance of normative reasons in order to distinguish them from motivating reasons. Recall that motivating reasons are reasons that *actually* explain the agent's behavior on the basis of the motivating attitudes the agent has (however construed). But the fact that normative reasons are not explanatory in this actual sense does not mean they are not "possibly" explanatory in the sense that Williams has in mind, i.e., explanatory if certain conditions are met. On the contrary, it is in terms of the conditions under which they would explain the agent's actions that we have defined normative reasons in the first place: a normative reason for A to ϕ is a truth such that, if A believes this truth and is self-governing in his agency, then A will ϕ.

One might still object that, according to the distinction, at least motivating reasons would only have a single role: they only explain, but they do not justify actions. In terms of the notion of justification we are here discussing, that would be correct. Nevertheless, there is still *something* of a justificatory aspect to motivating reasons. When an agent acts upon a motivating reason to ϕ, then his action can be made sense of in terms of that reason, even if it is not what he has a normative reason to do. Motivating reasons "rationalize" actions, as it is sometimes put, which is what makes explanations in terms of motivating reasons different from physiological explanations of behavior. Indeed, it is this very consideration which Williams also cites when he argues that every reason that we use to explain what an agent actually does must have a certain normative aspect:

> if we explain what A does in terms of his reason for doing that thing, which is one type of giving a reason why he did it, we rationalize his conduct (in the phrase familiar from Davidson's work): that is to say, we cite a consideration which was effective in his coming to act because it made normative sense to him. (2001, p. 93)

But it seems to me that this normative aspect does not make motivating reasons indistinguishable from normative reasons. After all, the attitudes that make the agent's actual actions intelligible need not be attitudes that he would ever act upon under the conditions in terms of which his normative reasons are understood. In fact, the attitudes which rationalize his current behavior might not even be part of his motivations *at all* under those conditions. Thus, we can accommodate the intuition that all reasons for action must have justificatory as well as explanatory aspects and still

make the distinction between motivating and normative reasons for action, such that an agent might be said to have a motivating reason to ϕ in the absence of a normative reason to do so, and vice versa. It turns out that normative reasons do have an explanatory aspect, but that it is different from that of motivating reasons, and conversely, that motivating reasons have a justificatory aspect that is different from that of normative reasons.

Another thing to note is that we have been using the notion of normative reasons in an 'all things considered sense,' such that the agent would have a resultant motivation to ϕ under the conditions in which such reasons explain the agent's actions. But sometimes Williams talks about having reasons in a sense that may be outweighed by other reasons:

> "A has reason to ϕ" does not mean "the action which A has overall, all-in, reason to do is ϕ-ing". He can have reason to do a lot of things which he has other and stronger reasons not to do. (1980/1981a, p. 104)

However, we can make the same distinction between normative reasons and motivating reasons at this non-resultant level of reason talk. Thus, a 'non-resultant' motivating reason explains a non-resultant amount of motivation that the agent actually has. And a 'non-resultant' normative reason is a truth such that the agent would have that 'non-resultant' motivation if she believed this truth and were self-governing in her agency.

Given the distinction between normative reasons in the all things considered and the non-resultant sense, there is now a further question we can ask about the interrelatedness of motivating and normative reasons. We have seen that an agent lacks in self-governance if he does not act in accordance with his self-adopted reason, i.e., with what he takes his all-things-considered normative reason to be. It could be the case that the agent did believe there was some, non-resultant, normative reason in favor of what he did, but that there were stronger normative reasons not to do it. We may wonder, however, if his action could still be explained by a motivating reason that he had, if the agent did not think there was any non-resultant normative reason in favor of it whatsoever. Perhaps it is this possibility that Williams wanted to deny when he said that the reasons which explain an agent's actions must always make some normative sense to him. And that it is in this sense that we cannot fully distinguish between motivating and normative reasons.

I have my doubts about whether this is plausible, however. Consider Harry Frankfurt's example of the unwilling addict (Frankfurt, 1971/1988b, p. 17). The addict may be wholehearted in his rejection of his drug use, it seems, and still fail to resist it. His desire for the drugs may be too strong, and in combination with beliefs about where to get the drugs, or how to operate them, explain how it came to be that, once again, he's heating up the heroine. That seems like a perfect case of a motivating reason without that reason having any normative credibility whatsoever in the eye of the agent himself. The justificatory aspect of this motivating reason is *solely* with reference to a disenfranchised desire that the agent has come to regard as fully external to his evaluative outlook. Under the idealized circumstances in terms of which we understand his normative reasons, the agent might not have any remaining desire for the drugs.[1] However, note that this possibility is not a necessary condition for the distinction between motivating and normative reasons. For without this possibility it would still follow that a motivating reason in the resultant sense only requires the agent to believe that he has a normative reason in the non-resultant sense, and that clearly distinguishes the limited sense in which motivating reasons are justificatory from the full-blown sense in which normative reasons are.

With these remarks in mind, we are now in a position to observe that the conditions on the explanatory role of normative reasons as I have explicated them (whether understood in the all-things-considered or the non-resultant sense) are the exact same conditions that Williams places on the reason statements that he is considering, including external reason statements:

> Does believing that a particular consideration is a reason to act in a particular way provide, or indeed constitute, a motivation to act? [...] Let us grant that it does—this claim indeed seems plausible, so long at least as the connexion between such beliefs and the disposition to act is not tightened to that unnecessary degree which excludes *akrasia*. (1980/1981a, p. 107)

Provided that we identify Williams's use here of the notion of *akrasia* with our notion of lackings in self-governance, it follows that we may apply

[1] For a different line of argument in support of roughly the same conclusion, see also Setiya (2010).

Williams's arguments about internal and external reason statements to our notion of normative reasons.[2]

2.1.3 *Self-Attributed External Reasons Entail Internal Reasons*

Given the explanatory requirement on external reasons that we have now established, it follows that once Owen would recognize that he has an external reason to join the army, because this would give him the motivation to act accordingly (absent lackings in self-governance), he would thereby also acquire an *internal* reason to join the army:

> The claim [that the agent would be motivated to act if he believed himself to have an external reason] is in fact *so* plausible, that this agent, with this belief, appears to be one about whom, now, an *internal* reason statement could truly be made: he is one with an appropriate motivation in his *S*. A man who does believe that considerations of family honour constitute reasons for action is a man with a certain disposition to action, and also dispositions of approval, sentiment, emotional reaction, and so forth. (1980/1981a, p. 107)

This implication is interesting for two reasons. First of all, it means that an agent can never coherently attribute an external reason to himself without also attributing an internal reason to himself. This gives us another way of understanding the difference between the Wingrave example and the famine relief example. According to the sort of externalist who rejects the Authority Principle, a person might coherently claim that, even though he should give to famine relief, he does not have a reason to do so. By contrast, the External Reasons View is not the view that Owen might coherently claim of himself that he has an external reason to join the army without having an internal reason to do so. Or that the person in the famine relief example might also have said, coherently, that even though he knows he

[2] This conforms to the loose sense in which the term "akrasia" is often used nowadays, as the lack of motivation to execute one's judgment about what one has reason to do. Instead, for Aristotle, akrasia involved a failure to translate one's general sense of how to act to the appropriate concrete action as a result of a misapprehension of the concrete situation. But that does not seem to be what Williams has in mind here, for such a misapprehension would simply be reflected in another false belief about which action the agent has reason to perform in the present situation. Despite his great affinity with the classics, I therefore take Williams to have been using "akrasia" in the modern sense here.

has an external reason to donate, he does not think he has an internal reason to do so. The External Reasons View only allows *others* to attribute an external reason to an agent without attributing an internal reason to him.

Second, the implication is interesting because it means that the External Reasons View is committed to the idea that external reasons are a kind of counterfactual internal reasons. If an external reason statement is true, then that truth determines certain elements that would be in the agent's S if the agent would believe that truth. The challenge for the External Reasons View is to explain what sort of reasons could have such implications for the agent's S without being internal reasons. In order to see whether this challenge can be met, we must first take a closer look at the precise requirements that constrain the notion of an internal reason.

2.2 The Deliberative Route

In his latest writing on the subject, Williams summarized the Internal Reasons View as follows:

> The formulation of the internalist position which I now prefer is: A has a reason to ϕ if and only if there is *a sound deliberative route* from A's subjective motivational set (which I label "S," as in the original article) to A's ϕ-ing. (2001, p. 91)

Note that Williams has now switched to talk of reasons in the all things considered sense. If the agent *would* ϕ after sound deliberation, then deliberation would have provided him with a resultant motivation to do so, rather than merely a certain non-resultant amount of motivation. And a resultant motivation after sound deliberation constitutes a reason in the all things considered sense:

> It is natural to take the condition as implying not just that A has a reason to ϕ, but that he or she has more reason to do that than to do anything else. (2001, p. 91)

From now on I will stick to the discussion of all-things-considered reasons, except when explicitly noted otherwise. Nevertheless, there is nothing in the above two remarks from Williams that would prevent us from translating his view back into non-resultant reason talk: if an all-things

considered reason is constituted by a resultant motivation after sound deliberation, then a non-resultant reason may still be understood in terms of a non-resultant amount of motivation after sound deliberation.

With that out of the way, let us now focus on the crucial notion in the formulation: that of a "sound deliberative route." Essential to the idea of such a route is that *it must start from somewhere*, and if the agent is to have an internal reason, then that starting point must be his actual *S*. However, it may seem strange that Williams is using the notion of soundness to depict this idea with. In logic, we say that an argument is *sound* when (a) it is valid and (b) its premises are true. Thus, in order for an argument to be sound, there has to be something right about the premises it started from. Instead, the Internal Reasons View seems committed to the opposite idea: that whatever *S* we happen to have *can* and *should* be the starting point for the deliberative route that we must take, and that there is no additional sense in which there has to be anything true about this *S* in order for us to be able, in principle, to reach our reasons by means of that route.

2.2.1 *Instrumental Deliberation*

Nevertheless, there is an important feature of deliberation that might explain why Williams may have used the word "sound," which is that it is in the interest of a deliberating agent that he takes the correct means to his ends. It follows that deliberation should lead him to correct errors in his beliefs. Williams gave the example of a person who wants to drink tonic and mistakes a glass of gin for a glass of tonic. The notion of sound deliberation is meant to cover the correction of such false beliefs, so that we can say that the person does not have an internal reason to drink from the glass. Thus, we can say that the deliberative route must lead to sound instrumental arguments that are based on true beliefs, even though the agent might have false beliefs at the starting point of his deliberations.

Note, by the way, that the gin and tonic example already presupposes our distinction between normative and motivating reasons. For suppose that the agent goes and drinks from the glass. Surely, there is now a reason why he did that (the motivating reason), even though he did not, according to Williams, have a reason to do so (the normative reason). The example does, however, lie within the requirement that the motivating reason must at least make normative sense to the agent himself. Given that the agent really believes that the glass contains tonic, it makes sense to him to drink

from it.

There is a further limitation that Williams places on the possible distance between the actual motivations of the agent and what he has normative reason to do. Williams distinguishes between, on the one hand, rejecting false beliefs and adopting true ones to replace them with, and on the other hand, the adoption of true beliefs on matters that the agent did not initially hold beliefs about at all. With respect to the first type of revision, Williams thinks that any considerations that depend on false beliefs are not really justified, and therefore, internal reasons must correspond to what the agent would do if all such falsehoods were removed. However, with respect to the second type of belief change, the mere *addition* of true beliefs without thereby removing any falsehoods from the original belief base, Williams proposed the following restriction:

> A may be ignorant of some fact such that if he did know it he would, in virtue of some element in S, be disposed to ϕ: we can say that he has a reason to ϕ, though he does not know it. For it to be the case that he actually has such a reason, however, it seems that the relevance of the unknown fact to his actions has to be fairly close and immediate; otherwise one merely says that A would have a reason to ϕ if he knew the fact. I shall not pursue the question of the conditions for saying the one thing or the other, but it must be closely connected with the question of when the ignorance forms part of the explanation of what A actually does. (1980/1981a, p. 103)

I mention this restriction in the interest of getting Williams right; nevertheless, I must admit that I find it wholly unconvincing, for several reasons. First of all, I would say that if an unknown fact is relevant to an element in the agent's S, then the agent has an interest in discovering this fact, just like the agent would have an interest in removing one of his false beliefs if that were relevant to an element in his S. If the latter type of interest is not constrained by how "close and immediate" the relevance has to be (whatever that is supposed to mean), then why should such a constraint apply to the former type of interest? The first argument that Williams gives is that it would be more natural to say that the agent "would have a reason," but this seems rather weak as a defense of a constraint with such normative implications for the agent.

Consider the following example. Suppose that the facts about global warming and the elements in our *S*s are such that, the more we learn about these facts, the greater our motivation becomes to reduce our carbon emissions. Because of the attention for the issue in the media, most adults in the West will hold beliefs about this matter nowadays, and a certain amount of the unwillingness to reduce emissions is due to widespread false beliefs about the matter. Nevertheless, the devil is in the details, and about many relevant details even our best scientists remain simply ignorant. Furthermore, global warming was well underway thirty years ago, when many people lacked any beliefs on the matter whatsoever. Suppose that such ignorance, in the absence of concrete false beliefs, would not be closely connected enough to the explanation of these people's carbon emitting behavior in the manner that Williams had in mind. Then, according to Williams, they had no reason to behave in ways that would have prevented the current climate crisis. They only "would have had" such a reason if they had known what we know now (or what the scientists of tomorrow will know in further detail).

I am not convinced that this would be more natural to say for a native English speaker. However, even if we would grant Williams this point in the linguistic sense, then it still seems that the ambitions of the Internal Reasons View should go beyond the contingent semantics of English natural language. Surely ignorance about global warming does not eliminate its normative relevance? Surely, if our *S*s contain attitudes about famine, flooding, and various other types of natural disaster, then a *sound deliberative route* should lead us to discover the facts about global warming?

Of course, we cannot expect ourselves to know everything that would be relevant to our interests if we did know it. We are finite creatures, after all, and thus it is only rational that we expect from ourselves and each other only what lies within the boundaries of our cognitive capacities and the state of knowledge in our times. However, a normative reason, in the all-things-considered sense, does not determine what may be expected from us, but merely what would be best for us to do. It does not follow from saying that it would have been best for us to reduce our emissions, that this could also reasonably be expected from us. The latter would be a claim about what we could be reasonably held responsible for, not merely about what we have reason to do. Thus, even if people could not be blamed for their failure to understand that cutting emissions would have been a correct means to their ends, doesn't mean we cannot say that

people did have a normative reason to cut emissions.

Perhaps Williams would have wanted to reject the global warming case as a relevant example, by simply responding that in this case, the facts about global warming simply *are* sufficiently close to our interests to warrant their normative relevance. But then he would run into his own second argument, which is that the ignorance must be part of the explanation of the agent's behavior. After all, it seems that we can simply explain why people drive cars, take planes, and use lots of electricity without bringing in their ignorance about global warming. We can simply explain their behavior given their interests and the things they *do* know. Of course, we need to explain their behavior in terms of their ignorance if we wanted to answer the question why they *didn't emit less carbon dioxide* given the dangers of such emissions. But if *that* already counts as a relevance to the explanation of behavior, then it becomes hard to see how any sort of ignorance that *would* fulfill the role of making a "would have a reason" statement true could possibly fail to also make a "has a reason" statement true.

Furthermore, the restriction requires a lot of distinctions in the logic of belief revision that I am not so sure about. First of all, the distinction between correcting false beliefs and adding new beliefs can, I suppose, be modeled in various ways, but would such a model capture the reality of the interconnectedness and holism of our beliefs? Second, the idea that a fact, of which the relevance to our interests can be established, might nevertheless be too "distant" to be granted normative significance presupposes not only that we can articulate such a scale of proximity, but also that we can make a case for a boundary value beyond which talk of "having reasons" would no longer be justified. I have just argued that the criterion of explanatory relevance is problematic, but other criteria might be equally problematic.

Furthermore, even if we could determine such a value, we would run into another problem. For suppose that the relevance of some fact F would be too distant from the interests in an agent's S. Then there might be another fact, G, that would be close enough to have relevance. However, it is conceivable that once the agent would learn about G, the relevance of F would become suddenly much more immediate. Hence, a sound deliberative route would, through the discovery of G, also lead to the discovery of F. But now we may wonder if there could be any sort of facts that would be relevant to us if we knew them, but which might not be

reached, in principle, through the discovery of intermediate facts that make the more distant facts more accessible to us, and their relevance clearer?[3]

It seems to me that Williams is getting himself into a lot of trouble over what should have remained an innocent and normatively irrelevant matter about when people are more likely to say, in everyday natural language, that an agent "would have had a reason" if he knew some fact, rather than that he "had a reason" in the light of that fact. Furthermore, if we accept the distinction between normative and motivating reasons, we can even accommodate the sense in which such agents do not have a reason by saying that they do not have a motivating reason. And we can be specific about the sense in which they do have a reason: they have a normative reason because it is based on facts that are relevant to their interests, which means that their sound deliberative route would not be completed if they remained ignorant of these facts. In what follows, I will therefore discuss the prospects of the Internal Reasons View without holding this restriction against it. Instead, I will understand the deliberative route as involving both the removal of all false beliefs and the adoption of all true beliefs that are relevant to the ends of the agent.[4]

[3]Perhaps if *G* could be discovered on the basis of our current "paradigm," in Kuhn's sense, while the discovery of *F* would require a "scientific revolution," then we might say that *F* would be too far removed. Indeed, the difficulty of understanding such a transition as rationally required from a deliberative point of view was at the heart of Kuhn's philosophical interests. However, this difficulty is closely related to Kuhn's famous incommensurability thesis, which makes it impossible to migrate the truth conditions of the beliefs in the old paradigm to those in the new paradigm. I am not sure that the Internal Reasons View is even compatible with that, but even if it were, then I would be unhappy about making it *dependent* upon the Kuhnian framework.

[4]Michael Smith's interpretation of Williams's view is that an agent has a normative reason to ϕ iff he would be motivated to ϕ under the following conditions: "i. the agent must have no false beliefs"; "ii. the agent must have all the relevant true beliefs"; and "iii. the agent must deliberate correctly" (Smith, 1995/2004d, p. 20). Condition iii. covers non-instrumental deliberation about the agent's ends themselves, to which I shall turn below. Conditions i. and ii. cover the types of belief revision which I have just discussed. By "all the relevant true beliefs" Smith means all beliefs that are relevant to the agent's ends. Nicole Saunders has noted (in a paper of which I only have a draft version) that condition ii. is therefore a misrepresentation of Williams's view, because it does not accommodate the proximity restriction that I have just discussed. She also argues that ii. is implausible because practical rationality should be on a par with the justification of beliefs, rather than with their truth. I think her argument equivocates on what "justification" means, however. In contrast, I have just argued at length *against* the proximity restriction, and proposed to include all true beliefs relevant to the agent's interests, which is identical to Smith's condition ii. Thus, in my view, Smith's portrayal of Williams, though exegetically inaccurate, is at least in this respect also an improvement.

2.2.2 *Deliberation about Ends*

The second aspect of the deliberative route is, of course, to establish those ends themselves. For now I will speak neutrally about "ends" as the possible elements of an agent's S that provide the intrinsic normativity, so to speak, without yet considering whether such elements should be thought of as Humean, non-cognitive desires, or as anti-Humean attitudes that are truth-apt and motivational at the same time. Now first of all, Williams remarks that deliberation about the ends themselves may involve considerations of planning and time-ordering so as to be able to combine the satisfaction of different ends. This is an interesting topic in its own right, but at the same time we can see how the normative relevance of such considerations is pretty much on a par with the normative relevance of instrumental reasoning. From a meta-ethical point of view, even if considerations of planning and organizing constrain plausible theories of practical reason, the reason *why* planning has normative relevance is not such a mystery. Furthermore, the normative substance that such considerations provide us with seems limited: even if the very fact that deliberators must be planners would provide us with some substantial values that all agents would have to recognize, then it still seems that (a) deliberators must also hold further ends in the interest of which they make their plans, and (b) the relatively formal considerations of good planning will be entirely neutral about what those ends might be.

A second possibility for deliberation about ends, which is more important for our present purposes, involves "where there is some irresoluble conflict among the elements of S, considering which one attaches most weight to" (1980/1981a, p. 104). Here we begin to touch upon the prospects for a dispositional solution that I mentioned at the beginning of this chapter. Note that this idea harbors an important assumption, however, which is that the relative "weights" of the different elements in the agent's S are not simply given as part of those elements themselves, at least not in a manner that is immediately transparent to the agent himself. Thus, one of the things that the deliberative route involves is the resolution of conflicts *within* the S by making the relative weights of the conflicting elements more explicit.

In fact, not only the importance of a certain end, but the very presence of that end in the S of an agent might not be transparent to the agent himself. Williams briefly mentions this possibility when he discusses the ways in which agents might misjudge their internal reasons due to

ignorance (1980/1981a, p. 103). However, this idea does not play much of a role in the rest of his account. Instead, a similar idea of *volitional opacity* will become the central element of the account that I will develop in chapters 7 and 9.

Besides making elements in the S more explicit or transparent, deliberation can also, according to Williams, add *new* elements, or subtract old ones, and this seems to include elements that constitute ends. One of his examples is that we might lose our desire to pursue a certain end once we arrive at a more concrete sense of what would be involved in attaining it. Conversely, deliberation might provide us with "constitutive solutions, such as deciding what would make for an interesting evening" (p. 104). These may initially be understood as a kind of means to more abstract ends ("an interesting evening"), but the means become ends in their own right once we have set our minds to them. For example, once I got the idea of ordering pizza in my head, I might no longer be happy with anything else, even though ordering Chinese takeaway food might have satisfied me if I had thought about it first. Note that this type of deliberation makes internal reasons to a certain extent *indeterminate:* on the basis of my S at time t_0, I merely have a reason to make my desire for something to eat more concrete, such that both pizza or Chinese food would do, but if I were to settle for pizza, then at t_1 I would have a reason to order pizza rather than Chinese. We shall be discussing this sort of indeterminacies extensively later on.[5]

Despite these suggestions, Williams remains intentionally vague about the various ways in which we deliberate on our ends. As we shall see later in this chapter, this vagueness poses a problem, and in fact this will be one of the problems that this thesis is meant to address. For now, however, we should note that whatever the principles or methods of deliberation may be, in Williams's view they must always work *from* the current S of the agent. For example, if a certain adjustment in the S resolves a conflict among its elements, then the adjustment would itself be motivated by the fact that the conflict in the prior S called for such an adjustment. Thus, even though the motivational set has changed, the change was motivated by the state of the initial set, maintaining a kind of motivational continuity. This is what I take to be the idea of a deliberative *route*. The constraint that this idea places on normative reasons is that the reason why they are reasons for a particular agent can always be traced back to elements in

[5]See section 9.2.3.

the S of that agent at any stage in the deliberative process. That's what makes reasons, understood along these lines, "internal" to the S of the agent. Given this explication of the notion of an internal reason, let us now take a look at Williams's argument for the Internal Reasons View—the view that all normative reasons must be internal in this sense.

2.3 Williams's Defense of the Internal Reasons View

The central objection that Williams raises against the External Reasons View is that it cannot explain what it would mean for an agent to *come to believe* an external reason statement about himself. As we have explicated above, an external reason statement has the following two features. First of all, the statement that A has an external reason to ϕ may be true even if A does not have an internal reason to ϕ. Second, however, this discrepancy can only exist as long as A does not believe the external reason statement about himself. If A would have believed in the truth of the external reason statement, then he would have had an internal reason as well. Hence, it seems that by merely coming to believe the external reason statement, A would have to *acquire* an accompanying internal reason to ϕ.

But how could that happen? If A acquires an internal reason that he did not have before, that means an element is being added to his S that bears no motivational continuity to the prior elements in his S. His new belief in the truth of the external reason statement would have to "inject" this element in his S entirely on its own accord, as it were, without using the existing elements as leverage, and furthermore, if A is self-governing in his agency, then the newly injected element would motivate his actions without those prior elements getting in the way, so to speak, regardless of what their content or their strength happened to be. As Williams put it:

> Given the agent's earlier existing motivations, and this new motivation, what has to hold for external reason statements to be true is that the new motivation could be in some way rationally arrived at, granted the earlier motivations. Yet at the same time it must not bear to the earlier motivations the kind of rational relation which we considered in the earlier discussion of motivation—for in that case an internal reason statement would have been true in the first place. I see no reason to suppose that these conditions could possibly be met. (1980/1981a, p. 109)

The question, how our rational appreciation of truths which do not already involve or link up with our existing motivations, could make us motivated in wholly new ways, resembles the old Humean question about how reason (or beliefs) could motivate in non-instrumental ways. However, in Williams's view it is also a refinement of that question, because it now focuses on the possibility of new motivations that have no deliberative connection to the old ones whatsoever, rather than merely on the possibility of new motivations that have no instrumental, means-end connection to the old motivations. Thus, the question can no longer be answered by merely rejecting a reduction of deliberation to means-end reasoning. Instead, answering the refined question requires showing how the rational appreciation of certain truths could require, and bring about, new motivations which were not *in any way* called for on the basis of the pre-existing motivations. But if they were not called for by his present motivations, then why would the agent be rationally driven towards the new motivation?

Another way to articulate this line of argument, which I myself find helpful, is to explore the implication that the agent would be deficient or irrational if he would *not* acquire this new motivation upon learning such a truth. So for any candidate truth T, consider an agent A whose motivations do not change one bit upon his learning that T. Now let us suppose that A is wholly coherent, unified, harmonious, and what not, in his motivations: his motivational set, in its present state, does not in any way call for changes in that set. Furthermore, from the perspective of this harmonious unity of motivations that he has, A could not care less about whether T or not T. Then surely we have run out of resources for giving meaning to the claim that A would be deficient or irrational for not having acquired a new motivation?

My intuition is that we are at least close to a sound line of argument here. However, many critics of Williams have felt otherwise, and it is not my purpose to discuss all their objections or to evaluate the merits of Williams's defense.[6] Rather, my purpose is to discuss Williams's view as a historical precursor to the view that I want to put forward myself in the chapter 4. Nevertheless, there are two objections to the above line of argument that I do want to discuss at this point, because they will help us to explicate premises of Williams's argument that are of crucial importance to my own view as well.

[6]For an in-depth discussion of the dialectical nightmare surrounding the Internal Reasons View, see Thomas (2006, ch. 4, pp. 67–97).

2.4 The Nonproceduralist Objection

The first objection exploits a possibility that Williams seems to have overlooked, or at least does not address, which is that the external reasons theorist might simply *agree* that it is impossible for someone who does not already have an internal reason to ϕ to come to have the belief that he has an external reason to ϕ. The objection would then be that this does not show that there is no such external reason for the agent. It merely shows that the truths about external reasons will be *epistemically inaccessible* to those agents who did not have the appropriate internal reasons to begin with.

In order to refute this objection, we might simply want to insist on the idea that it is essential to a reason that it can motivate. If external reasons are epistemically inaccessible to those agents who lack the corresponding internal reasons, then it seems no longer possible for those agents to be motivated by the external reasons, so that they no longer satisfy this requirement on reasons generally. We have already seen that the external reasons theorist cannot simply deny this requirement, because it is this requirement that distinguishes the External Reasons View from the sort of view according to which there is an external normativity that does not generate reasons at all.

Nevertheless, I think this sort of argument against the objection is invalid, because it equivocates on two different interpretations of the idea that reasons must be possible motivations. On a weak interpretation, the motivational implication is merely a counterfactual one: if the agent would have believed in the truth of the external reason statement, then the agent would have had the corresponding behavioral disposition. In contrast, on the stronger interpretation, there would be a further condition that we must add: namely, that it is also possible for the agent to acquire that belief.

In support of the strong reading, we might say that unless it is possible for the agent to come to believe the statement, the fact that he would be motivated if he would have the belief does not show that it is possible for him to have that motivation. But that argument simply trades on the same ambiguity in the phrase "it is possible for the agent to x." The external reasons theorist can admit that it must be possible in the counterfactual sense for the agent to have the belief, without admitting that it must be possible for the agent in the stronger sense, i.e. given his actual state. As

long as there is a possible world in which the agent has the belief, one can still say that it is possible for the agent to have the motivation, given the counterfactual implication from the belief to the motivation, even if the agent were unable to reach that state in the actual world. In order to distinguish the External Reasons View from the 'no reasons' versions of externalism, it would be enough to accept only this weak reading. After all, if external facts about normativity would not provide us with reasons, then they would not even support this merely counterfactual claim about the agent's motivation should he come to have true beliefs about those facts.

This distinction between a strong and a weak sense in which it may be possible for an agent to believe something which he actually does not is very similar to our earlier distinction between the procedural and non-procedural interpretations of the Disconfirmation Principle. The premise that Williams has been presupposing is a kind of proceduralism about normative reasons, and this is a premise that the external reasons theorist may want to deny.

Now, there are three remarks I want to make at this point. First of all, if we reject the possibility of a 'no reasons' form of practical normativity and accept the Authority Principle instead, then Williams's proceduralism about normative reasons follows directly from the proceduralist interpretation of the Disconfirmation Principle about practical judgments. The second remark is that nonproceduralism is an option which we should take seriously. However, I will discuss the prospects of nonproceduralism extensively in chapter 6. For now, we should simply keep in mind that Williams's defense of the Internal Reasons View is premised on proceduralism.

Third, and most importantly for our present purpose, even though nonproceduralism does seem to offer an escape route, not every external reasons theorist may be satisfied with it. After all, if the only external reasons that agents may have, in the absence of corresponding internal reasons, are those external reasons which they can never become convinced of having, then external reasons could play no justificatory role in accounts of the sort of interpersonal moral discussions in which we try to convince our opponents of our views. Suppose that Owen's father understood that Owen did not have an internal reason to join the army. Then according to the nonproceduralist escape strategy, Owen's father would also, upon reflection, have to conclude that all his attempts to convince Owen of his

external reason are utterly pointless. There may be cases in which such a conclusion is not *prima facie* implausible, such as when one is dealing with a psychopath, say, or with a hopelessly dogmatic religious fundamentalist who seems totally unresponsive to the sort of reasons that we see against the choices that he makes. However, the Wingrave example does not seem to be such a case, at least not at first approximation. Instead, the external reasons theorist may want to be able to claim that at least some external reasons can be understood proceduralistically, even when the relevant agents do not start out with the appropriate internal reasons.

2.5 The 'Non-Route-Like' Deliberation Objection

This brings us to the second objection. According to this objection Williams's argument simply begs the question against the idea that an agent *can* come to believe an external reasons statement, and thereby acquire the accompanying internal reason, without having had that internal reason before (Hooker, 1987; see also Thomas, 2006, p. 76). Recall that in Williams's view, our notion of deliberation needs to be broader than mere instrumental reasoning, but not so broad as to violate the requirement that the deliberative process has to be a kind of "route" and that the starting point of the route must be existing motivations. However, if that is the notion that Williams uses to *spell out* the Internal Reasons View, then it better not also be the premise of his argument in defense of that view. But when he tries to argue for the implausibility of the idea that acquiring a new belief could provide us with new motivations that have no deliberative connections to our old motivations whatsoever, it seems that Williams is simply inserting that same notion of deliberation into his views of how we can and cannot acquire new motivations. And it is precisely this notion of deliberation that the external reasons theorist is going to want to reject. His response to Williams could now simply be as follows: "just as Williams has liberated himself from a strictly 'instrumental' notion of deliberation, so I, the external reasons theorist, have liberated myself from a strictly 'route-like' notion of deliberation. My notion of deliberation simply does not require that it starts from the actual motivations of the agent." Or, as Hooker put it:

> the external theorist is likely to think that (at least some) rational deliberation about reasons for action starts not from the agent's own subjective present motivations, but from some objective

> ('external') values and requirements, fixed independently of the agent's present motivations. (1987, p. 43)

Thus, the objection is that Williams begs the question against this 'non-route-like' notion of deliberation. His route-like account of deliberation is premised upon a theory of motivation that seems in turn premised upon the route-like account of deliberation.

As I see it, there are two ways in which the internal reasons theorist might respond to this objection. The first is to explicate an account of motivation as a premise of the argument for the Internal Reasons View in such a way that we can provide independent support for this premise rather than deriving it in a viciously circular way from the Internal Reasons View itself. The key to such an approach would be to make sure that, whereas the Internal Reasons View is a view about normative reasons, the motivational account remains solely a claim about motivating reasons, such that any conclusions about normative reasons derived from it would be non-circular.

In contrast, according to the second type of response we should rather *agree* that the motivational story depends on the much more normative story about the soundness of deliberation and the scope of reason, but defend the idea that this normative story behind the Internal Reasons View can stand on its own.

Very roughly, the former approach is the one that I prefer and which has, I think, been presupposed by many philosophers who read Williams as a motivational 'Humean,' while the latter approach has been advocated by Thomas (2006, p. 76), who offers a much more 'anti-Humean' interpretation of Williams's argument. However, my intention is to remain neutral about Williams-exegesis in this matter. Furthermore, as we shall see below, the strategies are not as much at odds with each other as they initially seem to be.

2.5.1 The Descriptive Motivational Response

Let me explain the first response first. The trick is to think of the relevant premise of Williams's argument as a premise that is more or less *descriptive* about the ways in which motivational dispositions of agents can *actually* change across time. Therefore, it should be possible to evaluate such a premise independently of our ideas about how we *should* reason or how we *should* deliberate. When I say that the premise is "more or less" descriptive,

what I mean is that it lacks *that* particular type of normativity, however, not that it lacks normativity whatsoever. At this point, our previous discussion of the explanatory and justificatory aspects of normative and motivating reasons will be helpful once more. We have seen, not only that normative reasons also have an explanatory dimension (which played an important role in the argument for the Internal Reasons View so far), but also that motivating reasons have certain justificatory aspects: when we attribute a motivating reason to an agent we attribute a way in which his action 'made sense' to the agent, and we presuppose that the agent satisfies certain minimal conditions of rationality, conditions without which we would not be able to understand him as an agent in the first place.

From a 'Humean' perspective on motivation, what this means is that there are certain normative aspects involved when we make claims about how the motivations of agents might change, but that these are not yet the type of normativity that meta-ethics is concerned with, the type that moral philosophers must analyze rather than merely presuppose. Instead, the normative aspects built into the theory of motivation are the normative aspects of general talk about beliefs and desires, which are ultimately normative notions in their own right. However, this way of thinking about normativity can be generalized to anti-Humean ideas about motivation as well: regardless of whether one submits to the specifics of belief-desire psychology, one may still feel that a theory of motivation only needs to presuppose the general normative aspects of agency, knowledge, and meaning: the modes of normativity that other branches of philosophy are concerned with—primarily, philosophy of action and epistemology. Therefore, any theory of motivation that pretends to be dialectically independent from and prior to the meta-ethical question of practical normativity, essentially consists in an attempt to get the normative aspects of motivating reasons from the sorts of normativity that are at stake in epistemology and the philosophy of action. The first type of response to the 'non-route-like' deliberation objection represents the philosophical ideology that these branches of philosophy must come first, as it were, and that meta-ethics must be built upon them.

Provided that this is our take on theories of motivation more generally, what then is the specific premise concerning motivation that we must deploy in order to make Williams's defense valid and non-question begging? And is there going to be independent support for this premise? To begin, let me note that in order to make the argument *valid*, the motivational

premise need not be Humean—it doesn't require belief-desire psychology. In fact, one might even be inclined to think that it is incompatible with Humean motivation because of the generalization in Williams's argument from 'instrumental' deliberation to 'route-like' forms of deliberation that include non-instrumental deliberative moves. However, in the next chapter I shall argue that there actually is no such incompatibility and that the motivational premise that Williams's argument requires *can* be supported by a Humean theory of motivation, even if it may also be supported by non-Humean views.[7] For now, however, let us set the matter of belief-desire psychology aside, and focus on what it is that the Internal Reasons View requires.

With these remarks in mind, the premise that would make the argument non-question begging, if there is independent support for it, may now simply be formulated as follows: that by itself, reason or knowledge cannot produce new motivations that do not in any way follow up on pre-existing motivations. Let us call this the "Motivational Continuity Thesis." To be sure, the thesis does not claim that motivational changes are never discontinuous, but only that motivational discontinuities could never be the product of rational insight alone. Essentially, this thesis is simply a "descriptive" counterpart to the Internal Reasons View, in the "more or less" descriptive sense explained above, i.e., as a view about motivating rather than normative reasons. What it claims is that the rational production of new motivating reasons must be "route like" without thereby presupposing that practical deliberation about normative reasons should be "route like" as well. It merely claims that any instances of non-route-like deliberation are not going to actually produce new motivations on their own.

However, even though the Motivational Continuity Thesis does not presuppose the Internal Reasons View, it does *imply* the Internal Reasons View when we combine it with the other premises of the argument, thereby satisfying the demand for a non-circular defense. In particular, when we add the premise that true reason statements motivate self-governing agents if they are known by those agents, and the premise that true reason statements must not be made epistemically inaccessible by the absence of certain motivations, then it seems we can arrive at a conclusion about deliberation and normative reasons. After all, suppose that A has an external reason to ϕ without an internal reason to do so. From the proceduralist premise about accessibility it follows that A can come to believe that he

[7] See section 3.3.

has this external reason in the absence of his internal reason, and from the premise that self-governing agents are motivated by their beliefs about their reasons it follows that A subsequently acquires the internal reason. Which means that the disposition to ϕ under conditions of self-governance must have been added to his S. But from the motivational continuity premise it follows that he cannot have acquired an actual *motivation* to ϕ on the basis of his rational appreciation of his external reason alone, since that appreciation did not connect to his pre-existing motivations. Therefore, it is possible that A does not gain any motivation to ϕ, in which case the external reasons theorist must declare him to be lacking in self-governance. Furthermore, in the case that A does acquire the motivation to ϕ, this must be due to something over and above his rational appreciation of his external reason. But *there is nothing* that could play this role. Any causes of this new motivation that were neither based on earlier motivations nor rooted in the rational appreciation of his external reason would seem wholly accidental. And given their accidental nature, it would seem entirely stipulative to make such causes requirements for self-governance. Thus, given our premises, we can now reach the conclusion that in the case where A does not become motivated to ϕ, the external reasons theorist has no way of explaining his claim that A would have to be lacking in self-governance.

In other words, the argument shows that if we start with a 'route-like' account of the rational production of motivating reasons, and we add independent insights about self-governance and the accessibility of normative reasons, then we can arrive at a 'route-like' account of deliberation about normative reasons as well. However, even though this makes the argument non-circular, strictly speaking, the air of begging the question will remain if there is no independent support for the 'route like' perspective on the production of motivating reasons that the Motivational Continuity Thesis offers. Otherwise, we would merely have argued for the non-trivial route-likeness of practical deliberation on the basis of a presumed, but equally non-trivial, route-likeness of the rational production of motivating reasons.

As it happens, I think there is independent support for the Motivational Continuity Thesis, and that it may be defended on the basis of both 'Humean' and 'anti-Humean' theories of motivation. I will provide this defense in the next chapter.[8] For now, however, let us simply note that according to the first type of response, such a defense of the Motivational

[8] See sections 3.4 and 3.5.

Continuity Thesis would refute the 'non-route-like' deliberation objection.

2.5.2 *The Content Skeptical Response*

In contrast, the second type of response de-emphasizes the importance of such an account of motivation. In his own attempt to refute Hooker's objection, Alan Thomas draws on a distinction from Christine Korsgaard between two types of skepticism about the practical impact that reason can make. The first is "motivational skepticism," which he formulates as "scepticism as to whether a reason grounded on practical reason alone has motivational efficacy" (2006, p. 77). In contrast, "content skepticism" is a "Hegelian scepticism about whether Kantian formalism about practical reason yields any substantive conclusion" (pp. 76–77). A detailed discussion of Korsgaard's account would take us too far at this point, so I will simply focus on the way in which Thomas applies her insights. Now the idea behind the distinction, in Thomas's words, "is to argue that any scepticism about the pretensions of practical reasons must be a content based one and that motivational scepticism has no independent force" (2006, p. 77). He seems to be in agreement with Korsgaard about this, but disagrees about what it means for the Internal Reasons View. He presents Korsgaard as criticizing Williams for being a motivational skeptic, which in his view is a mistake, as Williams should be understood as having been "primarily" a content skeptic. According to Thomas, the independent plausibility of content skepticism allows us to refute the charge of begging the question.

Being a Kantian, of course, Korsgaard means to show that content skepticism fails, but as Thomas notes, she is not entirely against the Internal Reasons View, and especially in her later writings has showed sympathy for a Williams-like criticism of "dogmatic rationalism." Thomas writes:

> Korsgaard now no longer views it as acceptable simply to *stipulate* that rational agents are such as to be motivated by principles of reason. This seems to me to be a tactical withdrawal from the argument directed against the internal reasons theory in 'Skepticism about Practical Reason' and, indeed, to constitute a form of content scepticism in its own right. (2006, p. 81)

However, it seems to me that in order to present Korsgaard as a kind of content skeptic herself, Thomas has significantly widened the notion of

content skepticism compared to his previous definition of a skepticism about the implications of Kantian formalism. After all, the whole reason why Kantians want to keep their conception of reason *formalistic* is because they want to *avoid* being dogmatic in their rationalism. To be sure, the *term* "content skepticism" may seem to suggest a skepticism about any view that substantial content may be based on reason alone, especially if the term is meant in contrast to motivational skepticism. But in that case we should distinguish between two types of content skepticism: first, the skepticism about the possibility of pulling substantial content out of a purely formal conception of reason, and second, a skepticism about the plausibility of a conception of reason that is simply non-formal from the start and that may therefore include very substantial principles of reason to begin with. The former type of content skepticism is the one that Korsgaard means to prove wrong. Let's call it "anti-formal" content skepticism. The latter type is the skepticism about dogmatic rationalism, so let us call it "anti-dogmatic" content skepticism. Korsgaard subscribes to anti-dogmatic, but not to anti-formal content skepticism, while Thomas subscribes to both.

With this distinction in place, it should now be obvious that even if we reject anti-formal content skepticism, we need not reject the Internal Reasons View at all. After all, if the principles of reason are purely formal, then they need substance from elsewhere to be applied to, and the Kantian project is to transcendentally arrive at necessary conclusions about what reason would deliver regardless of what that substance from elsewhere would be like. In other words, Kantians may simply adopt the Internal Reasons View, and maintain that formal considerations of reason will yield certain deliberative results that can be arrived at from any subjective motivational set. That is why Williams thought of Kant's practical philosophy as the "limiting case" of the Internal Reasons View. As we shall see in the next chapter, however, anti-formal content skepticism does become relevant if we wonder whether the Internal Reasons View implies *relationalism*. The Internal Reasons View pushes us towards relationalism, a view that Williams seemed to endorse, but the Kantian project is an attempt to prevent that implication and reconcile an internal conception of reasons with a nonrelationalist view of practical judgment. It is for this reason, I think, that Williams referred to Kant's theory as a *limiting* case of the Internal Reasons View.

Rejecting anti-formal content skepticism is not only compatible with the Internal Reasons View, but conversely, *accepting* anti-formal content

skepticism also fails to help us refute the 'non-route-like' deliberation objection. After all, the external reasons theorist who claims that deliberation is not route-like, or that it does not have to commence from present motivations, can simply deny that reason has to be purely formal. If he claims that there are substantial principles of reason that deliberation can start from, then he does not have to squeeze the substance out of form, and thereby evades anti-formal content skepticism. What this move would make him vulnerable to, however, is the anti-*dogmatic* content skepticism, which Korsgaard and Thomas seem to agree on.

However, it seems to me that proponents of the idea that there are substantial principles of reason on which deliberations can be based independently of pre-existing motivations would rightly protest at being called "dogmatic" so easily. From their point of view, there might be very good reasons to think that reason is substantial, and any reasons they would give could hardly be refuted by objecting that those, too, depend on substantial principles of reason. This seems to be the sort of stalemate that Thomas has in mind when he remarks that each camp in the internal reasons debate accuses the other of begging the question.

Furthermore, it seems to me that if we want to provide a *defense* of the Internal Reasons View, then the burden of proof must at least *start* on our side of the divide, which gives the external reasons theorist's complaint of begging the question against that defense a certain merit if the stalemate cannot be avoided. Thus, it seems to me that anti-dogmatic content skepticism cannot stand on its own, at least not if standing on its own means it can do without independent support.

Personally, I think anti-dogmatic content skepticism is very plausible, at least with respect to deliberation about normative reasons, but I think it is plausible because of the line of argument from the first response, which is premised on the Motivational Continuity Thesis. However, that thesis seems a lot like a version of motivational skepticism, and according to the Korsgaard/Thomas view, motivational skepticism is besides the point. For the most part, this depends on whether any support for the Motivational Continuity Thesis would not ultimately presuppose anti-dogmatic content skepticism about normative reasons. I will return to this question in the next chapter.[9]

For now, however, note that Korsgaard's real issues seem to be not so much with the Internal Reasons View itself, but rather with a relational-

[9]See section 3.4.2.

ism that might be derived *from* the Internal Reasons View in *conjunction* with anti-formal content skepticism. Thus, her insight that motivational skepticism cannot by itself (i.e., without the help of content skepticism) establish such a relationalism may be compatible with my claim that a form of motivational skepticism *does* play a role in the defense of the Internal Reasons View itself—that is, in the form of which Kant is still a 'limiting case' rather than an opponent. In other words, motivational skepticism may not, by itself, harm Kantianism about practical deliberation, but that doesn't mean it cannot harm the External Reasons View.

In the next chapter I will discuss in further detail what the implications of the Internal Reasons View for meta-ethics might be, and on which additional premises those implications would depend. This will help us to relate the debate concerning internal reasons to the principles and distinctions from the previous chapter. Furthermore, I will also return to the question of whether the Motivational Continuity Thesis can be supported independent of assumptions about practical normativity. In particular, I will investigate whether this thesis might be defended on the basis of the Distinctness Principle from the previous chapter, and whether such a defense would be 'independent' in the required sense. Thus, whereas the current chapter has been focused on a discussion of the Internal Reasons View on its own terms, the next chapter will examine Williams's account from the point of view of the framework established in chapter 1.

3 *Internal Reasons, Relationalism, and Motivation*

In this chapter I investigate how Williams's Internal Reasons View, which I introduced and discussed in the previous chapter, may be related to the concepts and distinctions from chapter 1. The first two sections cover meta-ethical implications of the view to which Williams seemed sympathetic. In section 3.1 we combine the Internal Reasons View with the Facts and Authority Principles so as to make the view pertain to practical normativity as construed in chapter 1. Then, in section 3.2 I discuss the question of whether this comprehensive view would imply a relationalist interpretation of the Intersubjectivity Principle. This possible implication seems to be the central point of controversy surrounding Williams's account.

The final three sections of this chapter deal with the relation between the Internal Reasons View and the Distinctness Principle. I argue in section 3.3 that the two are at least compatible with each other, and that Williams does not 'revise' the Humean view. Then, in section 3.4, I claim that the Distinctness Principle can be used to defend the Internal Reasons View, by providing the independent support for the Motivational Continuity Thesis that, according to my diagnosis from section 2.5.1 in the previous chapter, is needed in order to refute the 'non-route-like' deliberation objection against Williams's argument. Finally, however, I will give a sketch of a similar defense on the basis of an anti-Humean theory of motivation in section 3.5, to show that my interpretation of Williams's defense is not committed to the exegetical claim that Williams himself had to be a Humean about motivation.

The purpose of this chapter, then, is to expand our understanding of the Internal Reasons View from the previous chapter in order to meet the following three criteria: first, to make it applicable to issues that were raised in chapter 1 concerning practical normativity; two, to make it defensible upon a Humean premise about motivation; and three, to do this

without reading the motivational Humeanism into Williams's own line of argument.

3.1 Implications for the Facts and Authority Principles

In the previous chapter I've elaborated on the Internal Reasons View and identified which premises we need in order to make Williams's defense of this view work. But why should we care about all this? So what, if reasons are internal? Even if Williams is right that the external interpretation applies to many people's reason statements in everyday life, why should it be a big deal if people ought to have meant their statements in the internal sense instead? By itself, this does not seem to be very interesting. However, the view becomes more significant if we explore its implications in conjunction with certain other assumptions that we may find plausible. In this section we take a look at the implications of the Internal Reasons View for the Facts and Authority Principles. In section 3.2 we turn to its implications for the Intersubjectivity Principle.

Everything that the Internal Reasons View says about *reason statements* will also apply to *practical judgments* if we adopt the Facts and Authority Principles from chapter 1. To be sure, the Internal Reasons View does not commit us to these principles. We have already seen that the *external* reasons view differs from the 'no reasons' varieties of externalism which deny the Authority Principle. Even though these two views may not exclude each other, defending both seems like overkill: if one would already think that we have reasons beyond the reach of deliberation from our actual dispositions, as the External Reasons View states, then what would be the point of postulating additional, further normative facts beyond the reach of those external reasons? Instead, it seems more plausible for 'no reasons' externalists to combine their externalism about practical normativity with the Internal Reasons View about reason statements, and for external reasons theorists to adopt the Authority Principle or a similar internalist criterion according to which reason statements—external reason statements, in their case—express practical judgments. This already shows that the Authority Principle and the Internal Reasons View represent two forms of 'internalism' that may be defended independently of each other. Nevertheless, combining the two is a much more viable option than the combination of their opposites, and several theorists in meta-ethics have understood the Internal Reasons View as a view about practical normativity

in the 'bottom line' sense, including Williams himself.

This does not mean that Williams understood the Internal Reasons View as a view about *morality*, however. As we have seen in section 1.4.1, some philosophers distinguish between practical normativity in the 'bottom line' or 'all things considered' sense on the one hand, and morality as a much more narrow or particular set of considerations on the other, which an agent may or may not have reason to care very much about. Williams refers to the former as the "practical *ought*," which "is to be taken to be equivalent to the 'all-in' or 'conclusive' answer to the question 'What ought I to do?'" (1981b, p. 119). In contrast, the latter—moral obligation—is understood by Williams as a particular form of the "general propositional *ought*," as he calls it, which is not established from the deliberative point of view. Williams is an internalist, in a manner similar to the Authority Principle, about the former, practical ought, which provides the agent with a reason, unlike the latter ought, which includes moral obligation. Hence, for Williams, the Internal Reasons View applies to practical oughts, even though it does not apply to moral obligations:

> In the practical or deliberative sense, '*A* ought to do *X*' will entail '*A* has a reason to do *X*,' in what I have called the 'internal' sense of that claim; the two are, however, not equivalent, since '*A* has a reason to do *X*' is not exclusive. (1981b, p. 120)

Note that in the above quotation, Williams is using the notion of having a reason in the non-resultant sense again. Since this non-resultance is the only consideration that he cites for the distinction between practical ought statements and internal reason statements, it seems safe to assume that if we switch back to reason-talk in the all-things-considered sense, then on Williams's view practical ought statements are equivalent to internal reasons statements. Note that this also commits Williams to the Facts Principle: if there are facts which make internal reason statements true, and internal reason statements are equivalent to statements that express practical judgments, then there must also be facts that make practical judgments true.[1] Therefore, I will from now on interpret Williams's arguments from the point of view of the Facts and Authority Principles. Given these

[1] Remember that my notion of 'facts' is so noncommittal as to merely rule out an extreme form of quietism about truth, and this quietism would be inconsistent, I think, with Williams's account of the truth of internal reasons, and moreover with Williams's views about truth elsewhere in his work.

two principles, any further implications of the Internal Reasons View must be implications for practical normativity as such.

With respect to the question of moral obligation, I suspect that the dispute about whether morality is practically normative involves a great deal of *verbal* disagreement over the use of *words* like "morality," "ethics," and "obligation." Even though Williams refuses to be an internalist about moral obligation, that does not mean he is committing himself to *sui generis* externalist metaphysical truths about morals. Rather, he is simply using the term "morality" to refer to norms that are essentially social, institutional, or perhaps even linguistic, in a manner similar to other varieties of the "general propositional *ought*," such as those of correct language use, or positive legal normativity, for example. However, this usage of the word "moral" is markedly different from the way in which I have used the word, for example, in section 1.3.2 in the first chapter, when I discussed the distinction between relationalist cognitivism and nonrelationalist *moral* realism. On my usage, which I think several philosophers share, "moral judgments" are simply full-blown, 'bottom line normative' practical judgments about *particular types of situations*, namely, those situations where one agent's acts may conflict with those of others. This characterization is admittedly sketchy, and I will return to this matter later. What matters for now is that unlike Williams, I use the term "moral" in such a way that every moral judgment is a practical judgment in the 'bottom line' normative sense, even if not every practical judgment is a moral judgment. Thus, on my view, if I agree that I have a certain obligation in some sense, but I disagree that I should, in the practically normative sense, fulfill this obligation, then I will deny that it is really a *moral* obligation. As I have already noted in section 1.4.1, I am not alone in this matter, as there are many philosophers who use the term "morality" in this practically normative sense. Now, if you are on our side in this verbal dispute, and if, furthermore, you should also subscribe to the Internal Reasons View and the Authority Principle, then your conclusion is going to be, *pace* Williams, that the Internal Reasons View applies to morality as well.

Finally, as we have seen in section 2.4, Williams's defense of the Internal Reasons View is premised on a proceduralism about reasons for action, and if the Facts and Authority Principles are true, then this proceduralism about reasons is equivalent to the proceduralist interpretation of the Disconfirmation Principle. Thus, it seems that if we combine Williams's defense of the Internal Reasons View with the Authority Principle, then

both the Facts Principle and the proceduralist version of the Disconfirmation Principle come along automatically as a kind of 'package deal.' Let us refer to the combination of these four items as the "Comprehensive Internal Reasons View," or CIRV. This view offers a picture of practical normativity as being a matter of fact about the motivational characteristics we would have if we took the path of sound deliberation from our current motivational characteristics. Even though CIRV depends on my preferred formulations from chapter 1, it seems sufficiently close to what Williams had in mind.[2]

However, what about our other two Principles: those of Intersubjectivity and Distinctness? Let us first take a look at the implications of CIRV for the dispute between relationalist and nonrelationalist accounts of the Intersubjectivity Principle. We shall return to the Distinctness Principle in section 3.3.

3.2 Implications for the Intersubjectivity Principle

As we have seen in chapter 1, the most straightforward way to account for the Intersubjectivity Principle is to adopt a nonrelationalist view of the content of practical judgments. However, if we also adopt the Facts and Authority Principles, then it follows from nonrelationalism that there are normative reasons for action which do not depend on any particular features of the specific, actual agents for which they *are* reasons. And the most straightforward way to account for this implication, it may now seem, is to adopt the External Reasons View. In fact, it might even seem that this implication simply *is* the External Reasons View. But that would be too quick, because there might be ways in which an internal reasons theorist could account for nonrelationalist normative reasons as well. Again, it all depends on what our premises are going to be. Let me explain.

As I see it, there are two ways in which one might try to combine CIRV with nonrelationalist realism. The first is to argue that even though different agents must start their deliberations from their own respective motivational starting points, which might differ very much in terms of their content, it might still be true that all agents would end up with the same ends after sound deliberation, simply because the features of rationality

[2] With one important exception, which is that I will treat CIRV as not including Williams's proximity requirement with respect to facts that are relevant to instrumental deliberation, as I have argued in section 2.2.1 in the previous chapter.

are such that, on the basis of *any* contingent motivational set, deliberative transformations would lead towards the same motivational dispositions eventually. Let us call this the "convergence strategy."

The second strategy would be rather to deny that the motivational starting points that agents actually have could be so contingent from the perspective of rational agency that no final ends will be featured in *every* agent's *S* from the start. If we deny this, so that there *are* final ends that we can place in every agent's *S* from the start, then any deliberative outcome based on such elements might therefore be nonrelationalistic. Usually, the idea behind such an approach is that we can place these elements in every agent's *S* because they are *constitutive* of agency: without them, one could not be an agent. Let us therefore call this the "constitution strategy."

3.2.1 The Convergence Strategy

Ingenious as both strategies may sound, Williams didn't buy either of them. The problem for the convergence strategy is that it cannot simply allow the method of deliberation to make the different motivational sets converge by presupposing *substantial* principles of rationality that simply rule out various motivational elements by being inconsistent with them. Because such an approach would already presuppose a 'non-route-like' conception of deliberation, which means that the view defended would be an External Reasons View. Instead, in order for the convergence strategy to yield an Internal Reasons View, it must be shown how deliberation, by operating *upon* the different elements already in the various motivational sets of wholly different agents, transforms those different sets into a common direction, bringing them closer together and ultimately making them exhibit shared attitudes on moral issues.

There are two ways in which one might try to make the idea of convergence plausible. The first way is to argue that practical reason is a pursuit of knowledge that, as far as its objectivity is concerned, may be understood as relevantly similar to other areas of knowledge in which we'd expect attitudes to converge when agents become more knowledgeable.[3] On this version of the convergence strategy, the converging motivational attitudes are essentially *beliefs*, or if they are not strictly identical to beliefs,

[3] I take it that this is more or less Thomas's view. He maintains that the objectivity of scientific knowledge is secured by epistemological contextualism, and then tries to argue that this contextualism also provides the key to defending objective knowledge in the field of ethics.

then their convergence is *driven* by the convergence of beliefs in a manner that requires an anti-Humean theory of motivation, and will be incompatible with the Distinctness Principle from chapter 1. Nevertheless, this anti-Humean approach is certainly compatible with the Internal Reasons View and the 'route like' conception of deliberation, because it may be argued that in certain areas of knowledge, if not all of them, advancing our knowledge is only possible by building upon our current beliefs. Even physical science seems to be like that, after all, since empirical observation is "theory laden" and the revision of our scientific theories should be thought of, in Neurath's famous metaphor, as the improvement of a boat while being already at sea. Given this general picture of belief-revision, there is no reason why the anti-Humean should not allow that new beliefs in the field of practical reason must be formed in a manner that continues upon old beliefs as well.

Furthermore, some proponents of this approach may argue that in the field of practical reason and ethics especially, there are no ethically neutral concepts in order to evaluate different ethical points of view with. Nevertheless, if theory-ladenness does not undermine nonrelationalism in physical science, then why should it give us reason to be relationalists in ethics? From their different theory-laden perspectives, scientists are still pushed towards the same direction by the objective facts that they are studying, and which underlie their observations. By analogy, certain moral philosophers have proposed the idea of "moral perception," a way to arrive at new practical beliefs about moral facts by 'perceiving' them from a perspective that is constituted by one's pre-existing attitudes and concepts.

At this point, however, we may start to worry that the analogy between moral knowledge, on the one hand, and scientific or everyday empirical knowledge on the other hand, has been taken too far. The Internal Reasons View was supposed to be a view specifically of the nature of practical reason, so it would be strange if all its implications would amount to no more than that which holds for belief formation generally. Williams's point was not that practical reasons are internal in a sense in which reasons to believe in the theory of evolution, say, would be 'internal' as well. Even if belief formation in other areas of inquiry would follow a pattern of continuous, 'route like' deliberation as well, then it still seems that the role of empirical evidence in natural science allows substantially more content to be *imported* into our belief sets during this process of inquiry

than the Internal Reasons View would allow in the field of practical reason. If practical deliberation could incorporate facts about normative ends that would be just as 'external' to us as the facts of empirical science, so to speak, then we should simply subscribe to the External Reasons View.

This means that this version of the convergence strategy must accomplish two things. The first is to show that other areas of inquiry are not as different from practical reason as we might have thought. But the second thing is to nevertheless also do justice to the spirit of the Internal Reasons View that practical reason is 'internal' in a manner that is dis-analogous to theoretical reason. In order to satisfy both requirements, Alan Thomas has tried to develop a notion of the 'distinctiveness' of internal reasons that is, if I understand it correctly, meant as something different from my notion of relationalism, and therefore compatible with nonrelationalism. Roughly, the view is as follows: skepticism about nonrelationalist moral knowledge is based on the idea that, whereas scientific knowledge can be justified on the basis of some kind of objective foundation (empirical evidence, sense-data, or whatever), there is no such foundation to be had in the field of ethics. However, the idea that scientific knowledge should be based on such a foundation has been challenged by contextualist and inferentialist developments in epistemology. Now, the features that inferential contextualists allude to in order to give their alternative account of nonrelationalist objectivity *are* also available in the field of ethics, or so Thomas means to argue. The difference between the relatively 'external' reasons of empirical science and the internal reasons of practical deliberation, however, is that an internal reason is *distinctive* of the particular agent for which it is a reason, in a manner in which the reason to believe that the earth is four billion years old would not be distinctive. It is this feature, Thomas argues, that helps to distinguish between the Internal Reasons View and the External Reasons View from a motivationally anti-Humean perspective.

In order to illustrate this, Thomas compares his own account to McDowell's account of practical reason. Both are anti-Humeans about motivation, both subscribe, in some form or another, to the metaphor of 'moral perception,' and both defend a nonrelationalist realism about the content of practical judgments. However, for McDowell, this means that belief revisions due to practical deliberation can be much more 'discontinuous' than the Internal Reasons View allows, because the justification of practical judgments is to be found in the "space of reasons" that the agent participates in,

rather than in any psychological idiosyncrasies that are distinctive of the particular agent. Hence, an agent may come to 'see' that he has to adopt a practical judgment which, even if it presupposes some of his previously held beliefs, does not continue upon any of his attitudes that were peculiar to *him* as an individual. In this respect, McDowell understands 'seeing the truth' in ethics in a manner that might be compared to 'seeing' the truth in, say, mathematics.

But for Williams, mathematical truths do not have the power to move us in the way that truths about our reasons for action can move us, and this difference must be explained by assuming that true practical reason statements must be internal in a manner which true mathematical statements are not. Thomas sides with Williams on this point, but he thinks that this insight can be incorporated in a more or less McDowellian scheme by stressing the fact that all psychological ascriptions are governed by normative principles constitutive of the space of reasons. What makes practical reasons distinctive, on Thomas's view, is that our knowledge of them is perspectival in a way that gives our thick ethical concepts a much more constitutive role with respect to those reasons than the concepts we use to do scientific research, say.

Now I must admit I do not see immediately how this argument against McDowell is supposed to work. Surely, insofar those thick concepts are constitutive of the space of reasons, they are no longer distinctive of any particular agent operating within that space, and insofar they are distinctive of any particular individual, they would no longer be constitutive of the space of reasons that all agents are bound to. Perhaps what Thomas really means to say is that on the one hand, internal reasons are not distinctive of us as individuals, but rather of us as human beings—as a species, and that on the other hand, the space of reasons is not an *a priori* entity that all conceptually possible agents would have to participate in, but rather a specifically human development. The thick concepts constitutive of *our* space of reasons are, in Thomas's own terminology, "relativised *a priori*"—they are conceptually necessary presuppositions of our current contingent practice, not of every conceptually possible practice.

Such a line of argument would seem to do justice to Thomas's Aristotelian outlook, and in fact, I am fairly sympathetic to such a proposal, but I cannot help but conclude that it would no longer be nonrelationalist. Because strictly speaking, it would imply that the truth conditions of practical judgments always carry a 'for human beings' subscript, and that in order

to determine whether such a judgment applies to a particular individual, we must know whether that individual is human, and that would be a particular fact about the individual that all conceptually possible agents may not share, which means that the content of practical judgments is essentially going to be relationalist.[4]

We will return to the problems of relationalism later on, to be sure, and it might be that Thomas's preference for "perspectival" moral knowledge over relationalist moral knowledge is merely a matter of semantics—of how to construe the content of practical judgments as we communicate and exchange those judgments in our linguistic practices. But as we will also see later on, I am not at all committed to the idea that relationalist subscripts have to be explicit in our judgments as we exchange them in those practices, and I think a 'quasi-nonrelationalist' semantics can be built upon a relationalist cognitivist metaphysics in a manner that might, in the final instance, amount to a view that is very similar to Thomas's proposal (with the key difference being that of my Humeanism vs. his anti-Humeanism about motivation).[5] I shall return to these issues later on.[6]

In any case, Williams did not even seem to think that internal reasons would have to be similar among all human beings in this sense, but if Thomas is right that Williams may not have been a Humean about motivation, then the reason why Williams was skeptical about such an

[4]Perhaps this objection might be circumvented by replacing the "for humans" subscript with a "for participants in our practice" subscript, such that whenever outsiders to our practice were encountered, they could be incorporated into a new, more encompassing practice, with the possible consequence that the constitutive principles of this new practice might change. Then it might be argued that the only objectivity we should aim for in ethics is the objectivity that all agents we'll ever encounter in the real world should submit to, rather than an objective validity that would range over physically impossible agents in remote *a priori* possible worlds. This would be a more 'Hegelian' rather than 'Aristotelian' line of argument, I suppose. However, while I do see how commonalities across human beings might provide substance for shared values, I do not at all see why mere physical possibility, or the difference between agents that we will and agents that we won't actually meet, are going to provide any reason whatsoever for making the possibility of shared thick concepts plausible. In fact, this line of argument simply seems to *presuppose* that (1) every agent can be incorporated in our practice and (2) our practice must always be based on constitutive ethical concepts. But accepting (1) might be a reason for rejecting (2), and vice versa, so the argument has not shown anything.

[5]Perhaps I could even subscribe to a 'quasi-anti-Humean' semantics of our everyday practice of giving and asking for reasons, and show that it is compatible with my Distinctness Principle. This idea will have to await another occasion, however.

[6]See section 10.5.4.

attempt to secure nonrelationalist ethical knowledge was not due to its anti-Humean motivational underpinnings. Instead, Thomas looks for arguments against nonrelationalism elsewhere in Williams's work. In particular, it is Williams's thought experiment about the "hypertraditional society" that would contain the key argument against nonrelationalism. Thomas means to show that whereas Williams's defense of the Internal Reasons View succeeds, his hypertraditionalism argument against nonrelationalism does not, because the latter presupposes the foundationalism that Thomas has rejected on contextualist grounds. Therefore, a nonrelationalist version of the Internal Reasons View, premised upon an anti-Humean theory of motivation and an inferential contextualist epistemology, can be defended. A discussion of the hypertraditionalism argument will be beyond the scope of this thesis, as is the discussion of the merits of inferential contextualism, but since my purpose will be to defend an account based on motivational Humeanism, I shall simply flag the anti-Humeanism as the *conditio sine qua non* for any 'belief-based' version of the convergence strategy.

However, there is a second way in which the convergence strategy might be made to work, and which is neutral about the Humean theory of motivation. This second approach is to show that we can pull the convergence out of a purely *formal* notion of rationality.[7] On such a view, no substantial information is 'perceived' or in any other way added to the agent's *S* from the outside, as it were, by the process of deliberation, but instead it is merely by reflecting on the formal consequences of the elements already in the *S* that a convergence among agents with different sets can be expected. However, even though formal constraints can certainly lead to various sorts of revisions in an agent's motivational set, it is hard to see how any substantial final end could in principle be ruled out, or ruled in, as the outcome of correct deliberation when there are no limits whatsoever on the content of the motivational set that the agent might start out with. This is the aforementioned worry of anti-formal content skepticism: skepticism about the possibility to pull substance out of form. In chapter 5 I will return to the prospects for the formalistic convergence strategy.[8] For now, we can flag the thesis of anti-formal content skepticism as a premise that rules this strategy out.

[7]This view is sometimes associated with Michael Smith's position, although as we shall see in chapters 5 and 6, Smith's view might actually be better understood as a nonproceduralist account, in which case it is not a version of CIRV.

[8]See section 5.4.1.

3.2.2 *The Constitution Strategy*

If unlimited variety in the motivational sets from which agents might start their deliberations would make nonrelationalist ends impossible, then the only other way to get the nonrelationalist ends is by placing limits on the possible deliberative starting points, as the constitution strategy attempts to accomplish. After all, who ever said that it is in principle possible for agents to desire everything? Or why should we assume, conversely, that there could be no motivational elements that every agent, in virtue of being an agent, must already posses?

In fact, Williams allows that there could be such elements. The interest that every agent has in finding out about the facts that are relevant to his instrumental deliberations is just such an element.[9] The very idea of being a goal-directed agent presupposes having this interest. However, Williams also claims that we can argue *why* all agents must have this interest: if agents would not desire to know the means to their ends, their actions would not be connected to their ends in means-end fashion and they would fail to be agents in the first place. And so it should be for any other interest that we might want to place in every agent's *S* regardless of the contingent features of each particular agent: we are only allowed to do this if we can argue on conceptual grounds that an agent would fail to be an agent without having that element in his *S*.[10]

Given that the interest in knowing the means to one's ends is hardly the sort of thing that will give us substantial, nonrelationalist moral ends such as those of altruism, Williams essentially employs a burden of proof argument against the constitution strategy. He invites his opponents to come up with such an argument, and concludes that no satisfactory proposals have been done so far.

However, I personally think a much stronger argument can be given, which I will discuss in more detail later on. Briefly put, the argument is as follows: if the interest in knowing the means to one's ends is to serve as a kind of 'existence proof' of universal motivations that are constitutive of

[9] "any rational deliberative agent has in his *S* a general interest in being factually and rationally correctly informed" (1989/1995, p. 37).

[10] "Somebody may say that every rational deliberator is committed to constraints of morality as much as to the requirements of truth and reasoning. But if this is so, then the constraints of morality are part of everybody's *S*, and every correct moral reason will be an internal reason. But there has to be an argument for that conclusion. Someone who claims the constraints of morality are themselves built into the notion of what it is to be a rational deliberator cannot get that conclusion for nothing." (Williams, 1989/1995, p. 37)

being an agent in the first place, then we should expect all such interests to be self-referencing in the same manner that the interest in having true beliefs about the means to one's ends is self-referencing. For note that agent *A*'s interest in having such knowledge is essentially a desire that it be the case that *A* has true beliefs about the means to *A*'s ends. In contrast, *B*'s interest concerns the beliefs of *B* about the means to *B*'s ends. Recall the distinction between practical beliefs that are *similar* and those that are *isomorphic* in section 1.3.2 of the first chapter. We can apply this same distinction to desires, or motivating states generally. It turns out that the aforementioned interests of *A* and *B* are *isomorphic*, because they have the same structure from the indexical, first person point of view, but they are not at all *similar*, because they need not be directed towards the same states of affairs. In fact, there might be cases where in order for *A* to find out the facts about the means to his ends, *A* must act in ways that happen to prevent *B* from finding out the facts about the means to her ends.

But as we have already seen in section 1.3.2, nonrelationalist realism implies that our true practical beliefs must be *similar* to each other, not (or not necessarily) isomorphic. Hence, the constitutive interest in finding out about the means to one's ends is universally shared in the wrong way, and even if there would be other interests that had this same, isomorphically understood universal presence, they could never provide us with nonrelationalist moral values because they do not satisfy the similarity condition across agents.

However, there is one further possibility that we should consider. It may be that even if the constitution strategy cannot deliver similarity on its own, it can nevertheless be used to give the idea of a formal rationality-based convergence a head start, as it were. We have reasons to doubt that formal rationality could squeeze substance *out* of an unlimited range of starting points, and we have seen that constitution considerations are going to have a hard time placing nonrelationalist substance *into* the starting points, but perhaps the constitution strategy can limit the range of starting points in such a way that the convergence strategy *can* develop nonrelationalist substance *from* it using exclusively formal methods. In other words, perhaps the two strategies can compensate for each other's shortcomings. This is, very roughly, what I think Korsgaard tries to accomplish. I will return to this idea in chapter 5.[11]

However, for now, let me note that with respect to this combined strat-

[11] See section 5.4.3.

egy, a burden of proof argument still makes sense. On the combined account, what formal rationality has to accomplish is a move from isomorphic states to similar states, and that still seems quite a daunting task. *Prima facie*, it is not at all plausible that this can be done. Essentially, this worry is still the worry of anti-formal content skepticism, admitting that "Kantian formalism" may include transcendental considerations about the motivational elements that all agents must *actually* have on the pain of no longer being agents, but insisting that due to the essentially isomorphic structure of such considerations, convergence onto similarity is still not to be expected. From now on, I mean to use the notion of anti-formal content skepticism in this transcendental sense.

With these explications made, it now follows that in conjunction with anti-formal content skepticism (which rules out Korsgaard's formalist combined strategy) and the Distinctness Principle (which rules out Thomas's anti-Humean strategy), CIRV implies relationalist cognitivism. It may be that Williams (who used the term "relativism" instead) was a relationalist for this same reason, or it may be that even though he was an anti-Humean, as Thomas suggests, Williams had an additional argument that would block the convergence strategy, such that the relationalist implication of CIRV would still follow through. This is an exegetical matter about which I want to remain neutral. But in any case, the relationalist implication is the major one, meta-ethically speaking, the one that everybody worries about. If CIRV implies relationalism, then how is it going to account for the Intersubjectivity Principle? This question will play an important role in the chapters to come, since my own view is going to be relationalist as well, and for more or less similar reasons.

There is now one final loose end from chapter 2 that we must take care of, and that is the defense of the Motivational Continuity Thesis, which must receive independent support in order to ward CIRV against the 'non-route-like' deliberation objection. Again, we shall have to keep the division between Humeans and anti-Humeans about motivation in mind. However, as I shall argue in the following sections, the thesis can be defended on both types of accounts.

3.3 No Revision of Humeanism Is Needed

In the remainder of this chapter I mean to do three things. First, I will show that it is a mistake to think of the Internal Reasons View as containing an

'update' or 'revision' of the purely instrumental Humean theory of motivation. Instead, I will argue that the two are fully compatible: nothing in the purely instrumental account of the belief-desire structure of motivating reasons, as captured in my Distinctness Principle, needs to be altered in order to be able to combine it with the Internal Reasons View.

Second, I will argue in section 3.4 that the Distinctness Principle is not merely compatible with the Internal Reasons View, but that it also provides the kind of independent support for the Motivational Continuity Thesis that we need in order to make the defense of the Internal Reasons View non-question begging.

Third, however, I will argue in section 3.5 that motivational continuity can *also* be defended within the framework of motivational anti-Humeanism, provided that certain further assumptions about motivation are made. This means I will be in partial agreement with Thomas concerning Williams's defense of the Internal Reasons View. I agree that the defense is not exclusive to motivational Humeanism, but I disagree that it does not require any motivational premises at all and that it could be based on content skepticism alone. As I have argued in section 2.5.2, anti-dogmatic content skepticism does rule out external reasons, but it cannot be defended independently from motivational considerations, while anti-formal content skepticism is not directed against external reasons in the first place, but rather against the formalist strategy required to make the Internal Reasons View nonrelationalist. However, let us first turn to the matter of the compatibility between Humeanism and the Internal Reasons View.

3.3.1 Motivational vs. Metanormative Humean Theories

For starters, note that if the distinction between normative and motivating reasons is sound, then that means that when people talk about the "Humean" theory of reasons for action, there are actually two very different claims that they might have in mind. Michael Smith has exploited this insight by advertising his view as a combination of a "Humean theory of motivating reasons" with an "anti-Humean theory of normative reasons" (1994, p. 130). This rhetorical way of putting things is certainly elegant, but it is also misleading, because it suggests that there is this particular feature that practical reasons might have—the property of being 'Humean'—which, according to Smith, motivating reasons have, and normative reasons do not

have. This is misleading because, at least in contemporary usage of these terms, when moral philosophers speak of "Humeanism" in the context of motivating reasons, they usually have a wholly different feature in mind from the feature that is considered "Humean" with respect to normative reasons.

When a theory is Humean about motivation, then according to the theory, a motivation to ϕ requires a *means-end* relation between ϕ-ing and some intrinsic desire of the agent. This is the idea that I have meant to capture in the Distinctness Principle (in the context of our current discussion about different forms of Humeanism, I will also refer to this view as the "motivational Humean" theory or simply "motivational Humeanism"). In contrast, when someone is considered a "Humean" with respect to normative reasons (a view to which I shall refer as the "metanormative Humean" theory or "metanormative Humeanism"), we usually only mean to say that he claims there to be an essential dependence between any normative reason that an agent has and his contingent motivational attitudes. In the context of cognitivism, this is basically a combination of relationalism and the Authority Principle.[12] But in order to qualify as a "Humean" about normative reasons, one need not think that the dependence is a means-end relation, nor that the attitude depended on has to be an intrinsic desire.

This dis-analogy between the meaning of Humeanism with respect to the normative and the motivational is perfectly logical, too. For suppose that normative reasons were to be analyzed in terms of the same belief-desire criterion as motivating reasons. Then by necessity, normative and motivating reasons would coincide, with the absurd consequence that one could never act against one's normative reasons. Lacking in self-governance would be impossible. I can think of no philosopher who would defend such a view, and even though I am not a Hume scholar, I doubt this was what Hume had in mind. Moreover, I also doubt that by calling his theory of normative reasons "anti-Humean," Michael Smith merely meant

[12]Of course, noncognitivists can also be Humeans about practical normativity, and I suppose they usually will be. However, strictly speaking they would not be Humeans about "normative reasons" according to the definition of such reasons from section 1.4.2, i.e. about truths that determine self-adopted reasons for action once they become known to the agents whose actions they are reasons for. According to the noncognitivist, after all, there are no such truths to be known. But what would make a noncognitivist a "Humean" in the metanormative sense, I suppose, is a claim to the effect that our deliberations must be geared towards ends that are constituted by our particular, contingent noncognitive attitudes, and that there will be no convergence from such attitudes across all possible agents.

to assert his belief in the possibility of weakness of will. Instead, what Smith means when he calls his theory of normative reasons "anti-Humean" is that he is a nonrelationalist about those reasons.

In similar fashion, when Schroeder speaks of the "Humean Theory of Reasons" (2007, p. 2), which he distinguishes from the "Humean Theory of Motivation" (p. 7), he means the claim that any "objective normative reason" (pp. 12–15) must be explained with reference to some psychological feature of the agent.[13] As Schroeder construes it, this theory is essentially a "parity thesis" which claims that all normative reasons depend on attitudes in a manner structurally similar to the manner in which reasons concerning matters of personal preference depend on attitudes of personal preference.[14] Now Schroeder is sympathetic to what he calls "hypotheticalism," the idea that the attitudes referred to by the Humean Theory of Reasons must be desires, but this is largely due to a very liberal definition of desires (similar to the one I have given in section 1.5). However, Schroeder defends himself against the charge of being committed to "instrumentalism," the aforementioned radical view that the relation between normative reasons and the desires they depend upon must be instrumental (pp. 179–191). That is simply not what being a Humean about normative reasons means.

Note that this also makes it valid to count Williams as a Humean about normative reasons, regardless of whether he was a Humean about motivation. The fact that Williams favored a relationalist version of the Internal Reasons View makes him a Humean about normative reasons, and since metanormative Humeanism is not restricted to an instrumental

[13] In Schroeder's terminology, however, the "Humean Theory of Motivation" is not understood as a theory of motivating reasons, because Schroeder uses the notion of a "motivating reason" in a different sense. In his usage, something counts as a motivating reason when it is both a "subjective normative reason" and an "explanatory reason" for an action (p. 14). His notion of a "subjective normative reason" is identical to what I have called a "self-adopted reason" in section 1.4, and the "explanatory reason" comes closer to what I am calling a "motivating reason." The advantage of Schroeder's terminology, I must admit, is that it preserves a certain symmetry between the notions of motivating and normative reasons that my own terminology lacks. However, the disadvantage is that in Schroeder's terminology, all cases of acting against one's better judgment are going to be cases in which the explanatory reason involves being motivated by a desire without this counting as a "motivating reason," which seems rather artificial and a counter-intuitive thing to say. In any case, what matters is that the "Humean Theory of Motivation" in Schroeder's framework is still dis-analogous to the "Humean Theory of Reasons" in the manner that I have made explicit here.

[14] In Schroeder's leading example, the fact that there will be dancing at some party is a reason for the protagonist, Ronnie, to go to that party, and the fact that this is a reason for Ronnie depends on the fact that Ronnie loves to dance.

dependence relation, Williams's notion of the deliberative route does not break with the Humean tradition in this respect. Furthermore, if Williams's arguments in favor of a relationalist Internal Reasons View were indeed independent from motivational Humeanism, then we can think of Williams's account as the diametric opposite of Smith's view: whereas Smith has been trying to show that he can be a Humean about motivation without being Humean about normative reasons, Williams was arguing for being Humean about normative reasons without (necessarily) being Humean about motivation. In any case, the fact that Williams is not revising metanormative Humeanism does not mean that he has nothing to offer to Humean thinking: what he does have to offer is a way of understanding how, on a Humean conception of normative reasons, such reasons could still be different from our self-adopted reasons.[15] Furthermore, even if Williams may not have been a motivational Humean himself, we shall see in the next chapter that the Internal Reasons View may also help us to solve the Facts Problem and explain how Humeanism about both motivation and normativity may be combined.

3.3.2 *The Instrumental Character of the Distinctness Principle*

Of course, one might still wonder why motivational Humeans *would* want to think that the dependency relation must be instrumental in the case of *motivating* reasons. Surely, people are motivated by the results of their non-instrumental deliberations as well? In order to avoid any further confusion, let me switch back to my own terminology from chapter 1, and speak of the Distinctness Principle in the context of motivating reasons and of relationalism in the context of normative reasons (provided that we accept the Authority Principle). Now the question is: if relationalism about normative reasons does not even presuppose a strictly means-end dependency relation, then why formulate a Distinctness Principle about motivating reasons that does involve such a strict relation? Or to repeat my previous, rhetorical question: surely people are also motivated by intentions that were formed by non-instrumental deliberation?

The answer, as far as I'm concerned, is that the Distinctness Principle is simply not a principle about deliberation in the first place. It is a principle about the limitations of *belief*, namely, that beliefs cannot motivate in their

[15] In the words of Schroeder: "One of the main contributions of Williams's seminal paper was to point out that it is possible to give a Humean account of *objective* normative reasons" (p. 13, n. 18).

own right. This means that whenever we are motivated, there is something in addition to our beliefs, and this something we call intrinsic desire. Now the crucial thing to understand is that the Distinctness Principle is a claim about the *synchronic* structure of an agent's attitudes. In order to describe this structure, the intrinsic desires of the agent are simply *defined* as the ends that we must attribute to the agent in order to understand her current motivation as directed towards what, according to her beliefs, are the means to those ends. Therefore, motivating reasons explain actions in a purely instrumental manner *by definition*.

But that does not mean that the Distinctness Principle cannot accommodate non-instrumental deliberation as something that we may refer to in order to explain what people do, because deliberation is a *diachronic* process, not a *synchronic* structure. This applies even to instrumental deliberation, which means that mental acts of instrumental deliberation are not identical to the things we attribute when we attribute motivating reasons, and I think that actually makes perfect sense. What we attribute when we attribute a motivating reason is, essentially, the *intelligibility* of the agent's behavior as an action that satisfies the requirements of instrumental rationality. On the assumption that the behavior is indeed an action in this sense, we may then look for a diachronic explanation of the action by attributing an instrumental deliberation on the basis of *prior* beliefs and intrinsic desires that, insofar the agent actually *was* behaving instrumentally rationally, will be equivalent to the beliefs and desires that constituted the motivating reason. Finally, we may explain the existence of the prior intrinsic desires by processes that produce and alter intrinsic desires, which include non-instrumental deliberation, but also processes that generate intrinsic desires that are at odds with our practical deliberations and undermine our self-governance, such as addiction, weakness of will, and the like.

So even though the 'route-like' character of the Motivational Continuity Thesis includes non-instrumental deliberation, that does not at all make it a revision of the synchronic Distinctness Principle, which merely defines intrinsic desire as a construct related in means-end manner to motivation in order to explicate the motivational inertia of belief. Nevertheless, the Distinctness Principle does put some pressure on the idea that certain changes amongst our intrinsic desires across time would be instances of deliberative correctness while others wouldn't. If there are truths about how our intrinsic desires should be changed, then getting those truths

right would be a matter of belief, which seems to undermine the fundamental independence that intrinsic desires should enjoy with respect to the supposedly inert sphere of belief. This apparent inconsistency ought not to surprise us, however: it is simply a reformulation of the Disconfirmation Problem from section 1.5.2. The solution to this problem will have to be deferred until later in this thesis, though: I turn to this issue in chapter 7.

Nevertheless, there is an important point to note right now: if the Disconfirmation Problem could not be solved, then this might be a reason to convert to anti-Humeanism and reject the Distinctness Principle, but it does not seem to be a viable alternative to simply try to keep the Principle by adding non-instrumental rationality to it. There are two arguments for this. The first argument is that it would make it impossible for the Distinctness Principle to explain various forms of non-self-governed agency. The kleptomaniac is driven by an intrinsic desire that explains his actions in an instrumental fashion, but we could no longer explain his motivation if motivation would require intrinsic desires to satisfy further constraints of non-instrumental rationality, or to have their origins in non-instrumental deliberations upon prior intrinsic desires, since the crucial thing about kleptomania is that it does not meet such criteria. The Humean Theory of Motivation must account for all actions, not just self-governed actions.

The second argument is that any revision of the Distinctness Principle that would replace the instrumental relation between intrinsic desire and motivation with a non-instrumental relation, would thereby simply have changed the *meaning* of the term "intrinsic desire," such that the new formulation would not really contradict the old formulation. What this means is that even if we would reformulate the Distinctness Principle so as to incorporate non-instrumental rationality, we could then simply re-introduce a notion of "intrinsic* desires," or "desired ends" perhaps, which we could define, again, as the ends that are related in strictly means-end fashion to the motivations of the agent. It seems that what our current Distinctness Principle says would still hold with respect to this re-introduced notion.

Now perhaps a 'revisionary' motivational Humean might want to accept this, and simply claim that the Distinctness Principle is to be *extended* by making the distinction between "instrumentally intrinsic desires," say, and what we might call "ultimately intrinsic desires," such that in order to explain how agents are motivated, beliefs can play two explanatory roles: they can explain the relation between the motivations and the

instrumentally intrinsic desires, but they might also, furthermore, explain the relation between the ultimately intrinsic desires and the instrumentally intrinsic desires. Such a view would keep the Humean intuition that beliefs cannot motivate in the absence of desires, while incorporating the idea that beliefs about what would be rational in a non-instrumental sense can play a role in explaining the agent's motivation.

I am sympathetic to this idea in principle, but it seems to me that because of the need to account for defects in self-governance, this expanded version of the Humean theory of motivation would not really constitute a revision of the Distinctness Principle, either. After all, there might be all sorts of behavior to which we can attribute intelligibility as agency that satisfies instrumental rationality, but which would not be intelligible on the assumption that the instrumentally intrinsic desires thus attributed reflected certain further, ultimate intrinsic desires in the light of the agent's beliefs about non-instrumental rationality. However, it seems that the weaker intelligibility attribution is already *sufficient* to attribute beliefs to the agent.

Now one might object that unless the agent would at least sometimes exhibit behavior that did satisfy the stronger type of intelligibility, we would not have reasons to attribute the non-instrumentally relevant beliefs to the agent. And this might help to distinguish between agents in a weak sense, like spiders, and agents in a strong sense, like human adults. However, the only beliefs which we may cite in order to promote the 'level of rational agency' in such a way would be beliefs about how to be a rational agent that, in Williams's sense, would 'come for free' in the same sense in which the logic of means-end rationality 'comes for free.' And all such concerns would essentially be concerns of self-governance: they correspond to the isomorphically universal interests that are constitutive of being an agent. And thus we are back at the first point about self-governance: any 'expanded' version of the Distinctness Principle which would incorporate non-instrumental notions of rationality would no longer be explicating a Humean theory of motivation, but rather a Humean theory of self-governed motivation. But there is nothing in the original Humean theory of motivation that explicitly denies that the requirements of self-governance might include certain beliefs. And even if there were, then it now seems we could simply reduce these beliefs to instrumental beliefs, by adding the interest in being self-governing as an instrumentally intrinsic desire constitutive of being a self-governing agent. Again, all this stuff is

okay as long as it 'comes for free' with the concept of agency, and when it does it won't really contradict the Distinctness Principle.

In contrast, consider now a principle of non-instrumental rationality, which says that whenever you desire *P*, then reason requires you to also desire *Q*, and which does not come 'for free' in the aforementioned sense. All parties agree that this would be a principle of substantial reason, which could not possibly be derived from formal considerations alone. Because of this, it seems that violating such a principle should no longer be considered a failure of self-governance by itself. Instead, the principle seems to be a matter of substantial belief: an agent who desires *P*, but does not himself believe in the principle, may be called unreasonable for not desiring *Q*, but it does not make him less of an agent in any sense. This means that, if we think that there are such principles, we cannot square them with the unrevised Distinctness Principle in the manner just described. However, let us now suppose that the agent does believe in the principle, and he desires *P*, but yet he does not desire *Q*. Now the agent does seem motivationally deficient: under conditions of full self-governance, we would expect him to desire that which he himself believes he has reason to desire. But the only way to 'revise' the Distinctness Principle in order to accommodate this, it seems to me, would be to simply give up on the motivational inertia of beliefs. For even though the belief in question cannot motivate on its own, as it depends on the presence of a desire that *P*, the belief is nevertheless supposed to bring about the desire that *Q* when no governance-undermining factors intervene, and to do this without the help of either a formal connection between *Q* and *P* nor that of an additional, intrinsic desire that *Q* when desiring *P*. Allowing this would amount to endorsing anti-Humeanism.

However, note that the latter argument is premised on the idea that there are such substantial principles. But there is nothing in the diachronic continuity thesis or the Internal Reasons View that says we should believe that. If anything, the Internal Reasons View casts a serious doubt on the existence of such principles, because we may wonder how 'route-like' they really are. In any case, the Internal Reasons View is perfectly consistent with the idea that all non-instrumental principles of reason must be just as formal as the instrumental principle of means-end rationality, and that anything else would not be required by reason alone. This, we have just seen, is compatible with the synchronic Distinctness Principle in its current form, provided that there is a solution to the Disconfirmation Problem.

It follows, therefore, not only that Williams's relationalist version of the Internal Reasons View is not a revision of metanormative Humeanism, but also that the Motivational Continuity Thesis, which is required in order to defend the Internal Reasons View, is not a revision of motivational Humeanism. The whole project of the Internal Reasons View is compatible with both dimensions of the Humean outlook. It might not demand either, but it will at least tolerate both.

3.4 A Humean Defense of the Motivational Continuity Thesis

Now that we have seen that the Motivational Continuity Thesis does not *rule out* the Distinctness Principle, let us see whether we can go a step further and deploy the Distinctness Principle in order to *defend* the thesis. Recall what the thesis requires: independent support for the idea that rational considerations can only produce new motivational attitudes in ways that continue upon what was 'called for' from the perspective of the pre-existing motivational attitudes. Less continuous developments in the motivational set are possible, but they would not be cases in which the development may be considered as demanded by a rational consideration, meaning that the absence of any such development would not have made the agent practically deficient or irrational.

The Humean defense consists of two steps. The first step is to argue that the Motivational Continuity Thesis follows from the Distinctness Principle. The second step is to argue that the reason for a motivational Humean to believe in the Distinctness Principle does not already presuppose some sort of claim about practical normativity—in particular, that it does not already presuppose anti-dogmatic content skepticism. The second step is needed in order to establish the idea that the Distinctness Principle provides *independent* support for the Motivational Continuity Thesis.

3.4.1 How the Thesis Follows from the Distinctness Principle

So what is the first step? Recall that in order to demonstrate the compatibility between the Motivational Continuity Thesis and the 'unrevised' version of the Distinctness Principle, I started out by noting that the latter is a principle about *synchronic* structure, whereas the former is about diachronic change. But insofar that distinction helps to put distance between the two claims, it would also seem to make it difficult to defend the former on the

basis of the latter. If the Distinctness Principle is purely synchronic, then nothing diachronic could follow from it, we might be inclined to think.

However, as I continued to discuss the matter in the previous section, I noted various interrelations between the synchronic and the diachronic matters, and it is in the light of those relations that our defense may now be formulated. Essentially, what the Distinctness Principle does is to explicate the inertia of belief by stating how belief is synchronically related to intrinsic desire, but the inertia of belief which is thus established carries implications for diachronic considerations as well. Suppose, for example, that an agent is motivated to ϕ and that his belief that ϕ would contribute to P seems part of his reason to do so. One may then adopt the Distinctness Principle and claim that this can only be the case on the grounds of a desire that P which did not depend on the belief just mentioned, and that furthermore, any belief on which the desire that P might in turn depend on would in turn presuppose a desire independent from that belief, such that there must be a Q so that the agent is motivated to ϕ because she desires Q intrinsically. But it would surely be odd if one would then go on to claim that this intrinsic desire to Q were diachronically and rationally produced by a prior belief "that Q should be the case" that did not depend on any desires whatsoever, and that the mere fact that the agent *believed* that he should Q would count as an explanation why, other things being equal, he would acquire the desire that Q. Given the idea that synchronic intelligibility requires beliefs to be inert in the sense postulated by the Distinctness Principle, it is completely unclear why it should be rational for any such belief to produce, without the help of desires, such an intrinsic desire at a later instant.

Given that the Distinctness Principle does not allow us simply to point at belief change in order to explain intrinsic desire change under conditions of sustained self-governance, what the Principle requires is that we give some kind of further explanation of why it would take a breach in self-governance to explain situations in which the belief change would not be accompanied by a desire change in this manner. This is, once again, the Disconfirmation Problem. Thus, the diachronic impact of the synchronic Distinctness Principle is that belief adoption cannot by itself explain intrinsic desire addition (as the anti-Humean might want to say), but that instead it now *requires* explanation why belief adoption should be accompanied by intrinsic desire addition. In other words, where the belief counts as an *explanans* for the motivational anti-Humean, it is part of the

explanandum for the motivational Humean.

Now, we have already seen in the previous section that with respect to certain general non-instrumental principles of reason, we might be able to give a straightforward explanation of their capacity to generate new intrinsic desires upon the basis of old intrinsic desires in non-instrumental ways by pointing to the fact that the satisfaction of such requirements is constitutive of the very idea of self-government. In such cases, the diachronic addition or removal of intrinsic desires comes 'on the cheap.' I then discussed the question of why we shouldn't then also allow such principles in the synchronic Distinctness Principle, but remarked that doing so would amount to re-defining the notion of intrinsic desire in a manner that would therefore not strictly revise the original principle. However, what that does show is that the unrevised version of the Distinctness Principle *permits* us to introduce a meaningful notion of non-instrumental rationality with respect to synchronic structure, precisely because it does not imply a revision of the original principle. Hence, the principle allows us to introduce the idea of synchronic *intelligibility* of behavior as action that satisfies the requirements of non-instrumentally rational self-governing agency.

It is such an attribution of the intelligibility of the synchronic structure of the agent's attitudes that would 'match' the non-instrumental deliberative process when the agent actually is functioning rationally, in the same sense in which the instrumental synchronic intelligibility attribution 'matches' the diachronic instrumental deliberative process when the agent actually is functioning instrumentally rationally. However, we have also seen that what may be synchronically attributed as intelligibility in the light of principles of rational agency is *constrained* by the Distinctness Principle to formal considerations. Even principles that are conditional upon desires, such as the principle that when you desire P, reason requires you to also desire Q, are ruled out by the Distinctness Principle if no formal connection between desiring P and desiring Q can be made. Instead, such a principle would be the object of substantial belief, and for intrinsic desires to be under a requirement of reason to comply with such a belief would violate the Distinctness Principle.

But now we can reason as follows: surely, if the motivational efficacy of such a *conditional* substantial belief is already ruled out by the Distinctness Principle, then the same would hold for *unconditional*, 'non-route-like' substantial beliefs about what sort of intrinsic desires reason would re-

quire us to have—for example, the belief that desiring P is a substantial feature of being reasonable. If we accept the Distinctness Principle, then it becomes impossible to explain why self-governance could be subject to such a belief, in the sense that having the belief and not desiring P would constitute a lacking in self-governance. This means that the idea of such beliefs rationally producing intrinsic desires *diachronically* will not come 'on the cheap' in the sense that we can simply explain such a deliberative process as 'matching' a synchronic intelligibility attribution permitted by the Distinctness Principle.

So given the Distinctness Principle, there can be no substantial principles of reason such that an agent might be considered to be unreasonable if the synchronic structure of his attitudes would violate such a principle. But now we may wonder: how could any substantial principle of reason require the diachronic generation of intrinsic desires, irrespective of pre-existing desires, if the *synchronic outcome* of that process would not be required by reason at all? If the S of the agent at some point in time would not, synchronically, violate any principles of reason, then what sort of requirement could there be upon the agent to *have arrived* at a different S if not on the basis of the state of his S at a prior instant? I don't see how these questions could possibly receive positive answers.

There is a different, but perhaps superficially similar question, which could receive a positive answer. That question would be as follows: how could any diachronic process be called irrational if it's outcome, considered synchronically, would not violate any rational requirements? The answer to such a question is simply that there might be requirements of self-governing rational agency that were due to the fact that self-governance is itself a diachronic concept. So far, I have mainly used the notion of self-governance as part of a synchronic requirement: that an agent who has a self-adopted reason to ϕ at time T, but who is not effectively motivated to ϕ at T, is lacking in self-governance. However, note that this principle is not meant to define or capture the notion of self-governance. On the contrary, it is meant to define the concept of a self-adopted reason by referring to a presumably already meaningful notion of self-governance. The synchronic implications of this notion do not preclude that it has constitutive diachronic features as well. And in fact, it seems very plausible that self-government is a thoroughly diachronic concept.[16]

However, there are two things to note about this insight. The first

[16]For various diachronic aspects due to intending and planning, see Bratman (1987).

is that diachronic features of self-governance are primarily about cross-temporal *coherence*, which puts various constraints on how we make plans and organize our schedules and priorities. Such constraints are usually rather conservative with respect to our pre-existing ends: if anything, they help articulating the costs rather than the benefits of major revisions in our subjective motivational sets. In the light of this, we can think of the synchronic-diachronic distinction in two ways. The first way is by thinking of the synchronic in terms of 'time-slices'—configurations of attitudes that have no duration associated with them. However, talking about an action *at a time* in the 'time slice' sense is at best highly abstract: it may serve a purpose when simplification is a necessity, but we should keep in mind that with respect to the 'time slice' notion of synchronicity, every action is already a diachronic phenomenon in itself. However, there is a second, looser and more pragmatic notion of talking about things being synchronous, and that is simply for two things to exist 'during the same phase,' without either being considered as temporally, and in particular causally, prior to the other. Talking about an agent having a structure of attitudes in this synchronic sense is also an abstraction, but in a more pragmatic sense: it allows us to abstract away from the complex underlying causal mechanisms which sustain, or realize, the intelligibility of the agent as having that structure of attitudes during the period of time under consideration.

In terms of this second notion of synchronicity, the distinction between the diachronic and the synchronic becomes a relative one, and the requirements of cross-temporal coherence that self-government places upon us helps to understand how we can shift between different scales of synchronicity. Thus, from a diachronic perspective, self-government requires us to update our attitudes, and respond to various events, so as to keep some of our 'overarching' attitudes intact across that period of time, allowing us to be intelligible as having a synchronic structure of those attitudes during that period of time.

However, whereas self-governance guides the diachronic constancy of attitudes over time, the idea of substantial principles of reason requiring the addition of intrinsic desires in 'non-route-like' fashion would rather represent a breaching of cross-temporal coherence. The sort of non-instrumental deliberation that would bring about major revisions in the agent's *S* would be diachronic in the sense that it moves from one self-governmentally synchronic 'phase' to the next one.

Now it does seems to me that certain conditions which are constitutive of self-governing agency in the diachronic sense would actually make it rational for us to seek changes that upset cross-temporal coherence to a great extent—I will discuss this idea in much more detail later on in this thesis. But given the way in which self-governance is 'prejudiced' towards coherence and conservatism, so to speak, it seems to me that any reason to 'rock the boat' which would be a reason for the sake of self-governance, would have to be justified in the light of *something* sitting uneasily, something being not entirely coherent or governmentally at rest, so to speak, within the *S* of the agent during the period of time before the decision to make a change. Which means that the relevant concern of self-governance itself would be formal, and the way it would make its impact would be 'route like,' by picking up on some tension within the *S*.

To conclude, diachronic considerations which may be constitutive of agency will not help to answer the question how a substantial principle of reason would make it irrational for an agent not to develop a desire that *P*, if his prior *S* did not in any way demand a desire for *P* to be added, and if the state in which he does not desire *P* does not violate any synchronic principles of reason, as must follow from the Distinctness Principle. The point is simply that there is nothing essentially diachronic about the idea of substantial principles of reason. It is the concept of deliberation that is diachronic, but if there are any 'non-route-like' substantial principles of reason then the only reason that an agent would have to come to adopt them is because he had not adopted them already. Since such principles would not depend on anything called for by the prior *S* of the agent, there are simply no diachronic considerations for such principles to call upon in order to escape from the clutches of the synchronic Distinctness Principle. Thus, it follows, that if the Distinctness Principle is true, there can be no beliefs about what reason requires which would explain how an agent could rationally arrive at some motivation that was not in any way called for by the state of his *S*. Which is what the Motivational Continuity Thesis claims.

3.4.2 Why the Distinctness Principle Provides Independent Support

On to the second step of our argument. Why should we believe that the Distinctness Principle is true? Now the project of this thesis is not to defend motivational Humeanism; it is rather to build a meta-ethical account

on the assumption that the Distinctness Principle is true. But without demonstrating that my reasons for this assumption are sound, I shall try to explain why they do not already presuppose a settled view on the scope of practical reason. I have already made the distinction between the normativity of normative reasons and the normative aspect of motivating reasons. In my view, the normativity of motivating reasons does not derive from that of normative reasons, but rather from that of the normativity of beliefs and desires—the concepts in terms of which, according to the motivational Humean, we must analyze motivating reasons.

Now the normativity of desires, in a sense which does not presuppose normative reasons, is simply that of *teleology*, of goal-directedness. This notion, it seems to me, is not at all restricted to deliberating agents like us, but is instead something that we share, on some level, with 'lesser' agents such as various members of the animal kingdom. I shall make no assumption about where on the evolutionary ladder the lesser agents begin, nor about where they get promoted to a higher status or what sort of intermediary concepts of agency we might discern along the way. What matters is that we need not invoke our concepts of normative reason or practical judgment in order to attribute goal-directedness, and therefore the notion does not seem guilty of presupposing our meta-ethical analysandum.

Regardless of how this notion of teleology is understood precisely, the intuition behind Humeanism is that *getting something right* has nothing to do with it. Belief may derive its meaning from the context of agency, as interpretationists about the mental like Dennett and Davidson have argued, but within that context, the role of belief is precisely that of representing the world in a manner which is not geared towards a goal. Of course, it may be in the interest of an agent, given his desires, to hold beliefs about the things which interest him, and in that sense his beliefs are geared towards his goals, but they are still disinterested in terms of their content.

And the reason for that, I submit, is that motivational Humeanism ultimately captures an intuition about *truth*: that truth itself is disinterested. By that I don't mean that "reality," or "the world" is disinterested (whatever that, or its opposite, would be supposed to mean), but rather that the *concept of truth* is disinterested: what it *means* to say that a proposition is true would not be different if one were motivated by different interests. It might have different meaning in the sense of what "it meant to you" but that is already a much richer notion of meaning, incorporating the interaction of the belief in the truth with one's desires. But in the pure

assertive sense of meaning, which the concept of truth is about, what it means to say that a proposition is true is completely independent from what one's motivations might be, as long as the proposition stays the same.

This also means that any results from cognitive science which point to the fact that our behavior is best explained in terms of attitudes or cognitive structures that do not represent, separately, our understanding of the world and our ideas on how to interact with it, are besides the point. From the interpretationist perspective on the mental, insofar as any such cognitive structure implements an understanding of the world, we may attribute to the agent the attitude of holding certain things to be true, and what that means, by itself, carries no interests one way or another. Now it may be remarked that truth, itself, is also a normative notion. One might even say that it is *the* normative notion. But any Humean who would agree with this could hardly be accused of building ideas about *practical* normativity into his understanding of truth. It should be obvious that the sense in which our concept of truth is normative is a much more theoretical one: it is the normativity that concerns all areas of knowledge, as studied by epistemology in the general sense.

If the reasons that motivational Humeans have for believing in the motivational inertia of belief are grounded in an understanding of epistemic normativity, then I think we have established that the Distinctness Principle need not presuppose anything practically normative. Therefore, anti-dogmatic content skepticism is not dialectically prior to the Distinctness Principle. Which means that the Distinctness Principle provides independent support for the Motivational Continuity Thesis. And that means, as I have argued in section 2.5.1 from the previous chapter, that with the Distinctness Principle on our side, we can refute the 'non-route-like' deliberation objection against Williams's defense of the Internal Reasons View.

Assembling the different arguments that we have seen in this chapter and chapter 2, we now have a defense of CIRV premised on various elements from chapter 1. On the assumption that the Disconfirmation Problem can be solved, we have seen that the Distinctness Principle implies the Motivational Continuity Thesis, which together with the proceduralist version of the Disconfirmation Principle, and the premise that captured our intuition about self-government, yields the Internal Reasons View. Coupled with the Facts and Authority Principles, this leads to the Comprehensive Internal Reasons View, or CIRV. Finally, we have seen two strategies

for combining CIRV with nonrelationalism. The former, Thomas's anti-Humean strategy, would be ruled out if the Distinctness Principle is true. The latter strategy, which consisted in a combination of formal considerations about convergence and what is constitutive of agency, is ruled out by Williams's anti-formal content skepticism. Note that this may still be an independent premise, of which I have not claimed that it could be derived from the Distinctness Principle. In any case, if we add this premise then relationalist CIRV follows, and we get the motivational Humean 'version' of Williams's view, so to speak. This is the view that I will take as the inspiration for the account that I am going to develop in this thesis. Note that it leaves a lot to be answered: how to solve the Disconfirmation Problem, a fuller understanding of the notion of self-governance that I have relied upon so often, as well as a more detailed account of what non-instrumental deliberation may consist in within the 'route-like' boundaries.

As noted before, however, I mean to stay neutral about the exegetical question of whether Williams actually had a motivationally Humean view in mind. In any case, the Humean defense that I have just offered in support of the Internal Reasons View is certainly my own. For the sake of completeness, I shall now turn briefly to the prospects for a defense of CIRV on the basis of an anti-Humean theory of motivation.

3.5 The Anti-Humean Defense of the Motivational Continuity Thesis

We have seen that the Motivational Continuity Thesis is compatible with the Distinctness Principle, and furthermore, that it actually follows from that principle. But does a defense of the thesis also *require* the Principle? A crucial step in the above line of argument was to show that, in order to accommodate the idea that substantial principles of reason might produce new intrinsic desires, we would have to reject the Distinctness Principle and adopt an anti-Humean theory of motivation. This may seem to suggest that, on the basis of motivational anti-Humeanism, we could reject the Motivational Continuity Thesis, and adopt the External Reasons View.

I think this is correct, in the sense that without further qualification, being an anti-Humean about motivation is compatible with being an external reasons theorist. However, from that it does not follow that all anti-Humean theories of motivation are compatible with the External Reasons View—only that some might be. It depends on the specifics of

the theory. Now if we ask ourselves what sort of anti-Humean theory of motivation Williams might have been most sympathetic to, then the most obvious candidate, it seems to me, would be the theory that we must be anti-Humeans in order to make sense of Williams's notion of *thick concepts*.

3.5.1 The Argument from Thick Concepts

Williams famously distinguished between "thin" and "thick" ethical concepts. Examples of thin concepts are *good* and *ought*, which can be used meaningfully to give a positive or negative evaluation of basically any action, state of affairs, or matter of fact. The concept of good only expresses *that* something is evaluated positively, not what sort of thing or why. By contrast, examples of thick concepts are *cowardice* and *bravery*, which not only express the positive or negative evaluation, but also *describe* the evaluated action as being, empirically, of a certain kind, such that it is in virtue of it being of that kind that it is to be evaluated in that manner.

The question with respect to thick concepts is whether we can give a "two-factor analysis" of them, such that the attribution of any thick concept to an action ϕ may be reduced to the combination of a purely descriptive characterization D of ϕ and a pure evaluation, using a thin concept, of ϕ's being D. Harcourt & Thomas (forthcoming) argue that the nature of thick concepts is such that they resist this kind of analysis. A discussion of their detailed arguments against specific two-factor proposals is beyond the scope of this thesis, but their general explanation for the phenomenon is that grasping the extension of the concept involves sharing a social perspective with its users that already constitutes the evaluative interest that gives ethical concepts their directive potential. There are two implications which this argument may be thought to have, both of which are endorsed by Thomas. The first is that this feature of thick concepts makes ethical knowledge possible—the knowledge contained in true and justified thick concept attributions. The second is that this feature of thick concepts disproves motivational Humeanism.

With respect to the first implication, as we have already seen, Williams was skeptical about the idea of *nonrelationalist* ethical knowledge. Nevertheless, Williams may have been more sympathetic to the second implication. The reasoning behind it should be more or less as follows. Under conditions of self-governance, our practical judgments have motivational implications. Therefore, if our practical judgments involve thick concept

attributions, then the evaluative aspect of thick concepts must be motivationally efficacious in some sense (i.e., a 'defeasible' sense that can only be thwarted by factors that explain the impairment of self-governance). However, if a two-factor analysis of such concepts is impossible, then this motivationally efficacious part cannot be separated from the belief-like attitude that the descriptive part of the thick concept attribution involves. Hence, the motivation provided by thick concept use cannot be analyzed by separating the non-motivational role of belief from the motivational impetus of intrinsic desires in the manner that the Distinctness Principle requires.

Of course, this way of presenting the argument is very sketchy, and in order to demonstrate its validity a number of premises would have to be explicated about what the two-factor analysis is supposed, exactly, to accomplish and whether that corresponds, exactly, to what the Distinctness Principle would require. I do not think the argument is sound, so I would probably reject one of those premises, but to figure this out is beyond the scope of this thesis.

In any case, what matters for my present purpose is that some versions of motivational anti-Humeanism may be based on this argument, and if they are, they might furthermore accept the restriction that it is *only* through this role of thick concepts that beliefs may have non-instrumental motivational efficacy. In other words, according to this version of the anti-Humean theory, a belief may motivate non-instrumentally if and only if the belief involves the attribution of a concept that resists two-factor analysis. Suppose furthermore, that the only reason why concepts may resist such analysis is Thomas's claim, that one cannot determine the extension of the concept unless one already shares the interest implicated in that concept.

Now the point is not that one would just lack all sort of accidental knowledge required to determine the extension of the concept. The point seems rather that one would simply fail to grasp the *intension* of the concept if one does not share the interest implicit in it—one can only grasp the concept from the *perspective* of the interest. But if that is true, then it would seem that for an agent who does not share that interest, there is no way of *acquiring* the concept in the light of its descriptive merits alone, such that the interest implicated in it would then come along and get added to one's *S*. Instead, it would seem that in order to develop the grasp of the concept, one would have to develop the interest needed to grasp that concept from the inside, as it were, on the basis of one's pre-existing interests. Which is

exactly what the Motivational Continuity Thesis requires.

3.5.2 *Does the Argument Provide Independent Support for the Thesis?*

There are two possible interpretations of the argument I have just sketched. According to one interpretation, which Thomas seems to have in mind, it does not provide *independent* support for the Motivational Continuity Thesis, because the notion of a thick ethical concept on which it relies already involves the idea of practical normativity. Therefore, the motivational skepticism about learning thick concepts on the basis of rational insight and acquiring new motivations as a result from it, would not be 'dialectically prior' to the content skepticism implicit in such an understanding of the nature of thick concepts.

According to a second version of the argument, even though thick ethical concepts are in a sense 'culturally holistic,' presupposing a historical background that is already rich with our idea of practical normativity, the reason why culture does function in this sense can be explained in terms of the psychological mechanisms that enable us to be cultural beings, and it is because of psychological reasons that we cannot distinguish between the ability to grasp the concept and the motivational interest that is implicated by the possession of the concept. On this version of the argument, the motivational skepticism *would* be prior to the normative skepticism.

I have no idea which version, if Williams was indeed a motivational anti-Humean, would be closest to how he meant to defend the Internal Reasons View. However, as I have argued in section 2.5.2, it does seem to me that insofar as Williams was an anti-*formal* content skeptic for 'dialectically prior' reasons, this skepticism was directed against the Kantian version of the Internal Reasons View, rather than against the External Reasons View. But anti-*dogmatic* content skepticism seemed ill-suited to stand on its own as a nonreducible meta-ethical claim about the nature of practical normativity, because it would continue to beg the question against the idea that reason might be intrinsically non-formal.

Thus, it seems that against the first version of the argument, an external reasons theorist might simply reply that thick ethical concepts are not the only way in which beliefs can non-instrumentally motivate, and that having beliefs about substantial principles of reason is a further way in which beliefs can do this. The proponent of the first version would then be back at square one in the question-beg stalemate over the scope of

practical reason. Instead, the proponent of the second line of argument might have some story about why our psychology does support the idea of motivational thick concepts, but not of being motivated by substantial principles of reason. If motivational anti-Humeanism were true for such psychological reasons, then it seems to me that the Internal Reasons View must be true as well.

I conclude, then, that the Internal Reasons View may be defended upon both motivationally Humean and anti-Humean grounds. Despite its possible exegetic inaccuracy, we shall see in the next chapter that the Humean defense of the Internal Reasons View provides us with a sketch of a solution to the Facts Problem. Interestingly, we have also seen that the Humean defense is premised on the assumption that a solution to the Disconfirmation Problem can be given. But as we shall see in the chapters to come, solving the Facts Problem along the lines of the Internal Reasons View does not make the Disconfirmation Problem go away. This means that an independent account of disconfirmation is required, which I am going to build in chapter 7. On the one hand, this account will stay within the boundaries of the Internal Reasons View, as there will be no room for external reasons. But the picture of the deliberative route will change radically: instead of Williams's focus on the scope of practical reason and the room for constitutive imagination, my own account will give a central role to the idea of empirical self-knowledge.

II Facts about Reasons : The Status Quo

4 *Outline of a Relationalist Solution*

In chapter 1 I have formulated five principles in order to articulate certain intuitions about practical judgment. I started with the Facts Principle and the Disconfirmation Principle, which led us to the Facts Question and the Disconfirmation Question. We have seen that in conjunction with the Authority and Distinctness Principles, the Facts Principle leads to a paradox. Thus we arrived at the Facts Problem: the problem of answering the Facts Question in a way that would resolve this paradox. In a similar fashion, we arrived at the Disconfirmation Problem: the problem of answering the Disconfirmation Question in a manner that would resolve a paradox resulting from the conjunction of the Disconfirmation, Authority and Distinctness Principles.

Because the Facts and Disconfirmation Principles are, in some sense, two sides of the same coin, the Facts and Disconfirmation Problems are not really separate problems. They are rather two different ways of approaching the same problem. Nevertheless, there is a certain asymmetry between these two perspectives. From the perspective of the Facts Question, the problem takes a metaphysical form: the Facts Problem requires us to explain the nature of facts about normative reasons, which seems like a very fundamental, but also highly abstract and theoretical undertaking. In contrast, from the perspective of the Disconfirmation Question, the problem takes an epistemological form: the Disconfirmation Problem requires us to explain what sort of things could rightfully make us change our minds and our ways of behaving, which seems like a much more concrete and practical issue.

Of these two perspectives, the perspective of the Facts Question may seem to be the most fundamental one: it is about what normativity ultimately *is*, whereas the Disconfirmation Question is merely about how we can correct our mistakes about it. Making mistakes about practical normativity already presupposes the existence of practical normativity, which may suggest that it is the Facts Question that really gets to the bottom of

things. This view seems to be implicit in the literature, where questions of disconfirmation are usually treated as derived from questions about moral truth or moral facts. The best known meta-ethical "isms" that moral philosophers defend, such as "cognitivism," "realism," and "relativism," are first and foremost about truth and truth conditions, not about the circumstances under which we should revise our practical judgments. I will briefly go along with this view in the current chapter and in chapters 5 and 6, discussing different solutions that may be extracted from the literature in response to the Facts Problem.

However, in the last two chapters we've already seen that in order to defend the Internal Reasons View, one of the most prominent views in the literature on these matters, we had to make assumptions about disconfirmation at crucial points in the argument. First of all, it turned out in section 2.4 that Williams's own argument for the view is premised on an implicit *proceduralism* about disconfirmation. Secondly, I have shown in section 3.3.2 how a defense of the view within a motivationally Humean framework must rely on the assumption that a solution to the Disconfirmation Problem can be formulated.

In similar fashion, we shall see in this second part of the thesis that the different possible solutions to the Facts Problem are often driven by assumptions or ideas that concern disconfirmation. In my own view, this means it will be much more fruitful to think of the Disconfirmation Question as the primary question, and to approach the Facts Question as derivative. Therefore, in part III of the thesis, and especially in chapter 7 the Disconfirmation Question will become our starting point, and solving the Disconfirmation Problem will be our primary objective. In the end, finding a satisfactory answer to the Facts Question might still be seen as the most fundamental goal, but it is through a theory of practical disconfirmation that we must get there. Or so I will argue.

Nevertheless, in order to get a grip on the *status quo*, let us first take a closer look at the Facts Problem. Because the problem involves a paradox, there are essentially two types of solutions: we must either reconcile the principles that gave rise to the paradox by explaining away their apparent inconsistency, or we must reject one of the principles and explain away its apparent plausibility. Let us call solutions of the former type "reconciliatory" and those of the latter type "dismissive." Among the reconciliatory solutions, we can make further distinctions depending on whether they are relationalist or nonrelationalist about the Intersubjectivity

Principle, and on whether they are proceduralist or nonproceduralist about the Disconfirmation Principle. Among the dismissive solutions, we must of course distinguish different views depending on which of the Principles they are dismissing. Thus, the Distinctness Principle is rejected by motivationally anti-Humean views, the Authority Principle by the 'no reasons' forms of externalism, and the Facts Principle itself is rejected by noncognitivism, quietism, and what I will call "radical error theory."

The solution that I favor, and which I shall defend throughout this thesis, is reconciliatory, relationalist, and proceduralist. If we compare this type of view to the other options that I have distinguished, then the fact that it is a *relationalist* type of view stands out as its most distinctive feature. As we shall see in the next chapter, there seems to be little reason for proponents of a nonproceduralist reconciliatory solution to subscribe to relationalism, and the same applies to those who are dismissive of the Facts and/or Authority Principles. It might be possible to defend relationalism from within a motivationally anti-Humean framework—Bernard Williams, as we have seen in the previous chapter, was a relationalist without being unambiguously committed to a Humean theory of motivation. Nevertheless, most philosophers who are more explicitly anti-Humean about motivation—including those anti-Humeans who do subscribe to the Internal Reasons View, such as Alan Thomas—are nonrelationalists about intersubjectivity.

In this chapter, I will present an outline of the relationalist view that I favor, and explain how that view solves the Facts Problem. The solution is based on a "dispositional" approach to practical normativity, an approach that has become very popular in the literature, in widely different varieties. The dispositional solution that I propose is based, more or less, on the relationalist, motivationally Humean version of CIRV that I defended in section 3.4 of the previous chapter, but in the present chapter I shall define the solution in my own terms, so that we can leave all the disambiguations and exegetical disclaimers with respect to Williams's work behind us. Nevertheless, the reasons why we are being pushed towards relationalism, at this stage, are basically the ones that I articulated in section 3.2, including Williams's anti-formal content skepticism, a further discussion of which must await chapter 5.[1] In the present chapter, I will simply stick with a relationalist understanding of the solution, and take stock of the challenges that the relationalist must be able to meet in order to make his solution

[1] See section 5.4.

work. One of the goals of this thesis will be to address those challenges. We return to nonrelationalist reconciliatory solutions in the next two chapters.

4.1 The Dispositional Approach

In order to reconcile the intuitions behind the Facts, Authority and Distinctness Principles, various authors have tried to defend a "dispositional" or "response-dependence" theory of practical normativity. According to such a theory, the facts about our normative reasons for action are facts about what our motivations and affective responses would be under certain ideal conditions of agency (Firth, 1952; Brandt, 1979; Williams, 1980/1981a; Smith, 1989, 1994, 2002/2004c; Lewis, 1989; Johnston, 1989; Jackson & Pettit, 1995). These conditions typically include self-government, correct reasoning, and access to the relevant information. The general idea is simple: we remove the mystery about why we would be motivated, under ideal conditions, in accordance with our normative reasons, by *analyzing* normative reasons in terms of the motivations that we would have under those conditions.

Note that the general formulation of dispositionalism that I just gave presupposes the Facts and Authority Principles (because it identifies facts of practical normativity as facts about normative reasons), but not necessarily the Distinctness Principle: it only refers to our motivations under ideal conditions, without saying anything substantial about how those motivations would have to be constituted or generated. This is reflected in the literature: some dispositionalists and response-dependence theorists are self-proclaimed Humeans about motivation (Lewis, 1988, 1989; Smith, 1987, 1989), some are anti-Humean (McDowell, 2001; Thomas, 2006), and some are difficult to classify (Bernard Williams is a notable example, as we have seen in the previous two chapters). Furthermore, we have already seen that CIRV, which we may now understand as a variety or subclass of dispositionalism, can be formulated and defended independently of the dispute between Humeans and anti-Humeans concerning motivation.

Nevertheless, the dispositional approach is a very attractive one for Humeans about motivation who wish to reconcile our Principles, because it provides a formula to synthesize facts about normative reasons with the analysis of motivation in terms of instrumental beliefs and intrinsic desires. The motivationally Humean dispositionalist idea is, again, simple: we remove the mystery about why we would be motivated, under ideal

conditions, in accordance with our normative reasons, by analyzing normative reasons in terms of the intrinsic desires, the instrumental beliefs, and the derived desires that we would have under those conditions. From now on, when I speak about dispositionalism, dispositional theories, or the dispositional solution, I will have this more specific, motivationally Humean version of the dispositional idea in mind.

If the Humean is right about motivation, then why, we may now ask, should the content of our intrinsic desires be any different under ideal conditions from what it is under the actual conditions? According to the Distinctness Principle, no intrinsic desire can be refuted by belief, no matter how true and justified. So what could there be 'less than ideal' about our actual intrinsic desires? The answer is that even though no intrinsic desire is subject to criticism when considered in isolation, there are certain *combinations* of different intrinsic desires that an 'ideal' agent would never have. Thus, it seems hardly ideal for an agent to have very strong intrinsic desires that are simply inconsistent with each other. This insight does not really go against the Distinctness Principle: the reason why it is undesirable to have such inconsistent desires does not depend on any beliefs about the state of the world.

By itself, this insight does not yet provide us with an answer to the Facts Question. From the mere fact that my actual set of intrinsic desires is less-than-ideal, it does not follow that there are positive facts about what my ideal set of desires would be. Instead, this might simply remain undetermined. After all, suppose that I desire both that P and that not P, that these desires are equally strong, and that the rest of my desire set is completely neutral about whether to prefer P or not P. In that case, it seems that we could think of two ideal versions of myself, of which one would still desire that P but not that not P, while the other would desire that not P and not that P. It seems completely undetermined which of these I would be under the ideal conditions. Hence, these considerations do not provide me with a fact about whether I should approve or disapprove of P.

However, in practice things are usually not so symmetric. Not just because one desire might be stronger than a conflicting desire, but also because one of the conflicting desires might be more coherent with further desires that we have. Thus, we might be able to understand facts about normative reasons as facts about certain *asymmetries* within the sets of our intrinsic desires. On the basis of such asymmetries, some idealized versions of ourselves might be considered 'nearer' to our actual self, and

then our normative reasons could be analyzed in terms of the desires of our 'nearest ideal self.'[2]

This line of thought leads us back to the dispute between relationalist cognitivism and nonrelationalist realism. If it depends on my *actual* set of desires which ideal self is nearest to the actual me, then it would seem that my normative reasons might have been different if my actual set of desires were different. Consequently, it seems no longer guaranteed that different less-then-ideal agents have normative reasons to approve of the same states of affairs. After all, since their actual desire sets differ, their nearest ideal selves might have different desires as well. Their conflicts in the actual world might persist, so to speak, among their nearest ideal selves. Which would commit us to relationalist cognitivism. In contrast, if the nonrelationalist realist wants to develop a dispositional solution to the Facts Problem, then he would have to show that the ideal selves of all agents must always desire the same states of affairs. We have touched on this matter in section 3.2, and I will return to the possibility for a nonrelationalist dispositional solution in chapters 5 and 6, while pursuing the relationalist solution in the current chapter.

Aside from the dilemma between relationalism and nonrelationalism, there may also be certain problems that affect dispositional solutions in general. Some philosophers may find the notion of ideal conditions troubling, or have skeptical worries about the very idea of dispositional facts and truth conditions. If the dispositional solution would fail for such reasons, then perhaps another type of solution might be devised—it is not obvious that the dispositional approach is our only option for reconciling our Principles. Furthermore, there are meta-ethical views in the literature that do not explicitly reject a dispositional solution, but which also do not explicitly make use of it, even though they do seem to combine all the Principles, which puts them in the 'reconciliatory' camp. Harry Frankfurt's view, which I will discuss briefly below, and extensively in chapter 8, might be an example of such a position. However, as I will argue in this chapter, one of the things that makes Frankfurt's view attractive is the contribution that it may have to offer to the dispositional account. And as we shall see

[2]That does not mean we should deny the possibility of 'symmetric' cases. Instead, we may simply conclude that the symmetric cases constitute genuine dilemmas, which cannot be resolved on the basis of sound deliberation. In order to uphold the Facts Principle, we need only show that there are facts about normative reasons that resolve some cases, not that they resolve all cases. I will return to this matter in section 9.2.2.

later on,[3] Frankfurt's view has problems of its own that are perhaps most easily solved along dispositional lines. With these considerations in mind, I shall proceed by treating the dispositional approach as our principal strategy for solving the Facts Problem in a reconciliatory manner.

4.2 Type-1 Dispositionalism

What are the ideal conditions of agency? In the light of CIRV, we may think of these ideal conditions as constituting the endpoint, or limit, so to speak, of the deliberative route that an agent might take: the point at which no further deliberative improvements could be made. In section 2.2 I have discussed Williams's take on this, which involved both instrumental reasoning and deliberation about ends. With respect to the former, I have argued against Williams's proximity condition on instrumentally relevant facts. In line with my argument there, I will include in my proposal that the agent holds all the relevant true beliefs, and no false beliefs, about the means to his ends, under the ideal conditions of agency.

With respect to non-instrumental deliberation, CIRV claims that it must be 'route like' and start from actual elements of the motivational set rather than from any substantial principles of practical reason. I have argued that this criterion can be defended on the basis of the Motivational Continuity Thesis, and that the latter thesis could in turn be supported independently on the basis of the Distinctness Principle. Furthermore, we have seen that the combination of 'route like' deliberation and proceduralism leads to relationalism unless we can make some kind of convergence strategy work, but the informal convergence strategy is incompatible with the Distinctness Principle, while the formal strategy would be ruled out by anti-formal content skepticism. Thus, even though Williams may not have been a motivational Humean himself, and even though his own reasons for defending relationalism may have been different, we can now see how the criterion of 'route like' deliberation plays an intermediate role in a chain of arguments from the premises of proceduralism, the Distinctness Principle, and anti-formal content skepticism, amongst others, to the conclusion of relationalism.

If we now omit the intermediate steps in this argument chain (the Motivational Continuity Thesis and the Internal Reasons View), and focus instead on the relationalist conclusion and the premises that gave rise to

[3] See sections 8.3 and 8.4.

it, then we get an answer to the Facts Question along the following lines. The facts which make the true practical beliefs of an agent true are facts about the *actual* intrinsic desires of the agent, such that those beliefs might have been false had those desires been different. More specifically, the fact that makes the practical belief in approval of P true is the fact that the actual intrinsic desires of the agent would give rise to the resultant desire that P if the agent would deliberate correctly upon those desires, rejecting all instrumentally relevant false beliefs and acquiring all instrumentally relevant true beliefs with respect to P in the process, and applying any valid methods of non-instrumental deliberation to his self-adopted ends in such a way that, had his intrinsic desires been different, such deliberations might not have given rise to the resultant desire that P.

This answer is sketchy at best, and further clarifications will be needed. However, I first want to propose a terminological modification. I have already mentioned self-government as one of the necessary conditions of ideal agency, and various considerations pertaining to self-governance have played an important role in the previous two chapters. An agent whose practical beliefs are true but who fails to act accordingly is not very ideal, and neither is the agent who is motivated in accordance with his normative reasons even if his beliefs are at odds with them. Now Williams, but also Michael Smith, for example, have used notions like "deliberation," "deliberative route," and "correct deliberation" in a very broad sense that already presupposes a match between motivation and practical belief at the end of the route, so to speak, and the formulations in the paragraph above are consistent with that. But a more common usage of the term "deliberation" does not seem to presuppose self-government. Consider Frankfurt's example of the "unwilling addict" (1971/1988b, p. 17). The addict is enslaved by a desire that he fails to control, yet his practical judgment opposes this desire. There seems to be no reason why the unwilling addict might not get his judgment right, and even have the justification for it.[4] Under those conditions, we might say that the unwilling addict knows he has a normative reason not to take the drugs, even though he fails to act upon it. His deliberations are correct, yet he doesn't have the desires to act accordingly. Therefore, it seems we should distinguish the requirement of self-government from that of deliberative correctness in

[4]This point is closely related to my argument, in section 1.4.1, for preferring the reference to "self-governance" in my Authority Principle to Smith's reference to "practical rationality" in his practicality principle.

order to specify the conditions under which the desires of the agent match his normative reasons for action. Thus, our modified answer to the Facts Question becomes as follows:

> Type-I dispositionalism. The fact that makes the practical belief in approval of *P* true is the fact that the actual intrinsic desires of the agent would give rise to the resultant desire that *P* if (1) the agent would deliberate correctly upon those desires, applying any valid methods of instrumental and non-instrumental deliberation to them, and rejecting all instrumentally relevant false beliefs and acquiring all instrumentally relevant true beliefs in the process, and if (2) his desires themselves were modified according to those deliberations, so as to maintain or establish the state of self-government, while (3) his actual intrinsic desires could have been different in such a way that (1) and (2) might not have led to the resultant desire that *P*.

I am calling this type of answer "type-I dispositionalism" because it is the first of three types of dispositionalism that will play a major role in the rest of this thesis. I will introduce "type-II" and "type-III" dispositionalism in the next chapter.

Now let us see whether type-I dispositionalism allows us to solve the Facts Problem: what sort of facts could make beliefs true in such a way that (a) self-governing agents who had those beliefs would have to be motivated in certain ways, if it is also true that (b) their motivation depends on intrinsic desires that could have been different regardless of their beliefs? Note first that the problem refers to an agent who is (i) self-governing and (ii) knows the facts about his normative reasons for action. Thus, the problem is about an agent under ideal conditions. But according to type-I dispositionalism, the facts about his normative reasons are the facts about his desires under those conditions. Thus, if the intrinsic desires mentioned under (b) had been different, then the facts about his normative reasons for action would have been different as well. Which means that the truth conditions for his practical beliefs would have been different too. But that means that in order for the agent to still be an agent who knew his normative reasons, his practical beliefs would have to be different too. Instead, if the agent had different intrinsic desires but the

same practical beliefs, then some of his practical beliefs would have been false.

This means that there is an ambiguity in the formulation of the Facts Problem. Yes, it seems contradictory to say that an agent who has to be motivated in a certain way, could have been motivated differently. But rather than one paradoxical claim, there are actually two perfectly compatible claims that now follow from our Principles. The one claim is that a self-governing agent who knows his normative reasons for action could have had different intrinsic desires and still be a self-governing agent who knows his normative reasons—but with different practical beliefs. The other claim is that a self-governing agent who knows his normative reasons for action could have had different intrinsic desires and still have the same practical beliefs—but then he would no longer be a self-governing agent who knows his normative reasons for action. On the basis of type-1 dispositionalism, these *two* claims follow from the conjunction of the Facts, Authority and Distinctness Principles, and there is no contradiction between them at all.

4.3 Semantic Pluralism

Note that from the claim that all reasons are internal reasons, it does not follow that the external interpretation never portrays what people actually *mean* when they utter statements about reasons. On the contrary, Williams's example of Owen's family is meant to illustrate that people really do make judgments about external reasons. Could this be a problem for the relationalist? If an agent A has a normative reason in the internal sense to ϕ, then if the relationalist line of argument is sound, this can only make it true that A should$_A$ ϕ in the relationalist sense, not that it should be the case in the nonrelationalist sense that A does ϕ. But the same does not apply to the statement that A has a normative reason in the external sense to ϕ. When Sir Philip Wingrave tells his son that there is a reason for Owen to join the army no matter what he desires, he seems to be implying that any man in Owen's position would have had the same reason. Thus, Sir Philip is not saying that Owen's belief that he should not join the army is false in a relationalist sense. Furthermore, it would also be strange to say that Sir Philip is expressing a relationalist practical belief of his own. We may have good reason to think that Sir Philip is trying to push his own preferences onto his son, but we should acknowledge that that is not

what Philip Wingrave *means*. At least on Williams's interpretation of the play, it would be very odd to suppose that even though Sir Philip claims that there is a reason for Owen to join the military, by that he really just meant that Owen should$_{\text{Philip}}$ join the military, and that, furthermore, he might have agreed that it was false that Owen should$_{\text{Owen}}$ join the army. It seems much more plausible to claim that Sir Philip simply believes that it should be the case that Owen joins the military, in the nonrelationalist sense. Hence, it would seem to follow from Williams's analysis that at least some practical beliefs are to be interpreted in the nonrelationalist sense, and that relationalism is therefore false.

Rather than as a relationalist about all practical judgments, we might think of Williams as a kind of 'partial error theorist,' because his view was that *insofar* an agent is saying that he has a reason in the external sense, his statement must be "false, or incoherent, or really something else misleadingly expressed" (1980/1981a, p. 111). However, once we have removed this error from our thinking, we may understand our practical judgments in the internal sense exclusively, and then, according to the Internal Reasons View, we might very well get them right, and adopt true beliefs about our normative reasons.

In fact, this may not be so far apart from John Mackie's view, even though his error theory is often presented as the thesis that all practical beliefs must be false, period. For in his own writing, Mackie merely stated that *most* people make moral claims in a manner that renders them false because of the error theory:

> The claim to objectivity, however ingrained in our language and thought, is not self-validating. It can and should be questioned. But the denial of objective values will have to be put forward not as the result of an analytic approach, but as an 'error theory', a theory that although *most* people in making moral judgements implicitly claim, among other things, to be pointing to something objectively prescriptive, *these* claims are all false. (Mackie, 1977, p. 35, my emphasis)

In other words, Mackie allowed that *some* people might make their moral judgments without being implicitly committed to this false claim. And indeed, he allowed for this possibility explicitly in order to demonstrate the logical independence of first-order judgments from meta-ethical judgments:

> A man could hold strong moral views, and indeed ones whose content was thoroughly conventional, while believing that they were simply attitudes and policies with regard to conduct that he and other people held. (p. 16)

We may wonder whether the "attitudes and policies" in this passage, which do not presuppose the error of objective prescriptivity, could nevertheless still be understood, on Mackie's own account, in more or less cognitivist terms. At first we may be inclined to think that they cannot, but note that later on in the book, Mackie did also write that the "attempt systematically to describe our own moral consciousness" is a "legitimate kind of inquiry" that "must not be confused with the superficially similar but in purpose fundamentally different attempt [...] to advance [...] to an objective moral truth" (p. 105). Surely, "describing" our moral consciousness sounds like a cognitivist enterprise, to which we might apply our Facts and Disconfirmation principles, and construing it as an "attempt" suggests that this enterprise is not trivial, but that there is really something to be gotten right or wrong about the moral consciousness of ours that Mackie is referring to. Let me stress that I am not saying that Williams was an error theorist in Mackie's sense or that Mackie was an internal reasons theorist in Williams's sense. But it follows from the cited passages that both authors recognized the need to leave some room for what may be understood as the central idea behind the other's view.

Note that all this does not really go against the spirit of the relationalist account, though. Sure, some people mean their judgments in a nonrelationalist sense on certain occasions, but if we can say that those judgments will always be false, then the tension between Williams's view and relationalism as I have defined it may be merely terminological. The point of the relationalist account is that *those practical judgments that satisfy* the Facts Principle must be understood in a relationalist sense, because the only facts that can make practical beliefs true are facts about relations between the judging agents and the objects of their judgments. In order to disambiguate our terminology, let us distinguish between the "R-practical" and the "NR-practical": R-practical judgments are judgments that have relationalist content, whereas NR-practical judgments have nonrelationalist content. We can then reformulate type-I dispositionalism as a view that provides (a) an answer to the Facts Question for R-practical beliefs, and (b) an error-theory for NR-practical beliefs. The real issue, of course, is that this involves what we might call a "semantic pluralism" about practical

judgment: the idea that different people follow different meta-ethical rules. I will discuss this idea in greater detail later in this thesis.

4.4 Problems for Type-1 Dispositionalism

We have now seen an outline of a relationalist solution to the Facts Problem. If type-1 dispositionalism is correct, then the paradox generated by the Facts, Authority and Distinctness Principles dissolves, which means that we can hold on to all three principles and thereby honour the intuitions that motivated these principles. However, the type-1 dispositionalist has his own problems to solve, in order to really make his account work.

The first problem is that in its current form, the account largely depends on a *promise*. The promise is that the desires of a person—the elements of his "subjective motivational set," understood in the context of the Distinctness Principle—are such, that valid deliberative strategies, when applied to these desires, will be able to resolve the conflicts among them in a rationally compelling manner. Not all of those conflicts, perhaps, but enough of them in order not to become skeptical about our deliberative efforts. However, it is not immediately obvious why this should be the case. Given the non-rational nature of intrinsic desires, it is a philosophical challenge to explain why deliberation could give us a rationally compelling reason to prefer one desire over another, when all that there is for deliberation to be based upon are simply those desires themselves. In section 4.1 above I have speculated about 'asymmetries' in the ways our desires are interrelated, which might allow us to construct a function of proximity between the actual, incoherent set of desires of some agent, and the different coherent desire sets that preserve elements of the original incoherent set. But this idea would have to be worked out in further detail, and we should be able to show that such an account makes sense of what we actually do, or at least try to do, when we deliberate in practice.

4.4.1 Four Concerns with respect to Unclarity

Williams does not offer such an account in the "Internal and External Reasons" article. On the contrary, Williams is purposefully vague both in his characterization of what goes into the subjective motivational set and in his characterization of the deliberative route from the actual set to the idealized set:

> But here it may be objected that the account of deliberation is very vague, and has for instance allowed the use of the imagination to extend or restrict the contents of the agent's *S*. But if that is so, then it is unclear what the limits are to what an agent might arrive at by rational deliberation from his existing *S*.
>
> It *is* unclear, and I regard it as a basically desirable feature of a theory of practical reasoning that it should preserve and account for that unclarity. There is an essential indeterminacy in what can be counted as a rational deliberative process. (p. 110)

However, there are a number of concerns involved here that need to be disentangled. The first concern is to be able to account for certain indeterminacies in the reality of our normative reasons themselves. As I have anticipated before, sometimes there may not be a fact that renders a single practical judgment true and all the alternatives false. Later on I will distinguish different sorts of scenarios that feature this type of indeterminacy, all of which need to be accounted for by a plausible theory of practical judgment. One of these sorts of scenarios is where *ex ante*, the best choice between two options is indeterminate, but where deliberation towards one of the options changes the underlying reality in such a way that *ex post*, the option arrived at has become the determinate right answer. Such a constitutive role for deliberation with respect to the truths about our reasons seems also prominent in Williams's account, especially in view of his appreciation for the role of the imagination. However, from the need to explain the vagueness or indeterminacy in certain matters of practical deliberation, it does not follow that our theory of deliberation must itself remain vague. There is no reason why a highly detailed, well-worked out theory of practical deliberation could not still produce the result that the correct answers to practical questions will be indeterminate in certain scenarios. The same applies to the idea that our deliberative activity itself might be partly constitutive of our reasons for action in certain scenarios. There is no reason why our account of what it means for deliberation to be valid should be vague in order to be able to account for this possibility.

The second concern, implicit in this passage, but nonetheless an important theme in Williams's work, is that we should be suspicious of any deliberative methods or normative ethical theories that attempt to reduce practical reason to the systematic application of a simple set of universal rules, principles or procedures. Although my purpose in this thesis is not

to take a stand on this matter, we shall see later on that the account that I will defend allows us to think of deontological, utilitarian and virtue-ethical approaches as "interpretative strategies" that may be more or less useful in different areas of practical deliberation.[5] Furthermore, I will acknowledge the important role that intuition and unconscious cognition have to play for us in order to better understand our normative reasons for action. However, from the fact that we should not be narrow-minded about the appropriate *psychological processes* of deliberation, it does not follow that we must be vague in our meta-ethical explanation of what it means to say that we have reason to act upon certain desires but not upon others. Even if we could spell out, for once and for all, with the utmost philosophical clarity, why certain desires could have a normative status that other desires lack, then we might still have to rely on methods much less spelled-out in order to judge which of our actual desires are the ones that have that status. The idea that our methods in *normative* ethics might be bound to be muddy and imprecise does not preclude us from at least striving for maximal clarity and explicitness in *meta*-ethics.

Nevertheless, there is a third reason why Williams might have wanted to remain vague in his meta-ethical specification of the deliberative function from the actual desires to the idealized set. His purpose in "Internal and external reasons" was first and foremost to argue that there could be no such things as external reasons. It was not to argue that there can be such things as internal reasons. The sort of view that Williams is attacking is the view that internal reasons are *not enough*, not the view that they are too much to ask for. The view that there are external reasons is the view that no account of internal reasons, regardless of its details and inner workings, could give us what we want in moral philosophy, and that we should therefore allow such things as external reasons. Against this view, Williams argued, first, that external reasons are impossible, and second, that there is nothing worth wanting in ethics for which internal reasons would not suffice. In keeping the deliberative function alluded to in the definition of internal reasons purposefully vague, Williams stays away from any restrictions that might be specific to certain accounts of the nature of internal reasons, but which do not follow from the general idea of an internal reason.

However, some philosophers may feel that internal reasons actually are too much to ask for. I am thinking of the skeptical sort of view according

[5] See section 7.4.4.

to which there is no rational way of rejecting any sort of desire, not even on the basis of other desires, simply because there is no rational way to take sides with respect to conflicting desires. This is what we might call a "flat" desire-based view of reasons for action, according to which every desire constitutes a reason for action, and the question which of these reasons are "normative" simply does not make sense.[6] The only measure of desire, on such a view, is simply that of strength: if two desires conflict, then the one that leads to action is apparently stronger, and that is the only relation of precedence among desires that reality allows. On such a view, moral reasons, for example, are not reasons derived from the construction of some ideal self on the basis of deliberative principles. Moral reasons are simply constituted by our "moral sentiments"—desire-type attitudes that happen to have a strong influence on our behavior as a result of either natural selection or cultural conditioning. But the only thing that counts in favor of such sentiments is that they are strong in those situations where they are strong. In other situations, however, where egoistic or violent sentiments motivate our behavior, there would be no way of arguing that the weaker sentiments of altruism or compassion would still be the more reasonable ones.[7]

The challenge posed by such a view is a fourth concern for our proposal, which we should distinguish from the three concerns mentioned above, and which is not addressed by Williams himself in "Internal and External Reasons"—perhaps because it might be specific to the motivational Humean assumption behind my current approach, although this is something I am not sure of.[8] In any case, if the type-1 dispositionalist is

[6]This is more or less the same as the radical account according to which the dependence relation between normative reasons and intrinsic desires is merely instrumental, which I mentioned and rejected in section 3.3.1. Note however that I rejected it for being a wrong interpretation of what metanormative Humeans usually *want* to defend. That does not mean, of course, that the view may not be considered as a *skeptical* alternative that non-instrumental metanormative Humeans must be able to defend themselves *against*. It is in that latter role that I am invoking the view in the present section.

[7]Note that such a view would be much more relativistic than the view that I have called "relationalism," which is one of the reasons why I prefer to use the term "relationalism" instead of "relativism" myself.

[8]In particular, I do not immediately see how the argument from thick concepts, which we discussed as part of a possible anti-Humean interpretation of Williams's views in section 3.5.1, would make one immune from this sort of skepticism. The motivational disposition implicit in my concept of bravery might clash with the disposition implicit in my concept of carefulness—why think there is anything over and above their occurrent relative motivational strengths to determine which of the two should precede the other on any particular occasion?

to meet this challenge, then he must be able to explain, and not merely promise, that there is a certain normative pressure that deliberation can bring to bear upon our intrinsic desires in a way that does not simply collapse into a "might is right" of the strongest desire at the moment in question. What are the normatively relevant relations between the elements of the subjective motivational set that deliberation should work with, and why? Does not any criterion that gives some desires a normative status at the expense of others, introduce some normative source beyond the desires themselves, thereby undermining the very idea of the relationalist solution?

4.4.2 Identification and Self-Disclosure

Here we may draw on the work of Harry Frankfurt. In his early essays, Frankfurt proposed that we should understand the different status of different desires in terms of how they relate to us as *persons*: some desires are "internal to the person," whereas others are "external to the person" (1971/1988b; 1976/1988c). When a desire is external to a person, and he is motivated by it, then he is motivated *against his will*. An example of this is the "unwilling addict" who does not really want to take the drugs but cannot overcome the addictive desire which motivates him to use the drugs anyway.

As Frankfurt put it, the unwilling addict does not *identify* with his desire for the drugs. In his view, a person faces two tasks whenever her desires pull her into different directions and she is not yet decided about what to do. First of all, she must *identify* with some of her desires—she must decide which desires express what she really wants as a person, and which do not. And second, she must *prioritize* among the desires she identifies with—she must decide what she wants *most* as a person. The second task implies that she may have to decide whether realizing one desire is more important to her than realizing another if they cannot be realized both, or not both at present, even though she does identify with both.

The idea that a person can really want something *as a person* is often taken to imply that if he wants something in this way, then what he wants must tell us something about *who he is*. It follows that if he acts upon

What is there to deliberate? I am not saying that there is no answer to this question—far from it—but only that it could still be asked, and that it would still be a sensible starting point for philosophical reflection.

what he wants in this way, his action *expresses* something about his identity, something which is characteristic of him being the person he is. Such an action, we shall say, exhibits *self-disclosure* (Watson, 1996/2004c, p. 261).

But what are we disclosing when we disclose our "selves" or our "identities" as persons? In his later work, Frankfurt has proposed that we should understand the self, or the identity of a person, in terms of what that person *cares about*. It is because we care about things in general, that we are persons, and it is what a person cares about specifically that determines his identity as an individual (1999c; 2004). In Frankfurt's view, caring is a complex motivational structure, and if a desire fits into that structure, then it expresses what the agent really wants.

Michael Bratman and Gary Watson have incorporated similar ideas about self-disclosure into their accounts of self-government. Roughly, their view is that the two go hand in hand: for an agent to really want something, or to care about it, is for that agent to adopt it as a goal for deliberation (Watson, 1996/2004c; Bratman, 2006). There is something to say for this view: clearly, when a person acts upon his self-adopted goals, this may tell us something about what kind of person he is. However, if we accept the Facts Principle and the Authority Principle, then it would follow from this view that what a person cares about must somehow be on a par with his *beliefs* about his normative reasons, and not necessarily with the *facts* about those reasons. Thus, it would follow that what we care about need not be what we *should* care about. But Frankfurt has complained that it does not make sense to ask what we should care about independently of what we actually do care about (Frankfurt, 2004, 23–28). Thus, Frankfurt seems to have taken a more or less Humean view about caring: what we should care about is itself a function of what we do care about. He concludes that caring is a *source* of reasons: it is on the basis of knowledge about what we care about that we should answer practical questions.

The challenge for such a view comes mostly from the direction of the Disconfirmation Principle: if the facts about our normative reasons are facts about what we care about, then it follows from that Principle that we can be mistaken about what we care about. Frankfurt embraces this conclusion (2006, p. 33–34, 49–50), but the question is whether he can account for it, and furthermore, whether mistakes about what we care about can fully explain the ways in which our practical judgments might be false. I will return to this matter in chapter 8.[9] For now, let us conclude

[9]See section 8.3.

that Frankfurt's ideas offer some prospect for a solution to the problem that we raised for the type-1 dispositionalist: the normative pressure that correct deliberation may bring to bear on some desires at the behest of others may be grounded in facts about how certain elements of an agent's subjective motivational set constitute what that agent cares about, while other elements may be considered external to him as a person and thus as having no intrinsic normative force.

Despite all this, there is a second problem for the type-1 dispositionalist, which is the problem that any form of relationalist cognitivism must solve: how to account for the Intersubjectivity Principle in the light of the claim that the content of a practical belief includes a reference to the agent having the belief. We have briefly discussed this problem in section 1.3.2: when different agents make conflicting practical judgments, then it seems that according to relationalist cognitivism, their beliefs do not really contradict each other, because each agent is merely making self-ascriptions about his own attitudes. This implication has seemed counterintuitive to many philosophers, including certain proponents of the dispositional approach. They prefer the alternative of a nonrelationalist version of the dispositional theory. In section 3.2 of the previous chapter, we briefly explored the difficulties that nonrelationalists must face. In the next chapter, we are going to discuss them in greater detail.

5 *The Nonrelationalist Alternative*

In the previous chapter I have outlined a dispositional solution to the Facts Problem according to which we should be relationalists about the intersubjectivity of practical reason. This is the view that I favor, and I will develop it in further detail in part III of this thesis. However, in the following two chapters I will first consider the alternatives to such a view. I will discuss nonrelationalist dispositionalism at length, because like the view from the previous chapter, it aims to provide what I have called a "reconciliatory" solution, accommodating all of our Principles from chapter 1. In section 5.1 I will discuss how this view attempts to solve the Facts Problem, and in section 5.2 what the challenge for this solution amounts to. Then, in section 5.3 we will see that in order to meet this challenge, two very different versions of the view may be distinguished, which I will call "type-II" and "type-III dispositionalism."

Type-II dispositionalism gives rise to many problems, which I discuss in the final section of this chapter, section 5.4. In chapter 6 I return to the type-III alternative. This will complete our overview of the *status quo* in the light of the principles from chapter 1. Although it is not my purpose to refute nonrelationalist accounts, I do intend to explain why type-I dispositionalism should not be easily dismissed, given the sort of problems that the other views are facing.

5.1 Nonrelationalist Dispositionalism

The idea of a nonrelationalist dispositional theory goes back at least to Firth (1952), and is being defended in the contemporary literature by Smith (1994, 2002/2004c) and Jackson & Pettit (1995). I shall focus on the version defended by Smith, but the problems we shall discuss apply equally, in my view, to any other account of this type. In a nutshell, the view is this: that P should be the case if and only if *every* agent would, under ideal conditions, desire in the resultant sense that P. If P is a proposition of the

form "A does ϕ," then, given our Principles, the view yields the claim that A has a normative reason to ϕ if and only if every agent would, under ideal conditions, desire that A would ϕ. So construed, however, the view may seem a bit odd, because first of all, it would be weird to require that under the ideal conditions, every agent would have to be acquainted with A, and second of all, it would seem strange why the ideal desire set of every agent would have to feature desires tailored to A. Therefore, in order to make sense of the view, we might want to add a requirement such as "when made to consider whether A should ϕ" or something to that effect.

However, in Smith's view it is an essential feature of reasons for action that they are universalizable along the following lines: that whenever A has a normative reason to ϕ, this is in virtue of certain circumstances in which A finds himself, such that *every* agent would have a normative reason to ϕ under the same circumstances. Therefore, the desires of our ideal selves about the actions of ourselves and others need not be understood, on Smith's view, as desires about individual, particular persons. Instead, they may be construed as desires about circumstances: that *any* agent would ϕ under circumstances C and that *no* agent would ψ under circumstances D. This means that we can reformulate the nonrelationalist dispositional account of normative reasons as follows, for any agent A, action ϕ, and set of circumstances C: A has a normative reason to ϕ under C if and only if every agent would, under ideal conditions, when made to consider C, desire that any agent would ϕ under C.

5.1.1 The Nonrelationalist Solution to the Facts Problem

The solution that this account provides to the Facts Problem is similar to the solution provided by the relationalist, type-1 dipositionalist view that we discussed in the previous chapter, although there is a slight difference. Again, the general idea is to remove the mystery of why agents under ideal conditions would be motivated in accordance to their normative reasons by analyzing those reasons in terms of their desires under those conditions. In the case of the relationalist version, we explicated this solution by disambiguating the apparently contradictory result of the Facts Problem into two consistent and perfectly compatible claims. The first claim was that a self-governing agent who knows his normative reasons could have had different intrinsic desires and still remain self-governing and knowledgeable, but only if his practical beliefs had also been different.

The second claim was that a self-governing agent who knows his normative reasons could have had different intrinsic desires without having different practical beliefs, but then would no longer have been a self-governing agent who knows his normative reasons.

Now, in the case of the nonrelationalist solution, we would have to change these two claims somewhat. The second claim needs the least alteration: the possibility of an agent having different intrinsic desires while having the same practical beliefs still obtains. The implication that the agent would then no longer be self-governing also still follows. But when it comes to moral (i.e., taste-independent) practical content, then the nonrelationalist will want to claim that such an agent would still be as knowledgeable as before; the object of his knowledge would not be different if his intrinsic desires were different.

The first of the two claims must be qualified in a much more significant manner in the case of a nonrelationalist solution. Because according to the nonrelationalist dispositionalist, there are intrinsic desires with contents of the form "that anyone will ϕ under circumstances C" that all agents would have under ideal conditions. And therefore, with respect to those intrinsic desires the first of the above two claims no longer holds: the nonrelationalist dispositionalist must simply *deny* that a self-governing agent who knows his normative reasons could have desired differently while remaining self-governing and knowledgeable of the relevant fact about normative reasons. Therefore, the nonrelationalist dispositionalist solution involves the claim that, for certain intrinsic desires, there are conceptual reasons why ideal agents could not possibly desire otherwise, despite the conceptual distinctness of these desires from matters of belief. This means that we do not really have a full solution yet. Like the relationalist proposal, the nonrelationalist solution depends upon a kind of promise: the promise that such conceptual reasons can be provided for these desires.

5.1.2 Weak Dependence

In fact, it might seem that this must now apply to all the desires of my ideal self, and that they do not depend upon the desires of my actual self at all, on this view, since the ideal self of every other agent is supposed to have the same desires, regardless of how the actual desires of that other agent might have been different. However, that conclusion does not follow, because some of my desires may be elements of the relevant circumstances

of my actions in non-moral cases where my particular tastes are relevant, as we already discussed in section 1.3.2. Thus, every agent might desire, under ideal conditions, that all agents who prefer green shall paint their walls green, and all agents who prefer white shall paint their walls white. If I prefer white, then that might give me a normative reason to paint my walls white, since even the ideal selves of agents who do not prefer white would still desire that the walls in my house would be white, if not in theirs. In terms of our terminology from section 1.3.2, what Smith's view requires is *similarity* between ideal selves, not *isomorphism*. Therefore, the first of the above two claims that resolve the paradox of the Facts Problem still applies to intrinsic desires that are matters of taste in this way: according to the nonrelationalist dispositionalist, a self-governing agent who knows his normative reasons could have had different tastes, and therefore different intrinsic desires that would reflect these different tastes, while remaining a self-governing agent who knows his normative reasons, provided that some of his beliefs would also be different (such as his beliefs about what color he should paint his walls).

Nevertheless, nonrelationalism does put severe restrictions on the way that our normative reasons may depend on our actual attitudes. First of all, if they depend on attitudes that others need not share, not even under ideal conditions, then there must be something essentially indexical about them. Their not being shared by others is brought out only if we represent them in a form like "I desire to $\phi(me)$." Thus, if I have a normative reason to paint my walls green, then other agents do not have to share my attitude in the sense that under ideal conditions, they would not have to be able to truthfully utter the sentence "I desire to paint my walls green." Second, a normative reason can only depend on such an 'indexical' attitude if it is 'backed up,' as it were, by a corresponding 'extensional' attitude: a resultant desire that any agent x will $\phi(x)$ if x has the aforementioned indexical attitude towards ϕ-ing and if x's circumstances do not count against ϕ-ing in any other relevant respects. Third, and most importantly, the normative reason can only depend on the actuality of the indexical attitude if the ideal selves of *all* agents would have that corresponding extensional attitude, *regardless* of whether those agents *actually* have that attitude.

In other words, according to nonrelationalist dispositionalism, my actual attitudes provide me with normative reasons only if those who do not share those attitudes would still have a reason to wish me the best of

luck in acting upon those attitudes. Let us call this type of dependence upon actual attitudes "weak dependence." In contrast, according to the relationalist dispositionalist, it is possible that my normative reason to ϕ depends on my attitudes in such a way that an agent who does not share those attitudes might have a normative reason to disapprove of my doing ϕ. Let us call that "strong dependence." Thus, the crucial disagreement between these two forms of dispositionalism is about whether normative reasons for action strongly depend on our actual attitudes.

5.1.3 The Advice Model

Through his deployment of the notion of circumstances, Smith has tried to incorporate as much actual attitude dependence into his account of normative reasons as he possibly can within the nonrelationalist restrictions that I have outlined. I have called this "weak dependence," and so far we have only seen examples of it that involved differences in taste. However, weak dependence allows another type of example that we have not touched upon yet: normative reasons may also depend on actual dispositions that constitute a lack of self-government. Smith discusses the example of a person who's suffered a humiliating defeat in a game of squash, and who feels intense anger and frustration as a result of this, which might lead him to smash his opponent in the face if he were to approach him close enough (Smith, 1995/2004d, p. 19). According to Smith, the ideal self of this person would presumably not have this problem at all, because he is "fully rational." In my view, this would rather be a matter of the ideal self being fully self-governing, but the result is the same: since there is no reason for the ideal self not to approach his opponent, we may suppose his ideal self would indeed approach him, out of good sportsmanship, in order to congratulate him on his victory. In contrast, the actual less-than-ideal person might not have a normative reason to approach his opponent. Instead, it might be wiser for him to take his own imperfect self-government into account, and to just leave as quickly as he can, without making a scene.

Smith argues that we should distinguish between the desires that the ideal self of this person would have *regarding himself* and the desires that the ideal self has with respect to his less-than-ideal self. In order to accommodate this distinction, Smith proposes an "advice model" of the dispositional solution: I have a normative reason to ϕ if and only if my

ideal self has a resultant desire that I—that is my actual, less-than-ideal self—would ϕ. Thus, I have a normative reason to do what my ideal self would 'advise' me to do. If our ideal selves would walk the earth, we would have to do as they say, not as they do.

But of course, they do not walk the earth. Smith conceptualizes the advice model in terms of a distinction between two possible worlds, the *evaluated* and the *evaluating* world. The former contains the actual self, while the latter contains the ideal self. We can then distinguish between what the ideal self would do in the evaluating world, and what the ideal self would have the actual self do in the evaluated world. I have some problems with this, because I do not think that ideal selves are possible, and so I do not think we can quantify over possible worlds in which they exist. I rather view the ideal self as a helpful construct, a perspective from the point of view of perfection that we can reason about in a manner similar to the way we reason about the limits of functions in mathematics—there is no possible value that instantiates the limit, but under the appropriate conditions, the limit outcome can be determined nonetheless.[1]

What I can agree with, however, is that the construct of an ideal self is helpful along the lines of the advice model, and that our imperfections need to be taken into account in order to determine what normative reasons for action we have. Note, by the way, that the advice model is not exclusive to nonrelationalist dispositionalism. Relationalist forms of dispositionalism, including our type-I dispositionalist proposal, may be understood along the lines of the advice model as well.

5.2 Nonrelationalism and Conceptual Possibility

According to nonrelationalist dispositionalism, it should be the case that P if and only if every agent would, under ideal conditions, desire in the resultant sense that P. It is important to understand that in this claim, "every agent" means every *conceptually possible* agent who is capable of having normative reasons for action. Because according to the dispositional approach, for every such agent we should be able to reason about a corresponding ideal self—a 'nearest' idealized version of the agent, a modification in which all his flaws have been adjusted for. What this means is that not only Nazis and psychopaths would have to share our values under ideal

[1] I will return to the matter of what exactly to make of an "ideal self" later on in this thesis.

conditions, but also extraterrestrial aliens whose psychologies might be as different from ours as they could possibly be. No matter what their actual "subjective motivational sets" would look like, if we would adjust for all the flaws therein, they would have to end up promoting the same values that we would end up promoting if the flaws in our motivational sets were corrected for. And because we are talking about conceptual rather than physical possibility, we should even include wholly fictional creatures, so long as they remain the sort of creatures capable of having normative reasons for action and their descriptions are conceptually coherent. Devils and demons, Sauron and the mighty Cthulhu—under ideal conditions of agency they would all be motivated to help those in need, to promote equal rights for minorities, and to incorporate environmental concerns into their intentions and policies. In terms of Smith's advice framework, if Cthulhu inhabits a possible world, then there is an evaluating world in which Ideal Cthulhu is feeling *really* sorry about the acts of his lesser self.

Note that this is not a consequence of nonrelationalism in general. The nonrelationalist externalist can simply claim that fictional creatures like Cthulhu represent the very possibility that externalists have been stressing: of a self-governing creature that simply does not care about being moral, or perhaps even desires to embrace evil for evil's sake. In fact, externalism about good and evil seems to be the 'preferred meta-ethics' of most fantasy literature: supernatural creatures are usually members of either Team Good or Team Evil, and the captains of Team Evil can be just as self-governing and aware of the evilness of their projects as their goody two shoes opponents.

However, despite this built-in externalism in the fantasy literature, it is not a general consequence of dispositionalism either that Cthulhu would have normative reasons to donate to famine relief. Because as long as dispositionalists are also relationalists, they can hold that what is referred to as 'good' and 'evil' in such stories is not to be understood in a practically normative sense, but rather in a purely descriptive sense, such that it may be practically normative for some agents to do the 'evil' thing. In the *Dungeons and Dragons* universe, for example, 'good' and 'evil' are best understood as a kind of supernatural energies that a creature can be 'aligned' with. And for an 'evil-aligned' creature, 'evil' is just as normative as 'good' is normative for a 'good-aligned' creature. From the type-1 dispositionalist perspective, we might say that the ideal selves of 'evil-aligned' creatures will desire their lesser selves to be as 'evil' as they

can. Of course, we do not believe that there is room for such supernatural properties in the actual world, but that is what makes it fantasy fiction.

The idea that all conceptually possible deliberators must have normative reasons to uphold the same values is, therefore, a specific implication of the combination of dispositionalism with nonrelationalism—the view defended by Smith. The implication may be summarized as follows: Smith must either bite the bullet and admit that Cthulhu *would*, under ideal conditions, use his powers to bless people with happy dreams rather than nightmares, or he must claim that creatures like Cthulhu are not conceptually possible. In order to block the latter option and to avoid Lovecraft exegesis, let us now consider a different type of creature.

5.2.1 The Case Against Nonrelationalist Dispositionalism: Mars Attacks!

In the comic movie *Mars Attacks!* our planet is invaded by aliens who lure us into thinking they come as friends, and then crush us with their advanced technology. A technology that also allows them to perform the most perverted experiments on human beings, animals, and combinations of both. Why are they doing this to us? First of all, it seems they simply have no disposition whatsoever to care about our well-being. But second, they have a further, positive reason to treat us like this: they find it simply *hilarious*. The Martians are driven by their sense of humor, and not susceptible to any mercy that would hold them back. Their psychology, although cruel and inhuman, seems conceptually possible. In fact, even though *Mars Attacks!* does not obey the laws of physics as we know them, I can think of no reason why it would even be physically impossible for alien beings to have these psychological features. Perhaps there are *exobiological* reasons why the evolutionary history of such creatures would require some pretty bizarre physical and social environments, which might not exist anywhere in the universe, but that does not make them *physically* impossible in the nomological sense.

With respect to the Martian invaders, then, Smith seems committed to the claim that there must be rational flaws within their sets of attitudes such that, if these flaws were corrected, they would no longer desire to torture us. This sort of claim is very tough to defend, for there seems to be no conceptual reason why the Martians could not be utterly without the disposition for mercy, and if they do not have such a disposition whatsoever, then it seems hard to explain how purely rational considerations could

make them adopt a merciful attitude after all.[2] Several of Smith's critics have rejected the idea that rationality alone could have such implications.[3] Let us now turn to Smith's options for defending this idea.

5.3 Smith on Systematic Justification

Smith defends nonrelationalist dispositionalism by extending the notion of "correct deliberation" in Williams's account of internal reasons (Smith, 1994, pp. 155–161, 164–174; 1995/2004d, pp. 17–18, 20–34). Indeed, as Smith presents his view, it involves "an endorsement of the claim that all reasons are 'internal,' as opposed to 'external,' to use Williams's terms" (1995/2004d, p. 17). Clearly, then, Smith thinks of himself as an internal reasons theorist. In chapters 2 and 3 I have been sketching how the Internal Reasons View can figure in a chain of arguments from a Humean theory of motivation to the conclusion of relationalism, flagging various premises along the way that nonrelationalists might want to deny. Before I attempt to classify Smith's brand of nonrelationalism within this framework, however, let us first discuss his account of deliberation in his own terms.

The crucial aspect of deliberation that he thinks Williams has underestimated is that of "trying to find out whether our desires are *systematically justifiable*" (Smith, 1994, p. 158–159). What does he mean by this?

> I mean just that we can try to decide whether or not some particular underived desire that we have or might have is a desire to do something that is itself non-derivatively desirable. And we do this in a certain characteristic way: namely, by trying to integrate the object of that desire into a more coherent and unified desiderative profile and evaluative outlook. (1994, p. 159)

By saying that he means "just" this, Smith seems to be suggesting that the first sentence contains a fairly trivial and innocent, perhaps even deflationary, explication of the sort of justification that he has in mind. However, I find the notion of something "that is itself non-derivatively

[2] Compare Street (2009, p. 293): "If a group of intelligent, ideally coherent aliens descended upon us and began trying to kill us for food or torture us for sport, would we feel intuitively convinced that they were making a mistake about the normative facts? Would we be more inclined to say 'They shouldn't be doing this' or rather just 'How can we stop them?'"

[3] E.g. Sobel (1999) and Enoch (2007), which I will discuss in section 5.4 below.

desirable" extremely puzzling, and I doubt that many people would simply nod their heads in agreement if I would tell them that this is what they try to get at when they deliberate. Instead, the first sentence seems loaded with one or two philosophical claims that are precisely at stake in our discussion. The first claim is that deliberation is about desirability regardless of who is making the practical judgment. This claim simply builds a nonrelationalist semantics of practical judgment into the account of deliberation, which begs the question against the relationalist. The second claim is that the notion of deliberation already presupposes a notion of "desirability in itself" which makes the dispositional solution a non-reductive one at best, and a viciously circular one at worst. After all, the dispositional approach explains desirability in terms of what an ideal agent would desire, and the Internal Reasons View in turn explains the desires of an ideal agent in terms of how that agent would deliberate—so if explaining correct deliberation would presuppose a notion of desirability in itself, then we would be back at square one.

For the moment, however, let us assume that this is just sound non-reductionism about normativity, and take a look at the second sentence. How does one integrate a desire into a "more coherent and unified" profile, and why would that be more rational? The basic idea, as I understand it, is that we should try to improve the simplicity and regularity of the set of our intrinsic desires, such that we end up having many derived desires that all derive from a few very general ones:

> Suppose we take a whole host of desires we have for specific and general things; desires which are not in fact derived from any desire that we have for something more general. We can ask ourselves whether we wouldn't get a more systematically justifiable set of desires by adding to this whole host of specific and general desires another general desire, or a more general desire still, a desire that, in turn, justifies and explains the more specific desires that we have. (1994, p. 159)

Note that this is not just a matter of removing logical inconsistencies from the desire set: even a desire set that is fully consistent might be eligible to the above sort of improvement. Furthermore, once a more general desire has been added, new derived desires may be arrived at which had not been part of the original set, motivating new behavior. A desire set that exhibits this kind of generality is "more unified," in Smith's view, than a

desire set that just contains a jumble of various logically compatible but highly specific and independent desires.

5.3.1 The Analogy Between Desires and Beliefs

But why, we may now ask, would it be rational to prefer such unity? The answer, according to Smith, may be found in a simple analogy with beliefs: in the case of beliefs we also prefer a system of beliefs that exhibits this kind of unity. The paradigm example, I suppose, would be that of parsimony in empirical science: we prefer a theory with a few very general laws to a theory that contains just a jumble of many different and more specific laws. The big question, of course, is whether this analogy really makes sense, and whether it makes the inference valid that what is rational for beliefs must be rational for desires.[4]

If it would be rational to strive for desire sets that are "*more* coherent and unified" in this sense, then we may infer that our *ideal* selves would have, as Smith has put it elsewhere, "*maximally* coherent and unified" desire sets (1996, p. 160; my emphasis). At this point, however, the analogy between desire sets and scientific beliefs begins to work against Smith's case. Because surely, in science, multiple maximally coherent and unified belief sets are conceivable, which contradict each other on substantial questions about the way the world is, and it is a matter of contingent empirical fact that most of these belief sets are false. However, Smith's nonrelationalist dispositionalism requires that deliberation will decide between conflicting desire sets on the basis of rational considerations alone—considerations which must be valid *a priori*, because they must be valid for *conceptually possible* deliberators. Hence, Smith is asking much more from rationality with respect to desires than scientists expect from rationality in the case of beliefs. In contrast, if the constraints exerted upon desire sets by the requirement of maximal coherence and unity were comparable to those exerted upon beliefs by a similar requirement in the case of empirical science, then we should expect the ideal selves of all conceptually possible deliberators to have widely different and severely conflicting maximally unified desire sets.[5]

If empirical science proves to be dis-analogous in this manner, then perhaps we should pursue an analogy with *mathematics* instead, which

[4]For criticism of this analogy, see Enoch (2007, section 2, pp. 103–105).

[5]Similar problems have been raised for Smith's view by Sobel (1999) and Enoch (2007, section 3, pp. 105–108).

does seem to involve settling matters of belief on conceptual grounds alone. The idea of understanding ethics in comparison to mathematics has a rich philosophical tradition, of course, but it is not clear to me that such a comparison can be used to *support* the idea that disagreements between conflicting desire sets can be settled on the basis of *a priori* considerations alone. Rather, such an analogy seems to presuppose this idea. Since our web of belief in general requires empirical justification, focusing on the small subset of those beliefs that could be settled conceptually would no longer be motivated by the wish to maintain a general symmetry between beliefs and desires, but rather by the specific idea that desires are more like *a priori* beliefs than like empirical beliefs, which does not follow from the idea of a general symmetry at all, but rather presupposes the idea that needed to be justified: that desire criticism must be *a priori*.

Furthermore, even if the analogy with mathematics could be supported, then it is not at all clear that it will even provide the sort of coherence constraint that Smith is looking for. At least at first sight, mathematicians employ deductive reasoning in order to determine which conclusions follow from their premises, and such reasoning does not involve the application of a principle of parsimony. One might suggest that mathematicians often work abductively in a manner that resembles empirical science, however, and that their aim is really to find the most parsimonious set of axioms that will still allow them to prove desirable theorems about mathematical concepts that strike us as intuitively plausible. A Platonist about mathematics might hold that such axioms are themselves *a priori* truths, and such a view would at least provide a mathematical analogue to what Smith seems to have in mind for ethics. Nevertheless, mathematical Platonism is deeply problematic, and many philosophers would rather argue that the only truths established by mathematics are truths about which theorems follow from which axioms, not about which axioms are correct in themselves. The merit of parsimonious axiomatization schemes, on this view, lies not in their Platonic correctness, but rather in their usefulness for the application of mathematics to real-life problems and situations.

I shall return to the question of whether ethics should be understood as a formal discipline (like mathematics) below. For now, the most important conclusion is that, even if it is plausible that desire sets should be like belief sets in the sense of being subject to a demand for maximal coherence and unification, then it is still not at all clear why this should provide us with one set of desires that all conceptually possible agents are rationally

required to have. In most, if not all, areas of belief formation such considerations of unity and coherence leave the choice between various sets of belief undetermined. Hence, further argument would be needed to show that coherence and unity can do so much more in the field of ethics.

5.3.2 *Type-II vs. Type-III Dispositionalism*

It is hard to see, therefore, how merely extending the repertoire of rationality from strict logical consistency towards such requirements as "maximal coherence and unity" could block the line of reasoning from motivational Humeanism towards the relationalism that I have been sketching in the previous chapters. In the end, even maximally coherent and unified desire sets seem to allow for more diversity than nonrelationalism permits. Let us now try to situate this problem within the framework that I have developed, and figure out which premises Smith would be forced to deny.

As it turns out, the view that a demand for maximal coherence and unity will push us towards a single normative desire set, which I have discussed in Smith's own terms in the previous subsection, may now be translated into two rather different theoretical options within our framework. As we have seen in section 2.4, the defense of the Internal Reasons View required a procedural understanding of normative reasons for action, which I could derive from the procedural interpretation of my Disconfirmation Principle in conjunction with the Authority Principle. Because the argument for relationalism was in turn built upon this defense, the *nonproceduralist objection* against this defense might also be employed to undercut the relationalist argument. In the light of this, Smith's extension of the Internal Reasons View in terms of his notion of systematic justification may now be understood in one of two ways. Either it remains within the boundaries of proceduralism, in which case it must attempt to block the argument towards relationalism without making use of the nonproceduralist objection in this manner. Or it revises the Internal Reasons View in a more fundamental manner, by dropping proceduralism and understanding nonrelationalist practical normativity in a nonproceduralist way.

Let us call the former option "type-II" and the latter "type-III dispositionalism." We have now divided the motivationally Humean dispositionalist approach into three versions, depending on whether we accept proceduralism and/or relationalism:

dispositionalism:	*proceduralism?*	*relationalism?*
type-I	yes	yes
type-II	yes	no
type-III	no	no

Because relationalism does not seem to entail proceduralism, a fourth motivationally Humean variety of dispositionalism may be defined, which rejects proceduralism while accepting relationalism. From the perspective of our current discussion, this view does not seem very attractive however, as it combines the burdens of both the type-I and type-III views without offering in return any obvious advantage over either. Perhaps there could be independent reasons for wanting to defend such a view, but if there are, then they are beyond the scope of this thesis. I will therefore proceed on the assumption that the types I, II, and III are the most plausible candidates for a dispositional, reconciliatory solution to the Facts Problem.

Because of Smith's explicit allegiance to the Internal Reasons View, and because of the implicit proceduralism in at least Williams's defense of the view, one might be inclined to interpret Smith's proposal as proceduralistic. Furthermore, there are several passages in Smith's work, both in *The Moral Problem* and in subsequent articles, that seemed to me to suggest a type-II account, especially where he employs the notion of "convergence" (more on this below). However, from personal communication it now seems to me that Smith actually means to defend a type-III account. In his view, in order for us to be able to know what we should do we must be blessed with a kind of *epistemic luck* that some of us might actually lack.[6] For the epistemically unlucky, there are no internal justifications which, from their own point of view, could rationally compel them to see the errors of their ways.

5.3.3 Proceduralism and Epistemic Luck

I think this appeal to the notion of epistemic luck is interesting, but it needs to be qualified in order to properly distinguish between type-II and type-III dispositionalism. The type of epistemic luck that we are interested in is known as "veritic luck," which may be characterized as luck on the part of the agent that he has adopted a true belief, given the internal justification that he has for that belief (i.e. the justification from his own perspective). Much of the discussion on veritic luck has focused

[6] Again from personal communication.

on beliefs about contingent facts, and a popular way to define it is in terms of possible worlds in which those facts are different: an agent is veridically lucky, according to such a definition, if (a) he holds a belief that is true in the actual world and (b) his internal justification for this belief leads him to hold the same belief in nearby possible worlds in which it is false (Pritchard, 2005).

However, such a "world-centered" definition does not allow for veritic luck with respect to beliefs that are true in all possible worlds, including beliefs that are necessarily true on *a priori* grounds. In order to accommodate such "armchair luck," as he calls it, Nenad Miščević has proposed an "agent-centered" account of veritic luck (2007). The idea behind agent-centered accounts is not that the same belief might have been false, but that the same agent might have held a different belief on the basis of the same, or a sufficiently similar, internal justification. Interestingly, Miščević uses the term "procedure" to capture this idea of a way of thinking that sufficiently similar internal justifications have in common, which fits in nicely with our distinction between proceduralism and nonproceduralism, as we shall see below. His proposal for "procedural veritic luck" is as follows:

> Procedural veritic luck: It is a matter of luck that the procedure used by the agent has resulted in true belief.
>
> The agent's belief is true and has been justifiably arrived at in the actual world, but in a wide class of nearby possible worlds in which the relevant initial conditions are almost the same as in the actual world—and this will mean, in the basic case, that the agent at the very least forms her belief in the sufficiently similar way as in the actual world—the agent arrives at a false belief (or no belief at all). (2007, p. 61)

Note that this account allows for armchair luck: the procedure that I followed in the actual world to arrive at a belief that is true *a priori* might have resulted in a different belief, and one that is false, if the initial conditions of my efforts of reasoning had been slightly different.

Before we can apply this account to the case of type-III dispositionalism, however, a further qualification must be made. For much of the discussion on epistemic luck is held in the context of the more general epistemological problem of how to account for the possibility of knowledge for agents with *finite* and *imperfect* cognitive capacities, like us. Put differently, it is

about the problem of how to come up with a standard for justification that is on the one hand high enough to license talk of *knowledge*, while on the other hand keeping that standard low enough so as to include the sort of justifications that we are actually capable of giving for our beliefs. This problem does not only arise with respect to knowledge about the muddy empirical world, but also with respect to the crystal realm of the *a priori*. Miščević gives the example of a mathematician, Jane, who makes two mistakes in her calculation that cancel each other out, resulting in the correct solution. The errors are "extremely hard to detect, so that (...) Jane is justified in trusting her calculation" (p. 49). The point of the example, of course, is that we never know that we are making such mistakes until we discover them, which means that from our own perspective, we can never know whether our justifications are flawed. Hence, we must either accept a standard for justification that allows some such possibility for error, or conclude that we're never in a position to know that we know.

The Jane example is an example of armchair luck, but it is not an example that can serve as an analogy for Smith in order to clarify type-III dispositionalism. Because surely, it is less than ideal for an agent to make these kinds of mistakes. In other words, the ideal self of Jane would not be susceptible to this type of error, and therefore, neither to this kind of luck. This means that we must make a further distinction: between veritic luck that we are susceptible to in virtue of our imperfect capacities, and veritic luck that even our ideal selves would be susceptible to. Let us call the former "eliminable" and the latter "ineliminable" luck. Of course, the point of ineliminable veritic luck is not that our ideal selves might still have beliefs that happen to be false, because ideal selves hold true beliefs by definition. What ineliminable luck means, rather, is that our ideal selves have 'unlucky twins,' so to speak, who share the cognitive capacities of their lucky siblings, but nevertheless arrived at false beliefs instead.

However, in order to explicate the latter idea using Miščević's procedural account, we would once again have to place ideal selves inside possible worlds, assuming that ideal selves are possible agents. For the sake of explicating Smith's view, this is not a problem, as Smith is already committed to the idea that ideal selves are possible agents. Nevertheless, I did express skepticism about the possibility of ideal selves, earlier on, and I do not think that the case for nonrelationalism should depend on it.

In order to circumvent this problem, we might want to think of ineliminable luck as a type of veritic luck that we must *necessarily* have, for

conceptual reasons, if our beliefs are to be true. This way, we do not need to refer to ideal selves at all. Then, using Miščević's procedural account, ineliminable veritic luck may be defined in terms of the absence of a possible world in which an unlucky twin manages to correct his false belief on the basis of procedures that could be justified from his own perspective at the time when he still held that false belief. Undoubtedly this will give rise to various technical issues that need to be resolved in further detail, but to do so would go beyond the scope of this thesis.[7] For the sake of the argument, let us for now simply grant the type-III dispositionalist that the general concept of ineliminable veritic luck can be worked out properly, and focus instead on the question of whether it makes sense in the case of practical judgments. In the light of this and in the interest of simplicity I will also continue to speak loosely about veritic luck in the case of ideal selves.

With this terminology in place, we may redefine the procedural view of disconfirmation as the view that practical beliefs are not susceptible to ineliminable veritic luck. In contrast, the nonprocedural view holds that practical beliefs are susceptible to this kind of luck. Note, however, that proceduralism is perfectly compatible with the idea that practical beliefs may be susceptible to *eliminable* veritic luck. Note also that proceduralism is compatible both with the view that eliminable veritic luck does, and with the view that it does not, rule out knowledge. The same goes for nonproceduralism, but nonproceduralists must at least accept that *in*eliminable veritic luck is compatible with knowledge. Otherwise they would have to give up the idea of knowledge in matters of practical normativity.

In the case of nonrelationalist dispositionalism, we may narrow this down a bit further to a dispute about the existence of ineliminable *armchair* luck. According to type-II dispositionalism, practical normativity does not give rise to this type of luck. In contrast, type-III dispositionalism is the view that in order for our practical beliefs to be true, we must be blessed with ineliminable armchair luck. Furthermore, in order to avoid epistemological skepticism, the type-III dispositionalist must defend the claim that ineliminable armchair luck is compatible with knowledge.

With this further clarification of the difference between type-II and type-III dispositionalism at hand, consider how Smith originally formulated some of his premises at the beginning of *The Moral Problem*:

[7] An alternative would be to reformulate the agent-centered account of veritic luck in such a way that it does not invoke possible worlds in the first place.

> [T]here exists a domain of moral facts; facts about which we can form beliefs and about which we may be mistaken.
>
> Moreover, the way in which we conduct ourselves in living the moral life seems to presuppose that these facts are in principle available to all; *that no one in particular is better placed to discover them than anyone else*. That we have something like this conception of moral facts seems to explain our preoccupation with moral conversation and moral argument on the one hand, and novels and films in which the different reactions people have to moral questions are explored on the other. (1994, p. 5, my emphasis)

To me, this passage reads like a straightforward denial of ineliminable epistemic luck in moral deliberation. But perhaps Smith changed his mind on this matter, or perhaps the passage can be read in a different way. Nevertheless, we can think of the distinction as presenting him with a dilemma, because as we shall see below, both accounts have their own difficulties. Either he must accept proceduralism, and face the problems of the type-II account, or he must reject proceduralism and solve the difficulties that the type-III account will give rise to. I will finish this chapter with a discussion of the former account in the next section, and return to the latter in in the following chapter.

5.4 Problems for Type-II Dispositionalism

There are two sets of problems that I want to raise for type-II dispositionalism. The first set of problems is specific to this type of account, and concerns the worry that it does not have the resources that are needed to make nonrelationalist realism plausible. I have already briefly discussed these problems, and some attempts to solve them, in section 3.2. I will now discuss arguments put forward by David Sobel and David Enoch that undermine such attempts. These arguments support the *anti-formal* content skepticism which I introduced in section 2.5.2. I will elaborate on their criticisms, but even though I myself consider them to be very strong, I do not mean to present them as conclusive. My current purpose is merely to explain why, in view of the literature, type-II dispositionalism does not provide us with an easy way out of the *status quo* surrounding the Facts Problem.

The second set of problems concerns the account of disconfirmation that type-II dispositionalism requires, which I shall call the "Principles of Reason View," and which type-II dispositionalists share with proponents of other views, as it is a fairly standard account of disconfirmation. Nevertheless, as I will argue in chapter 7, the Principles of Reason View faces serious problems.[8] Furthermore, there are independent reasons for accepting an alternative account of disconfirmation that I shall propose, the "Affective Response View," which is hard to combine with type-II dispositionalism. However, all of this must await chapter 7. For now, I will restrict my discussion to the first set of problems.

5.4.1 Problems for the Formalistic Convergence Strategy

If Smith means to defend a type-II dispositionalist view, then his account would be a nonrelationalist, motivationally Humean version of what I have in section 3.1 called the "Comprehensive Internal Reasons View" (or CIRV), which combines the Internal Reasons View with proceduralism and rejects the 'no reasons' form of externalism (in my framework through the Authority Principle, in Smith's own framework by means of his Practicality Principle). Recall also that my definition of CIRV is in line with Smith's formulation of the Internal Reasons View in that it does not accept Williams's proximity requirement on instrumental deliberation (a choice that I defended in section 2.2.1). In order to see whether nonrelationalism can be made plausible on the basis of CIRV, I discussed two strategies in section 3.2: the convergence and the constitution strategy. With respect to convergence, I made a further distinction between a belief-based convergence approach which required an anti-Humean theory of motivation, and a 'formalistic' convergence strategy that would also be available to motivational Humeans. I then discussed the first approach, as defended by Thomas, in some detail, but the second approach I mentioned only briefly. Instead, I shall now focus on the second approach, as it is the only type of convergence available to the type-II dispositionalist given his commitment to the Distinctness Principle. Below, I will also return to the constitution strategy, and to the prospects for a strategy that combines both aspects of convergence and constitution.

But first, let us briefly rehearse why only 'formalistic' convergence can aid the motivationally Humean nonrelationalist. According to Thomas's

[8] See section 7.1.

anti-Humean convergence view, there are substantial beliefs that produce motivations without the help of intrinsic desires, and it is in the light of convergence of such beliefs under appropriate conditions that motivational convergence may be expected under such conditions as well. But the motivational Humean cannot explain convergence of motivations in this manner. Instead, he must explain why rational deliberation will change the *intrinsic desires* of agents in ways that lead to convergence between them. Perhaps he can still argue on the basis of an *analogy* between beliefs and desires, as Smith has proposed, but we have seen in the previous section that such an analogy would only explain convergence across conceptually possible worlds if we restrict the analogy to the small subset of beliefs that are true or false *a priori*, and such a restriction demands a further argument for the very idea that is at stake here: whether or not intrinsic desires are the sort of things that would converge on the basis of conceptual requirements alone. Hence, if the type-II dispositionalist wants to argue for convergence, then he must employ a formalistic convergence strategy, which involves the idea that no substantive information gets imported into the subjective motivational set of the agent during the process of convergence, and which therefore rather aims to show that through reflecting on the formal consequences of the elements already in the S, the required motivational changes can be achieved.

Even though the distinction between what counts as "formal" and what as "substantive" in this context may ultimately rest on an intuitive understanding of these notions, there is one crucial ambiguity in the distinction that we can and must address at this point, because it involves once more the difference between proceduralism and nonproceduralism. Recall that when I defined proceduralism, I hastened to add that it does not require 'procedures' in the sense of discrete sequences of concrete steps that we can capture in a formalized system. This same disclaimer now also applies to the sense in which I have been using the term 'formal' itself: the formalized convergence strategy is not meant to be 'formal' in the sense of formalized language. It is merely 'formal' in the sense of 'lacking substance,' i.e., being due entirely to conceptual relations and considerations. However, even with this clarification made, we might still think of that sense of formality in two ways.

The first way in which we might understand it is that, even though we might not be able to fully capture formal considerations as discrete steps in a finite procedure, we may nevertheless think of them as requirements

of rationality that could in principle be justified to an agent from the perspective of his prior attitudes to which these requirements need to be applied. This understanding of some requirement's being formally rational captures the idea that it must lack substantive content in terms of its being internally justifiable from all conceptually possible sets of prior attitudes. Thus, its internal justification does not depend on the specific content of the agent's prior attitudes, and in that sense lacks all substance. It is therefore consistent with a proceduralist understanding of disconfirmation: every set of attitudes that violates a formal requirement in this sense is eligible to disconfirmation in the proceduralist sense.

By contrast, the second way in which some requirement might be said to be "formal" is simply that the requirement be valid *a priori*, but without being committed to the idea that this validity can be demonstrated or arrived at in the proceduralist sense. This understanding captures the idea that formal considerations lack substance in terms of their not having empirical commitments. To be sure, note that the first conception of something's being formal *also* rules out dependence on empirical results, but in addition to that, it has the further requirement that formal considerations can be demonstrated. Thus, the first conception is stricter than the second (though of course proponents of the first conception are not required to acknowledge the second as coherent or intelligible).

With this distinction between two conceptions of formality in place, we might perhaps now also construe two conceptions of the idea of convergence on the basis of formal considerations, depending on which notion of formality we invoke. The first notion leads to a procedural idea of convergence, where every agent could in principle, by revising his set of intrinsic desires on the basis of needs for improvement justifiable from the perspective of that set, converge onto the same subset of desires that all maximally coherent and unified desire sets must include. Because type-II dispositionalism is proceduralistic, only this type of convergence is available to it.

By contrast, we might wonder whether a type-III dispositionalist might not favor a notion of convergence based on the second concept of formality. I will return to the merits of this idea in the next chapter. For now, however, let me remark that without such qualification, the prototypical idea of convergence, to me, already suggests a kind of proceduralism. Convergence implies that all agents change from widely different states towards states that are more similar, with all changes pointing in the direction of a

common focal point. Convergence upon formal considerations, therefore, suggest that such considerations could in principle make all agents change their initially different states. But that is precisely what the nonproceduralist denies. In other words, nonproceduralist convergence would mean that there are ways that all agents ought to have changed their states in the light of those states that were epistemically lucky. Which is not the picture I initially have in mind when I think of convergence.

I think this is another reason why I took Smith to be defending a proceduralist, type-II account when I originally read *The Moral Problem*, and I suspect many others with me. For example, in his critique of Smith's view, Sobel introduces the account as follows:

> Michael Smith ... claims that (1) "convergence in the hypothetical desires of fully rational creatures is required for the truth of normative reason claims" and (2) we have reason to "have some confidence [... that] there will be a convergence in our desires under conditions of full rationality" ... The plausibility of the claim that the desires of all agents will converge after proper deliberation hinges crucially on how one characterizes such deliberation. One could simply claim that a person only counts as having deliberated properly if she reaches certain approved conclusions. This path would assure Smith's first thesis at the cost of invoking a substantive, nonproceduralist conception of proper deliberation. The interest in Smith's claim stems from his willingness to invoke an understanding of proper deliberation which is not conceptually tied to the deliberator arriving at any particular motivations. (1999, p. 136)

Should Smith favor a nonproceduralist, type-III dispositionalist account instead, then he might perhaps object that nonproceduralism does not commit him to substantial *particular* approved motivations that are conceptually prior to the convergence. However, he would be committed to *general* principles of motivation-revision that must be established prior to the convergence without being justifiable from the perspective of the initial states of all agents. And that, at least, does not seem to be the view that Sobel has, in the above passage, set out to engage with.

Let us now focus on the type-II account, however, and the procedural idea of convergence on the basis of formal considerations. The problem for this approach, as I already indicated in section 3.2.1, is how to pull

substance out of form, especially in view of the seemingly unlimited variety in the conceptually possible intrinsic desire sets that agents might start their deliberations with. To get a bit of grip on the problem, we may understand it as involving the following three aspects: (1) the diversity of the starting points that agents must converge from, (2) the difficulties in arriving at any formal considerations beyond logical consistency at all, and (3) the difficulty in showing that any such considerations would yield convergence upon ethically interesting conclusions.

The first aspect greatly diminishes the relevance of our experiences with interpersonal convergence in real life, because such convergence is almost always explained against a common background that includes shared attitudes. Thus, Sobel points out that Smith's appeal to thick ethical concepts, in this context, is not helpful: thick concepts indicate the absence of certain diversity in attitudes, not the power of rationality to overcome such diversity (Smith, 1994, p. 188; Sobel, 1999, p. 146). As we have already seen, thick concepts might drive convergence if an anti-Humean theory of motivation could be made plausible, but motivational Humeans must always allow the conceptual possibility of rivalrous systems of thick concepts that correspond to the sheer logical possibilities for intrinsic desire variation.

Smith has also pointed to the massive *disagreements* among human beings in the past that seem to have been more or less settled in the present. Here, at least, the process of reasoning may have had a role. However, it may still be in the light of certain attitudes that already were shared amongst us that the desire sets of some parties in such disagreements turned out to be inconsistent. As I have indicated in section 5.2 above, the challenge for nonrelationalist dispositionalism lies especially in the realm of counterfactual, but conceptually possible, agents such as the sardonic aliens from Mars who lack any intrinsic desire towards our well-being whatsoever. In the words of David Enoch:

> There are, it seems to me, infinitely many coherent sets of beliefs and desires. Perhaps Smith's own beliefs and desires comprise one such set. Perhaps the null set (or perhaps a set with many beliefs but no desires) is another, for where is the incoherence there? And there seems nothing in the very idea of coherence to exclude infinitely many other coherent sets of beliefs and desires. How is it, then, that all rational agents converge on the same set? Isn't this an amazing miracle? Surely,

> it cries out for explanation. And absent such an explanation it will be too much to believe. (2007, p. 106)

Perhaps Smith need not even deny that infinitely many coherent desire sets are possible. Depending on how we construe the notion of a desire set, he may not need to claim that all agents would convergence on the *same* set, but merely that there is some non-empty set of intrinsic desires D such that all agents would converge on a *coherent superset* of D. Then all desires in D would give rise to normative reasons for action, and no other desires in the superset of any agent would be in conflict with them. Furthermore, all desires that express particular tastes of some agent may also give rise to normative reasons for action, as long as they are 'sanctioned' by general, non-indexical desires in D that make those tastes part of the circumstances of action in the manner which I discussed in the section on weak dependence above. Finally, there is nothing in Smith's view that prevents him from allowing a third class of 'dangler' desires, so to speak, which are not in conflict with anything but are too arbitrary to be shared by all fully rational agents.

The hard part, of course, is to show why all coherent desire sets would have to be supersets of D. This brings us to the second aspect of the problem, as by now it seems obvious that if we merely restrict ourselves to the class of all *logically consistent* desire sets, the intersection of all those sets would be empty, if only because this class would seem to include the empty set, as well as desire singleton sets. However, an agent with only a single desire, let alone with no desires at all, might not be conceivable, due to the holism of the mental for example, and various other reasons that I will discuss below when we turn to the constitution strategy. Nevertheless, conceivable agents, it seems, might in principle be fully consistent in their desire sets and still be in conflict with each other interpersonally, which is why Smith has employed the notion of a "maximally coherent and unified" desire set. We have already seen that the idea of maximal coherence and unification of desires may be explained by analogy to the epistemic virtue of parsimony for belief sets, and perhaps this would take care of the second aspect of the problem, i.e. why there would be formal considerations in favor of more unified desire sets. Nevertheless, we have also seen that this analogy cannot explain, but rather presupposes the really hard part of the view to defend: that intrinsic desire sets converge on a single (sub)set when made to conform to the constraints of maximal unity and coherence.

Perhaps it is actually trivial that D will be non-empty. Perhaps, for example, it could be demonstrated that D includes the desire to be coherent and unified in one's desires. However, this offers little help. First of all, this seems only trivial when we construe these desires in indexical form: I would desire that my desire set be coherent. However, as I have already argued, nonrelationalism requires similarity of desires in extensional form. But even if Bob would desire that his desires were unified, why should John also desire that Bob's desires were unified? Furthermore, and this brings us to the third aspect: if we want to make it plausible that there are normative answers to ethical questions, then the type-II dispositionalist must show that D includes desires that settle such questions. And that is clearly not a trivial matter.

For example, consider the following formulation of Smith's principle of parsimony for desires:

> Reason requires that ... (If someone has an intrinsic desire that p, and an intrinsic desire that q, and an intrinsic desire that r, and if the objects of the desires that p and q and r cannot be distinguished from each other and from the object of the desire that s without making an arbitrary distinction, then she has an intrinsic desire that s). (2007, p. 138)

He points out that this is the sort of principle that deals with "Future Tuesday Indifference," Parfit's example of an agent who cares about his future experiences except for those he will have on future Tuesdays (Parfit, 1984, pp. 124–125). The strength of the example is that we immediately feel that such a preference scheme would be unreasonable, even though we do not immediately see why it should be inconsistent.[9] Nevertheless, the strength of the example in showing how coherence might exceed consistency, is at the same time its weakness in making convergence towards morally relevant desires plausible. Not only because actual moral disagreements or dilemmas never feature such proposals. But also, and more importantly, because we have no trouble constructing thought experiments, such as my example of the Martians, in which the villains are not susceptible to these types of unreasonableness. My Martian invaders might be oddballs,

[9]But see Street (2009) for a defense of the idea that an ideally coherent agent *can* be Future Tuesday Indifferent. For the record I suppose I am undecided about this matter. But for the sake of the argument, my purpose here is to argue that if Future Tuesday Indifference cannot be coherent, that does not give us enough to establish nonrelationalism.

but they're not stupid. If morality really does derive from rationality, it is not because all sinners are fools. So if we keep the meaning of the term "arbitrary" in the principle so extreme as to only cover cases as absurd as Future Tuesday Indifference, then the principle does not offer interesting guidance. But if we loosen the meaning of "arbitrary," then we may feel that the support derived from the outlying examples is no longer sufficient to justify the principle.

The same problem threatens other principles, even those which may seem more ethically ambitious. An example is the idea that moral judgments should be *universalizable*. As John Mackie has argued, this notion is deeply ambiguous (1977, ch. 4, pp. 83–102). He therefore distinguished between different "stages" of universalization, where each stage was more restrictive compared to the previous one. But whereas the early stages could be considered candidates for conceptual requirements, they turned out far too weak to be ethically significant, whereas the later stages were ethically substantial, but impossible to derive from conceptual requirements.

Recently, Smith has attempted a more ambitious line of argument, based upon the idea that there is an ethically relevant symmetry between, on the one hand, the fact that an agent must trust his past self, and that his future self must trust him, and on the other hand, the fact that different agents must trust each other. In particular, the idea is that as a matter of conceptual necessity, I must trust my past self not to have intended to deceive me, or I could not any longer trust the attitudes that I have inherited, so to speak, from my past self. Smith has tried to make the idea plausible that it would be more unified for my desire set to therefore feature equal interest in not deceiving my future self and not deceiving others.

Now I am just sketching an outline of the argument here, and Smith's detailed version of it has yet to appear in print. Nevertheless, I do not see how the requirement for one's desire set to be symmetric in this sense could be justified from the perspective of someone whose desire set does not, initially, contain any attitudes geared towards such symmetry. Suppose that there is one Martian amongst us, appearing like a human, with many people caring for him and making sure not to deceive him. Nevertheless, he deceives them all the time, constraining his actions only in ways that prevent others from discovering his intent. If there are no attitudes in his desire set that are geared towards being symmetrical in Smith's sense, then

how could we possibly demonstrate to him the errors of being egoistic in his sense? At least in the proceduralistically formal sense, this does not seem to work as a formal requirement.[10]

5.4.2 Problems for the Constitution Strategy

The second strategy available to the type-II dispositionalist is to argue that the conceptual possibilities for variation in attitude sets are not as wide as they might initially seem to be, because they are constrained by certain requirements that are constitutive of being an agent in the first place. The purpose of the strategy is to derive a set of attitudes from these requirements that all agents must have.

In section 3.2.2 I raised the problem for this strategy that even if such common attitudes can be derived, they turn out to be common in the wrong way. They would be *isomorphic* between agents, whereas nonrelationalism requires attitudes that are *similar* between them (a distinction I have introduced in section 1.3.2). The example of the desire for one's desires to be unified, which I discussed in the previous section, illustrates this: if it should be the case in the nonrelationalist sense that Bob's desires are unified, then we must show that John would also desire Bob's desires to be unified, not that John would desire that John's desires be unified.

I have not much more to add to my earlier argument, except this: even if there might nonetheless be ways to arrive at similar attitudes across all agents through some constitutive requirement, then they would have to satisfy the further demands of being proceduralistically formal and consistent with the Distinctness Principle, in order for them to help the type-II dispositionalist. But this reduces their likelihood even further. A motivational anti-Humean might, for example, argue that having certain beliefs about the world is constitutive of being a believer at all, and that these include motivationally efficient beliefs that yield similar motivations across all agents. This is no help to the type-II dispositionalist, who must demonstrate similarity of intrinsic desires.

Perhaps nonproceduralists could propose a 'rich' concept of agency that would itself already harbor somewhat substantive intuitions about what it means to really be an agent, or a self-governing agent, for example, from which some nonrelationalist implications could then be derived. Again, the

[10] Smith, in personal communication, seems to agree with this, opting for the nonproceduralist, type-III dispositionalist interpretation of this argument instead.

type-II dispositionalist cannot make us of this approach, unless he could show from the perspective of those who profess not to share the respective attitudes that it is impossible for them not to have them.

But Enoch (2006) has employed a strong argument to show that the latter cannot be done, since any concept of agency that is substantive enough to imply similar attitudes shared by all agents will also be so substantive as to invite the response from some that they simply are not, and do not care to be, agents in that sense. Instead, the are happy being "schmagents," thereby re-introducing a less substantive conception of agency under a new name.

Enoch's own view is that all dispositional accounts of practical normativity must be rejected, and there of course I disagree with him. Furthermore, I think that the schmagency-argument does not establish the uselessness of a dispositional approach based on ideas about what is constitutive of agency: in particular, I have already relied on various ideas about what is constitutive of self-governing agency to build the type-I account that I favour, and I shall be relying on such ideas even more in the chapters to come. But at least with respect to the type-II account, the schmagency-argument helps to show, in combination with my own argument based on the similarity–isomorphism distinction, that we cannot put nonrelationalist content in the subjective motivational sets of all conceptually possible (schm)agents.

5.4.3 Problems for the Combined Strategy

At the end of section 3.2.2 I suggested that the convergence and constitution strategies might be combined in a way that would reduce each of their problems somewhat. We have seen that it is a problem for the convergence strategy that the class of logically possible intrinsic desire sets includes such immense variety that convergence to substantial results seems hopeless. However, if the constitution strategy can narrow that class down a bit, from the class of logically possible desire sets to the class of desire sets that only conceptually possible agents might have, then perhaps there will be more for the convergence strategy to latch on to. Enoch has made a similar suggestion:

> There is another possible explanation of the miracle of convergence that is, it seems, available to Smith, an explanation Smith himself hints at (2004a, pp. 27, 205). According to this

> explanation, convergence of all possible rational agents emerges as the result of features that are necessarily shared by rational agents, as the upshot, that is, of features that are constitutive of rational agency. If the details of such an explanation could be filled in, convergence would, of course, no longer be at all surprising. (Enoch, 2007, p. 107)

Conversely, we have seen that the constitution strategy fails to accomplish the jump from isomorphic attitudes to similar attitudes, but perhaps it can outsource that task to the convergence strategy. In other words, the combined strategy is a two stage process: in the first stage, the type-II dispositionalist attempts to squeeze as substantial isomorphic intrinsic desire structures out of the conditions constitutive of agency as he can, and then in the second stage, he must try to show why the process of rational deliberation, when applied to these structures, would make all agents converge on a set of similar attitudes.

So what sort of isomorphic desires can we derive from the concept of agency? In fact, we need not even look at the general concept of agency, but we can move a bit further by starting out with the concept of *self-governing* agency. I have hinted at this idea before; let me now explain why. Recall that I have argued for the notion of self-government as an element of the ideal conditions of agency that should be distinguished from the requirement of rationality, on the grounds that certain motivational deficiencies are intuitively compatible with the agent being fully capable of thinking rationally. The problem of such agents is not that they fail to deliberate properly, the problem is that they fail to control themselves.

Since ideal agents do not have this problem, whereas actual agents do, the convergence process on the basis of rational deliberation need not depart from our actual attitude sets, but rather from what our attitude sets would be if we were fully self-governing. Thus, it is from the class of intrinsic desire sets that conceptually possible self-governing agents might have, that the convergence must be achieved. A similar strategy has been proposed by Christine Korsgaard, who argues that Kant's transcendental analysis of agency as *self-legislation* yields the interest in acting in a law-like manner (1996, p. 97–99), from which Kantian ethics may then attempt to deliberate rationally towards nonrelationalist altruistic conclusions that all agents would be rationally required to accept (e.g. in Korsgaard, 2006).[11]

[11] Although her strategy for establishing her brand of moral rationalism bears this similarity to the combined strategy in support of type-II dispositionalism, it is not clear to me whether

There may well be all sorts of interesting intrinsic desires that ideal agents would have as a result of such deliberation upon what is constitutive of their self-governance, but there are two reasons why I think they still cannot give rise to nonrelationalist realism. The first, and most important for our current discussion, is that the gap between isomorphic attitudes of self-governing agents on the one hand, and similar attitudes that settle interesting moral questions on the other, will still be much too large for the formalized convergence strategy to breach.

It might be that every self-governor will turn out to have a normative reason to be critical of his past assumptions, for example. Or to strike a wise balance between experimenting with new policies for action on the one hand and sticking with old policies that have proven worthwhile in the past. But it is hard to see how we go from these kind of interests to a view about whether we should care for the fate of future generations, say, or about whether we should refrain from harming animals that have the capacity to suffer, if that transition must be made on the basis of formal principles of reasoning alone. Remember that even the Martian invaders, when made to consider their self-directed interest in being the best self-governors they can be, would have to be forced by correct reasoning from their own perspective, to conclude that they had better let us live in peace.

So what about Kantian ethics, then? Briefly, I think many problems that have often been raised for Kantian ethics are ultimately related to the same issues I have been summarizing here, though the various distinctions and disambiguations that I would need at this point in my framework in order to do justice to Kantian theory are beyond the scope of this thesis. Nevertheless, the bottom line, I suspect, for a marriage between Humean motivation and Kantian deliberation would be the aforementioned worry about universalization. Many philosophers doubt that the categorical imperative (which, if ascribed to all agents, would yield interpersonal similarity) could be derived from the mere idea of subscribing to action that is regular in a law-like manner (which by itself merely involves interpersonal isomorphism), and furthermore, that the different formulations of the categorical imperative could even be derived from each other. The

I should classify Korsgaard as a type-II dispositionalist herself. In her own terminology, Korsgaard rejects moral realism in favour of constructivism, but I do not know her work well enough to determine whether that rules out practically normative facts on the extremely minimal notion of facts that I have employed in order to state the Facts Principle. Nor is it clear to me, despite her being a card-carrying Kantian, whether her account would really violate my Distinctness Principle.

reason to be skeptical about this, in my view, is closely related to Mackie's dissection of lesser and more substantial stages of universalization.

Furthermore, it is also often doubted that even the categorical imperative yields enough moral substance to guide action, and it is controversial whether Kant intended it to do so. The categorical imperative rules certain maxims of action *out*, but it is not clear whether it rules anything *in* such that convergence would ensue. Finally, we should question the idea that self-governance requires one's rules of conduct to be law-like in the manner that Korsgaard implies. Self-legislation is just another word for self-governance, but not in a sense from which an austere legalism about the language of practical reasoning can be derived. According to Williams, this is a core problem for the Kantian project:

> Why should one adopt such a picture? Why should I think of myself as a legislator and—since there is no distinction—at the same time a citizen of a republic governed by these notional laws? This remains a daunting problem, even if one is already within ethical life and is considering how to think about it. But it is a still more daunting problem when this view of things is being demanded of any rational agent. (1985, p. 63)

But perhaps it is no surprise that we cannot derive Kantian ethics in this manner, preventing convergence from our proceduralist perspective, as many Kantians take the "Factum der Vernunft" (the moral law as known to us) to be an example of *synthetic a priori* knowledge. In their view, we just find this knowledge within ourselves, but we cannot demonstrate its truth analytically to the theoretical skeptic who professes to know no such thing. The question whether nonproceduralist epistemology and the epistemology of the synthetic *a priori* amount to the same thing is a complex one, however, as the analytic–synthetic distinction is by no means univocal any more in the contemporary literature. I will remain neutral about this matter, and stick to my own terminology of proceduralism versus nonproceduralism.

So to summarize our first reason for doubting the combined strategy, what all these considerations show is that, even though the constitutionalist stage of the combined strategy may seem to address the first two aspects of the problem for the convergence strategy—the wide range of priors and the establishment of conceptual requirements beyond logical consistency—it does not offer much solace in view of the third aspect—the problem of

arriving at morally significant conclusions. And so our worries about the convergence strategy still motivate anti-formal content skepticism: no formal derivation of substantial moral conclusions from premises whose conceptual necessity can be demonstrated seems possible.

The second reason is that I am not even sure that constitutive requirements of self-governing agency ever provide us with isomorphic normative reasons for concrete actions or choices. Consider the example of the interest in being critical about one's own views. Too little of this motivation, and one loses important opportunities for disconfirming one's false beliefs. But too much of this motivation, and one would spend too much time and energy on it. However, in particular cases, the optimal balance between these two extremes may not depend on empirical facts about the case alone. They may also depend on one's contingent intrinsic desires in ways that, once again, do not converge across all agents. I shall return to this matter later on in the thesis.[12]

For now, however, the first of these two reasons is the most important one. Despite all the intricacies that we have discussed in this chapter—the weak dependence construction to accommodate different tastes, the adviser approach to incorporate awareness of our imperfections into deliberation, the belief-desire analogy in order to establish a parsimony principle for practical reason, the two-stage combination of the constitutive and the convergence strategy, and the license to deliberate on what is constitutive of self-government—there *still* does not seem to be any rational argument that could in principle convince my Martians to stop torturing human beings for fun, given that they lack the intrinsic desire to promote or respect our well-being.

One way for Smith to accommodate this result is to adopt the type-III account instead. On that view, normative reasons for action are still determined by purely rational considerations alone, but it is no longer seen as a feature of rational considerations that they can be justified internally to all conceptually possible agents. However, as I will explain in the next chapter, this approach faces difficult problems as well.

[12]See section 10.3.4.

6 *Dispositionalism Without Proceduralism?*

In the previous chapter I have distinguished two interpretations of Michael Smith's nonrelationalist alternative to the type-I solution from chapter 4. The type-II approach accepts proceduralism, but we have seen that this is very hard to combine with nonrelationalism. Instead, type-III dispositionalism rejects proceduralism. On this view, some desire sets are irrational even though the conceptual necessity of the principles in virtue of which such desire sets ought to be revised cannot be demonstrated from the perspective of those desire sets themselves.

The epistemological upshot of type-III dispositionalism, as we have seen at the end of section 5.3.3, is that for every agent, there are intrinsic desires that the agent is either epistemically lucky or unlucky to have, in the ineliminable armchair sense. This suggests two ways in which the view might be criticized by proceduralists. The first would be to argue in general against the very idea of ineliminable armchair luck. The second is to argue more specifically against the idea that intrinsic desires are susceptible to such luck.

Although I am sympathetic to the first line of argument, discussing it properly would require me to address some of the most fundamental questions in epistemology, which would take us far beyond the scope of this thesis. Instead, I will restrict myself to a brief discussion in section 6.1 of what it *would* mean for a dispute in theoretical philosophy to involve ineliminable armchair luck. The defense of type-III dispositionalism relies on the idea that such a theoretical dispute will be analogous to certain disagreements in normative ethics in relevant respects.

In section 6.2, I will address the question whether type-III dispositionalism is still an Internal Reasons View, which will help us to get a better sense of how this view relates, exactly, to the various philosophical positions that we have been considering so far. In section 6.3 I articulate

various worries that I have concerning the aforementioned analogy, which make the solution unconvincing for me even if the notion of ineliminable armchair luck would be coherent.

6.1 Armchair Luck in Theoretical Philosophy

Smith has suggested that the dispute between nominalists and realists on the existence of universals might be a matter of ineliminable epistemic luck.[1] He imagines that their debate might reach a point where both sides have succeeded in making their views internally coherent and completely immune from objections. If that would happen, he insists, we would have to conclude that one side must have reasoned from ineliminably unlucky priors.

Note how this differs from verbal disagreement. Suppose that A claims that determinism rules out free will, while B holds that they are compatible. However, as it turns out, A holds that free will is the ability to do otherwise, whereas B holds that free will is that which makes us responsible for our actions. If B agrees that determinism rules out the ability that A has in mind, while A agrees that this ability is not required for the responsibility that B has in mind, then their disagreement may be resolved using what David Chalmers (2011) calls the "method of elimination": A substitutes "free will" with "free will$_A$" while B substitutes it with "free will$_B$."

Metaphysics seems to resist this method. Chalmers discusses the dispute over the existence of mereological wholes (2009). If there are two cups on my table, does that mean there are two or three objects on the table? If one applies the method of elimination, one might say that there are two objects$_A$ and three objects$_B$ on the table, such that objects$_B$ exist$_B$ while they do not exist$_A$, but those who subscribe to *ontological realism* will still disagree amongst each other about the question whether existence$_B$ is *real existence*.

The same applies to the existence of universals. However, if both sides on the mereology dispute would be capable of formulating internally fully coherent theories, then it seems we could simply choose how to use the 'real' existential quantifier after we've agreed about the relevant empirical facts, without this choice affecting much of importance, except that in certain contexts it may seem more practical to talk about the mereological wholes as existing while in other cases it seems the other

[1] From personal communication.

way around. This leads Chalmers to defend *ontological anti-realism* with respect to mereological existence: the view that there are no ontological facts about this matter that make the claim that mereological wholes exist true or false. Which rules out ineliminable armchair luck.

Perhaps it is less intuitive that the existence of universals might be indeterminate in this sense, because nominalism and realism about universals seem less philosophically isolated, so to speak: they are intertwined with many other theoretical issues. However, this also makes it less plausible to expect that fully coherent formulations of both theories will be possible: the more dependencies they have elsewhere in philosophy, the more justified we can be in hoping that a requirement of coherence in another discipline might ultimately settle the universals debate as well.

Furthermore, note that such resolution does not presuppose that all philosophy must be *analytic*. Many philosophers have argued that our knowledge has certain assumptions that we are *a priori* justified in making, because their negations, even though not formally contradictory, would nevertheless be practically or performatively self-defeating for us to endorse. If realism about universals could be shown to rely upon such an assumption while nominalism would violate it, then the dispute could be settled.

Hence, to claim that the universals dispute involves ineliminable armchair luck is to claim that it is on the one hand not so isolated from other issues in philosophy as to make ontological anti-realism about the issue itself plausible, while also claiming on the other hand that these philosophical interrelations cannot ultimately be exploited to resolve the debate on the basis of assumptions that no one could reasonably deny.[2]

[2]I am inclined to think that these requirements cannot be combined, because I favor a pragmatist philosophy of language according to which disputes can only be understood as meaningful in the context of the sort of inquiry that would in principle allow us to resolve them, if not in actual practice. The defense of such a general account of meaningful disputes in philosophy is surely beyond the scope of this thesis, though nevertheless closely related to my main line of argument here: that we should think of understanding disconfirmation as the primary issue, and of understanding facts as derived from it. The pragmatist approach that I have in mind would generalize this idea from meta-ethics to questions about meaning, knowledge, and existence in theoretical philosophy.

6.2 Is This Still an Internal Reasons View?

In section 2.4, we have seen that Williams's defense of the Internal Reasons View required a proceduralism concerning our knowledge of reasons for action. This made it possible to formulate a nonproceduralist objection against the defense, based on the idea that external reasons might exist in the light of facts that would remain epistemically inaccessible to us should we have started out with certain incorrigibly misguided attitudes in our subjective motivational set. Therefore, if Smith is also rejecting proceduralism, and if his argument also turns on the idea that the facts about their normative reasons may be inaccessible to some agents in this sense, doesn't that then make him an external reasons theorist? Isn't that simply what external reasons are?

The answer is that it depends on how nonproceduralists understand the notion of correct deliberation in terms of which internal reasons are defined. If the nonproceduralist agrees with the proceduralist that correct deliberation is a procedural notion, then the nonproceduralistically construed reasons for action may lie outside the reach of deliberation, depending on the prior attitudes of the agent, which would make them external reasons. However, if the nonproceduralist maintains that correct deliberation may itself be nonprocedural, in the sense that it will be impossible for certain agents to revise their deliberative practices from their own perspective so as to arrive at methods of deliberation that are correct, then the nonproceduralist can still defend the claim that all reasons for action are within the reach of a sound deliberative route, even if they may not lie within the reach of procedural inquiry for some agents. If Smith means to reject proceduralism, then this would seem to be his view.

But does it make sense, we may now ask, to understand the notion of a deliberative route in this sense? Doesn't the very idea of a deliberative route imply a proceduralist understanding? What would be the difference between an internal reasons view construed in terms of this nonproceduralist route, and the nonproceduralist defense of the external reasons view sketched above? I think the reason why Smith thinks of himself as an internal reasons theorist is that he really wants to stress the *transitional* function of the principles of reason that he believes in. Consider again his parsimony principle which we discussed in section 5.4.1:

> Reason requires that . . . (If someone has an intrinsic desire that p, and an intrinsic desire that q, and an intrinsic desire that r,

> and if the objects of the desires that *p* and *q* and *r* cannot be distinguished from each other and from the object of the desire that *s* without making an arbitrary distinction, then she has an intrinsic desire that *s*).

I have argued that the problem for this principle lies in the notion of an "arbitrary distinction" which may receive more or less substantial interpretations. Whereas the less substantial interpretations may be justified to all agents from their own perspective, the more substantial interpretations may not, while the latter would be needed to arrive at ethically interesting conclusions. In order to circumvent this problem, Smith may defend a nonproceduralist type-III account according to which this principle, under a fairly substantial interpretation, will still be *a priori* correct, but nevertheless epistemically inaccessible to some agents from their own perspective. Note, however, that despite being nonproceduralistically substantial in this manner, the principle would nevertheless still be *transitional* in that it requires a prior set of attitudes to operate on, yielding a revised set of attitudes as a result. And the same would apply to a substantial version of a universalization principle, for example. It is in this sense that type-III dispositionalism may still be understood as a 'route-like' account.

In fact, we may think of this account as the diametric opposite of the sort of external reasons view that one might wish to defend in the light of the 'non-route-like' deliberation objection. As we have seen in section 2.5, this objection was based on the idea that agents can acquire new attitudes in a rational manner that does not yield these attitudes on the basis of some deliberative continuation upon the previous attitudes of the agent. Hence, the 'non-route-like' External Reasons View is compatible with proceduralism, while rejecting the idea that principles of deliberation must be 'route-like.' By contrast, type-III dispositionalism rejects proceduralism, while being compatible with the idea that principles of deliberation are 'route-like.'

Nevertheless, the distinctions involved in carving out the type-III view as an Internal Reasons View are somewhat vague. Consider a deontological prohibition such as "Thou shalt not lie." Is this a concrete, ethical imperative? Or it is a transitional principle that only yields substance in its revisionary operation, eliminating attitudes in support of lying from the subjective motivational set in order to produce a motivational configuration that will not lead the agent to tell lies? Both answers seem somewhat arbitrary.

Furthermore, the distinction between nonproceduralists who take deliberation to be procedural and those who do not is also somewhat suspicious. Are these really different logical options, or is this another example of a verbal disagreement? Using Chalmers's method of elimination, type-III dispositionalists might distinguish between deliberation$_P$ (for proceduralistically reachable deliberation) and deliberation$_{NP}$ (for deliberation on the basis of principles that may not be justified proceduralistically to all agents), on the grounds that both conceptions capture some of the intuitions about deliberative correctness that nonproceduralists may want to take into account, while each notion can be employed to play a different conceptual role in their theory. In that case, the distinction between the Internal and the External Reasons View turns out to be ambiguous: in relation to deliberation$_P$ the type-III dispositionalist must be an External Reasons Theorist, but in relation to deliberation$_{NP}$ he can be an Internal Reasons Theorist.

6.3 Problems for Type-III Dispositionalism

Is type-III dispositionalism a convincing alternative to the type-II account we discussed in the previous chapter? We have just seen that the type-III dispositionalist may still want to insist that normative reasons for action are determined by principles of reason that are transitional in nature, giving rise to a 'route-like' conception of correct deliberation. Note that on the basis of this assumption, the type-III dispositionalist would still need a convergence, constitution, or combined strategy in order to establish nonrelationalism. This means that the problems that the type-II dispositionalist is facing will not automatically melt away before the type-III dispositionalist. However, the latter may seem better suited to handle them, because he may employ more substantial interpretations of such principles of reason, as I have explained in the section above.

But in return for this theoretical advantage over the type-II account, the type-III dispositionalist must face a number of additional problems that proceduralistic dispositionalists need not worry about. Furthermore, upon closer examination the advantage over type-II dispositionalism may turn out to be a bit of a poisoned chalice, as it immediately gives rise to a new dilemma that restores the predicament of the type-II view. I will explain this latter problem first, and discuss the other problems afterwards.

6.3.1 A Poisoned Chalice: The Economy Problem

Let us once more suppose, for the sake of our discussion, that there is such a thing as ineliminable armchair luck in theoretical philosophy, and that the nominalism–realism dispute over universals turns out to be a matter of such luck. Suppose furthermore that universals exist: the realists have been lucky in their prior beliefs; the nominalists unlucky. Because it is armchair luck we are talking about, it is important that we do not think of realism as being made true by some brute metaphysical fact that, despite its being inaccessible through empirical investigation, could have been different. The realist in our example is claiming that it is conceptually necessary that universals exist—not even God could have made a world without them. On a conceptual level, the story of the nominalist just does not make sense, even though it cannot be proven why.

Belief in such a scenario would have to be justified on the basis of the idea that for various philosophical reasons, it would not make sense if both stories made sense, nor if neither made sense, and so it follows that one and only one of them must make sense, even if we cannot refute either. However, we should be wary of the philosophical temptation to invoke this strategy whenever we want to postulate conceptual truths that we do not know how to demonstrate logically. If the dispute over universals involves ineliminable armchair luck, then we must accept that there is a certain 'gap' between that which makes sense conceptually and that which cannot be disconfirmed internally, but even then we should consider it a theoretical virtue to be able to keep this gap as small as possible. Accepting something as a matter of conceptual truth without being able to explain to one's critic *why* it is conceptually necessary should not be our favorite dialectical move in the game of philosophy. Note, once more, that we are not talking about *actually* being able to convince one's opponent, at a conference, say, or in the journals. That only shows that there is *eliminable* epistemic luck, which all parties may, for our present purposes, accept. The point is that one would be claiming a certain truth to be conceptually necessary while being unable, in principle, to attribute the belief of one's opponent in the opposite claim to any imperfection in their thinking. It involves biting the bullet that one's theory is based on a certain dogmatism, which I think philosophers should want to avoid whenever possible.

In other words, to suppose that one knows that one's opponent is wrong as a matter of ineliminable armchair luck is *theoretically expensive*. Allowing certain disputes in philosophy to be lucky in this manner is a very big deal,

the sort of result comparable in its significance to Gödel's incompleteness theorem in mathematics, say, except that no philosopher has ever made a proof of it. Furthermore, note that even though Peano arithmetic is incomplete, completeness in logic has not become an outdated commodity. Complete formalisms are valued for their theoretical neatness, so to speak, and if a complete formalism can do a certain job, then this is still preferable. In general, conservative formalisms are always preferable to more exotic systems of logic when they are equally useful, and when they are not, then there is always a trade-off between the desirable properties of conservative formalisms versus the added expressiveness of the more exotic frameworks. In essence, this is just another application of a principle of parsimony in mathematics, but now applied at the meta-theoretical level. A similar principle, it seems to me, should be applied at the meta-theoretical level in philosophy: we should try to be *economical* about the *a priori*.

Hence, it might be that a conceptually indemonstrable truth about the existence of universals is still acceptable, while believing in such a truth about the existence of mereological wholes goes too far. But in the light of this, it now seems quite a leap from being a nonproceduralist about the universals debate in order to keep it *a priori* to being a nonproceduralist about *morality* in order to keep it *a priori*. This means that the third aspect of the problem for the convergence strategy—the problem of arriving at ethically significant substance—is coming back in the form of a new dilemma. Because the more substantial we make our interpretation of transitional principles of reason, the less economical it would be to include them in the realm of ineliminable armchair luck. But the less substantial we make them, the harder it becomes to establish convergence upon ethically interesting results.

And here the analogy to the nominalism–realism debate offers little solace. For suppose that the Martians subscribe to a weak interpretation of the universalization principle that allows them to torture humans, while their victims subscribe to a stronger interpretation that would allow them to argue that the Martians should not torture them. Intuitively, does it really seem on a par with the idea that there must be an *a priori* truth about the existence of universals, that there must also be an *a priori* truth about which interpretation Martians and humans should give to the universalization principle, if neither truth can be arrived at through internally justifiable methods of inquiry? Without further argument, I do not see why I should accept that the moral prohibition of torture is just as conceptually necessary

as the existence of universals.

At this point, the type-III dispositionalist might want to insist that he does see these issues as properly analogous, and wonder why the burden of proof would be on him. However, the reason why he does have the burden of proof is because of the economy problem: other things being equal, the position that implies the least amount of ineliminable luck in philosophy is preferable. Against this insight, it is hardly a convincing move to simply declare that the price paid for making altruism nonproceduralistically *a priori* feels just as cheap as the one paid for the existence of universals. On the contrary, that would just seem like an *ad hoc* assumption of the very thing that the type-III dispositionalist has to make plausible.

6.3.2 Epistemic Luck and Direction of Fit

In the case of a motivationally Humean theory such as type-III dispositionalism, there is a further dis-analogy between the dispute over universals in theoretical philosophy and the normative ethical dispute between Martians and humans concerning the latter's extermination. According to the dispositional solution, we explain the truth conditions of our practical beliefs in terms of our intrinsic desires under ideal conditions of rational, self-governing agency. In the case of the type-II account, the idea was that just like there is a logic for beliefs, so there is also a logic of desires, and if we apply this logic to the intrinsic desires that we start out with, then we will ultimately converge upon the intrinsic desires that constitute our normative reasons for action. Now the type-III account is supposed to be similar, except that it makes the principles of rational desiring more substantial, such that they have at least to some extent already to be built into the prior desire set, with the implication that other prior desire sets will not converge proceduralistically upon the desires that rational agents must have.

Hence, if the desires of the Martians will not converge proceduralistically upon the desires of their ideal selves (but rather upon the desires of the unlucky twins of their ideal selves), their intrinsic desires are 'getting it wrong' in a manner that cannot be, in principle, explained to them from the perspective of those desires. But this feature, it seems to me, is hard to combine with the very idea of an *intrinsic* desire. For what is this proceduralistically unreachable 'it' that the desires are getting wrong, if not something extrinsic to them? And if there is something external to

them that these desires have to get right, then it would seem their *direction of fit* (Smith, 1994, p. 111) is no longer exclusively that of 'fitting the world to the mind.' Instead, nonproceduralist dispositionalism seems to imply that desires must have both directions of fit: even though they aim, motivationally, at making the world fit their content, they must also be understood as aiming, epistemologically, at having their content fit the substantial *a priori* principles of desirability.

In response to this, the type-III dispositionalist might wish to object that the formulation of the Humean theory of motivation in terms of directions of fit is to be understood only as a motivational, and therefore explanatory, rather than a normative, justificatory thesis. Thus, beliefs differ from intrinsic desires in their having a 'fit the mind to the world' direction of fit *in the sense that,* when actually encountered with evidence indicating a mismatch between the world and his belief, the agent will be moved to alter his belief, a move which is explained by the insight that beliefs have this direction of fit. Clearly, the type-III dispositionalist's commitment to nonproceduralism does not require him to say that intrinsic desires are belief-like in this sense, because insofar an agent is ineliminably unlucky in his desires he will never be moved to alter them in the above sense.

However, note that the same thing would apply to the belief in the existence of universals that cannot be justified to the defenders of nominalism from their own perspective, if the analogy is to hold, and we would still consider this belief to have the 'fit the mind to the world' direction of fit in a manner that distinguishes beliefs from intrinsic desires. Perhaps the type-III dispositionalist would be tempted, at this point, to argue that such a belief is not really about 'the world' because of its *a priori* character, but what he cannot deny is that it is a *belief*, that it has *truth conditions*, and that therefore, the belief is to be understood as an attempt to be 'fitting' rather than a goal to be 'fitted,' so to speak. The idea that the nominalist might never be moved to change his belief about universals does not make a difference in this respect. So the analogy on which the type-III dispositionalist is relying seems to require that there is something that the desires must fit in a manner analogous to the way in which there is something that the belief in the existence of universals must fit.

Furthermore, we have already seen that the explanation of action has a justificatory aspect as well: even without presupposing the normativity of normative reasons for action, explaining actions in terms of beliefs and desires does presuppose the normativity of beliefs and desires themselves:

the normativity of the sort of minimal rationality that we must ascribe to an agent in order to be able to attribute beliefs and desires to her in the first place. The different directions of fit are articulations of the different ways in which beliefs and desires have to make sense in explanations of the agent's actions, and what the type-III dispositionalist seems to have committed himself to is that something similar to the direction of fit needed to make sense of a belief in the existence of universals may also be applied to the concept of an intrinsic desire.

Is this still compatible with the Distinctness Principle? Strictly speaking, it may depend on how one wishes to construe the notion of "belief." In Michael Smith's view, beliefs aim at the truth, while desires aim at the good. However, one might argue that the very idea of aiming at something, or of having a 'mind to something' rather than an exclusively 'something to mind' direction of fit, involves the idea of truth. Furthermore, we have already seen that the very idea of being ineliminably lucky in the armchair sense involves the idea of truth, because it is a form of veritic epistemic luck. So if intrinsic desires must aim at something that lies beyond that which can be justified from their own perspective, and if an agent can therefore be epistemically lucky or unlucky in the desires that she starts out with, then we may wonder how distinct such desires still are from what we might attribute to the agent as her beliefs about what is true and false. Instead, it seems rather that this view is going in the direction of the idea of having "besires," which is something that Smith has clearly not wanted to defend. We may therefore wonder whether the motivationally Humean type-III dispositionalism is such a stable option. Can it still be distinguished from motivationally anti-Humean dispositionalist accounts, and even if it can, does it still account for the Distinctness Principle in 'spirit,' so to speak—does it harbor the intuition *behind* the Humean theory of motivation? Of course, motivational Humeans may differ a bit amongst each other about what that intuition really amounts to. This matter is beyond the scope of this thesis, but as I have briefly speculated in section 3.4.2, my own view about this intuition may ultimately boil down to the idea of the disinterestedness of truth, which does seem hard to reconcile with the idea of there being a truth about the rationality of desire that certain desires can be ineliminably unlucky about.

6.3.3 Nonproceduralism Does Not Explain Intersubjectivity

I have introduced the distinction between relationalism and nonrelationalism in the context of the Intersubjectivity Principle (section 1.3.1). One of the advantages of nonrelationalism, I claimed, was that it offers a straightforward account of intersubjectivity. However, in a sense, this only applies insofar one subscribes to proceduralism as well. If one appeals to nonrelationalism in order to explain the purpose of a moral discussion in terms of reaching a conclusion that is valid for all through reasoning and argument, then this purpose would be defeated if nonrelationalism merely holds in virtue of the postulation of truths that are beyond the reach of such methods when these truths are themselves disputed.

In a sense, what type-III dispositionalism delivers is nonrelationalistic objectivity in the *absence* of nonrelationalistic intersubjectivity. If humans and Martians cannot overcome their difference, in principle, through methods of inquiry starting from their different initial perspectives, then whatever intersubjectivity there is between them is not explained by the facts that make the views of only one of the species true. Against this, one might hold that they are at least agreeing that what they are disagreeing about is a nonrelationalist matter of fact, in the same way that the nominalist and the realist agree that they disagree about facts considering real existence, and that it is this agreement that explains the intersubjectivity.

However, I find this sort of intersubjectivity rather meager. If I were a nonproceduralist nonrelationalist about my disagreement with the Martians, for example, there would no longer be a point in my trying to convince them of their mistakes through reasoning and argument. By analogy, if realists and nominalists about universals really believed their dispute to be explained by a difference in ineliminable epistemic luck, it would no longer make sense for them to try to eliminate this difference.

Of course, they would still be actual philosophers and not their ideal selves, so their discussions might still serve the purpose for both sides to perfect their respective theories. In similar fashion, a putative discussion between the Martians and us might help us make our moral theories more coherent before they finally kill us. Furthermore, in practice the Martians do not exist, and even if we subscribe to type-III dispositionalism we may never know which actual disputes with real human beings are due to ineliminable luck. In fact, we might even believe that all human beings have started out with the lucky priors, while admitting that Martians which started out with the wrong ones are at least conceptually possi-

ble. In the end, it is really the problem about conceptual possibility that nonproceduralism needs to address.

These arguments make sense, but they carry one implication that may be unfavorable to type-III dispositionalism: namely, that the type-I dispositionalist can make use of the exact same explanations to account for the Intersubjectivity Principle. If the type-III dispositionalist can say that the practice of moral discussion still makes sense, in the absence of a proceduralistically reachable nonrelationalist truth, because it helps both sides make their respective theories more coherent, then the type-I dispositionalist can say the same thing. The only difference is that he believes that the truth cannot be reached because there is no such truth, rather than because of its unreachable nature, but this does not affect the explanation of the usefulness of the discussion. And the same applies to those discussions where in practice all parties may have sufficiently similar prior attitudes to be able to reach, in principle, conclusions that will be valid for all participants. After all, what matters for the explanation of the intersubjectivity, in this case, is that they contingently happen to have attitudes that are sufficiently similar to make commonly valid conclusions reachable through proceduralistically understandable means. Whether those attitudes are successfully aiming at further nonrelationalist facts about what is *a priori* rational to desire, as the type-III dispositionalist believes, or whether they are just the attitudes they happened to be without aiming at such nonrelationalist facts, as the type-I dispositionalist holds, once again does not affect the explanation of intersubjectivity.

In fact, these are precisely the sort of explanations that I will develop in further detail, later on in this thesis, to make my own type-I account plausible in the light of the Intersubjectivity Principle.[3] But regardless of how successful these explanations will turn out to be, what we can conclude for now is that the type-III dispositionalist will be equally successful or unsuccessful at explaining the principle as the type-I dispositionalist.

Against this conclusion, the type-III dispositionalist may want to object that his explanation does offer something extra: the common understanding between any two agents discussing reasons for action that there is some nonrelationalist fact about such reasons at stake in their discussion. Meager or not, this does seem to be a feature that type-I dispositionalism

[3]I discuss the idea that we contingently share dispositions in section 10.2 and the idea that we can improve our own theories by contrasting them with the views of others in the absence of common ground in section 10.3.1.

must lack. However, we should wonder whether this feature is really part of the explanandum that all meta-ethical accounts must share. First of all, in order to claim that this common understanding is part of the actual moral discussions that we have involves ascribing nonrelationalism to lay men as their implicit 'folk meta-ethics,' and I will argue later on that this assumption is unjustified.[4] Second of all, if the argument is not meant to be about what people actually think and mean, but rather about what they should think or mean in order to make sense of their discussion, then it seems to me that the argument simply begs the question against the very idea of relationalism. If people do not always actually believe there is something nonrelationalist at stake in their moral discussion, then the assumption that they should believe this is not something that the *relationalist* has to account for. Rather, it is precisely the thing that the *nonrelationalist* has to make plausible.

But what if type-I dispositionalism would be equally successful in explaining intersubjectivity? As long as the type-III dispositionalist is able to explain it too, and just as well, why does this pose a problem for type-III dispositionalism? The reason is that the ability of nonrelationalists to give an account of intersubjectivity that is more straightforward than the one that relationalists must cook up is, at least in my understanding of the meta-ethical debate, one of the major selling points of nonrelationalism. We have already seen that, where two theories can explain the same phenomena, we should favour the one that is most economic about ineliminable armchair luck. Because type-I dispositionalism does not postulate such luck, losing his advantage over that view in explaining intersubjectivity makes it that much more difficult for the type-III dispositionalist to justify his metaphysical expenses. So the argument about explaining intersubjectivity is not so much an independent objection, but rather a consideration that intensifies the economy problem.

6.3.4 Type-I Dispositionalism Isolates Nonproceduralistic Values

We can elaborate a bit further on the last argument in order to make it even stronger. The comparison between type-III and type-I dispositionalism becomes especially important if we can work out a type-I theory that is plausible on its own terms. In section 4.4 I have formulated two problems for type-I dispositionalism. The second problem was the problem about

[4] See section 10.5.1.

intersubjectivity, which as we have just seen does not seem to discriminate between the two views. The first problem was the problem of coming up with a plausible story about how to reject certain intrinsic desires in favour of other intrinsic desires from the perspective of the subjective motivational set as a whole. It may seem that this, at least, is a problem specific to type-I dispositionalism, and that a failure to solve this problem would give us a reason to opt for type-III dispositionalism instead.

But actually, I do not think this follows at all. If realists are right about the existence of universals, and nominalists are ineliminably unlucky about it, then it follows that nominalists must at least be able to come up with a pretty impressive theory about why universals do not exist, and how to construe various other philosophical concepts and claims in the light of that. In terms of proceduralistically distinguishable measures of success, there would have to be an ultimate nominalist theory that performs just as well as the true realist theory. For one thing, this impressively powerful and wonderfully coherent nominalist theory must clearly outperform lesser nominalist theories, and get a lot of things right that they get wrong, or at least make a lot of things work that other theories fail to make work. If this were not the case, then the nominalist project would remain stuck at some point, whereas realism would continue to improve, eventually persuading all philosophers to reject nominalism for proceduralistically justifying reasons.

In other words, it turns out that the nonproceduralist, in order to make sense of the idea of ineliminable luck, is committed to a somewhat substantial notion of proceduralistic success. This also applies in the case of ethics. To return to our favorite example, consider a bunch of Martian invaders who have a rather poor ethical theory, conceptually speaking. It is just a mess and incoherence is all over the place. But suppose that this is not because they are not serious inquirers—rather, they just have not been very good at it so far. Or perhaps ethics is just hard. In any case, suppose that we manage to demonstrate some of their incoherent conceptual commitments to them. Impressed with our arguments, they cease their practices of torture, and start working to improve their theories about their reasons for action. If nonproceduralism is true, and if their prior intrinsic desires happened to be ineliminably unlucky where ours have been lucky, then it might be that with a lot of conceptual effort, they will reach a theory about what their normative reasons are that is now immune against all our arguments, and nevertheless still tells them to

torture and eliminate us. And so they do, and we still die.

The type-I dispositionalist will simply allow that agents which would have normative reasons to torture us are conceptually possible. Whatever story the type-I dispositionalist comes up with in order to solve the aforementioned problem of discrediting some intrinsic desires at the behest of others must be a story that explains how Martians can go from a poor theory to a much better theory in the above example. Because the type-III dispositionalist also needs to be able to explain this, it may now seem that the latter could in principle agree with the former on such a story, except that the type-III dispositionalist would not want to call the successful theory of the Martians a correct theory about their normative reasons. In order to be able to express their common ground, the type-I and type-III dispositionalist may now agree that the best theory the Martians could arrive at proceduralistically on the basis of their intrinsic desires would correctly describe their "P-reasons." Furthermore, the type-I dispositionalist may agree to refer to the nonproceduralistically justifiable reasons that the type-III dispositionalist believes in as "NP-reasons." Then, the disagreement between the two views may be reformulated as a disagreement about whether P-reasons are normative reasons, and about whether NP-reasons exist.[5]

However, given the fact that there is room for P-reasons in the type-III theory—a *need* for them, even—the dispute between Martians and humans, from the perspective of type-III dispositionalism, would now seem less analogous to the dispute over universals, and more to that over mereological wholes, in the sense that it has become relatively isolated. Humans and Martians may now agree with each other about what their respective P-reasons are. In particular, we may have to agree that the Martians do have a P-reason to torture and exterminate us. There is only a residual disagreement about whether their reason to do so is also an NP-reason, but what is the import of that, if we already know that if they do

[5]This distinction between P-reasons and NP-reasons bears a certain resemblance to our earlier distinction between the R-practical and the NR-practical in section 4.3. Whereas that distinction introduced a semantic pluralism regarding the difference between relationalism and nonrelationalism, the distinction employed here introduces such a pluralism with respect to the difference between proceduralism and nonproceduralism. Furthermore, since type-I and type-III dispositionalists agree that proceduralism cannot reach nonrelationalism, they may also agree that the type-I dispositionalist's R-practical judgments map onto P-reasons, while NR-practical judgments presuppose NP-reasons. By contrast, the type-II dispositionalist would hold that NR-practical judgments are made true by P-reasons, and agree with the type-I dispositionalist that NP-reasons do not exist.

have an NP-reason not to torture us, they are in principle precluded from ever knowing that they do? Furthermore, the Martians will be perfectly capable of understanding why we feel violated, and why this feeling is internally justifiable from the perspective of our intrinsic desires in terms of our P-reasons (or P-values). The only role left for the NP-reasons dispute is that it would allow us to say to ourselves: "they are not only wronging us in the light of our *human* values, they are wrong *conceptually*!" Will that help us maintain our posture in our final moments of anguish? I am not so sure.

Metaphysically speaking, any truths about NP-reasons may now seem to have become just as disconnected from the real world, and from our understanding both of Martian and human moral practices in terms of their knowledge of their respective P-reasons, as the truth about the existence of a mereological whole consisting of two cups on a table seems disconnected from both the empirical facts about the cups and our linguistic options for referring to them.

Summarizing, the problem is that just like the isolated situation of the question concerning the existence of mereological wholes pushes us towards a "metametaphysical" theory of ontological anti-realism about an ontological fact of the matter in this matter, so the isolated situation of the question concerning the existence of NP-reasons seems to push us towards the anti-realism concerning NP-reasons defended by the type-I dispositionalist.

6.3.5 Contemplating a World Without Value

This brings me to a feature of Michael Smith's position that has puzzled me a great deal. Smith has appreciated the skepticism from others about the ability of his account to really explain the existence of nonrelationalist value and normative reasons for action, and indeed already anticipated such skepticism in *The Moral Problem* as something that may turn out to be justified. Thus, even though he is usually "optimistic" about the prospects for his brand of realism, he wants to take anti-realism very seriously as an alternative to his account. However, what puzzles me is that he feels that we should subscribe to *error theory* in such a case, and reject our entire moral vocabulary, including our concepts of value and normative reasons for action, as having been misguided:

> In deciding whether or not our moral talk is legitimate, then, it

> seems to me that we have no alternative to admit that we are venturing an opinion on something about which we can have no cast-iron guarantee. (1994, p. 187)

> If someone asks me why I believe that there are rational principles that underwrite the fact that we would all desire the happiness of our loved ones if we had a maximally informed and coherent and unified desire set, then I find myself unable to give a good answer. I am unable to give a good answer because I cannot think of any convincing reasons to suppose that there are rational principles capable of delivering anything that we would all desire if we had a maximally informed and coherent and unified desire set. And when I contemplate this fact, I find my confidence that anything has value diminishes by the second. (2006, p. 102)

What puzzles me about this, in particular, is why he would want to stick to his nonrelationalist conceptual analysis, if it would turn out that this analysis delivers no truths for our practical judgments to get right, given that he might instead also switch to a relationalist analysis in order to make sense of our moral practices in the light of the absence of any nonrelationalist values in the fabric of the world or the realm of the *a priori*. Now of course, Smith would not be the only error theorist in meta-ethics. However, there are two aspects which make the error theory that Smith seems open to rather different. The first is that it is defended in the light of a *failure* to make sense of our moral vocabulary. A meta-ethical theory such as the one defended by Mackie is best understood as being in some sense revisionist about how we should understand our own usage of moral language. In contrast, Smith's contemplation of error theory is the result of a stark *conservative* attitude towards his own conceptual analysis of the linguistic practice of ethics.[6]

Secondly, whereas Mackie seemed quite happy about making the theoretical move that he proposed, Smith really considers the prospect of error theory to be deeply disturbing. As I read Mackie's book, there is a sense of optimism in it, and the meta-ethical part breathes the sort of light-heartedness typical of analytical philosophers when dealing with

[6] I return to the distinction between conceptual conservatism and revisionism in sections 10.5.2 and 10.5.3.

meta-theoretical questions. Furthermore, the rest of the book is very constructive and positive about doing normative ethics. Now in general, Michael Smith has a similar style in his metaphysical writings, *except* in some of those passages where he discusses error theory. To him, it seems, this is not a meta-ethical question about how to construe moral talk. No, it is an existential question about whether we live in a bleak world without value. The error theory he considers is a form of *nihilism*:

> Similarly, the coherence of our evaluative concepts is a condition of the importance of everything that we do. This means that if our concepts are incoherent, then nothing we do is important at all. That is why abandoning all evaluation has nothing positive to recommend it. To be sure, if we were in a position to know that our evaluative concepts are incoherent then we would have to come up with something else that we want to do independently of our beliefs about what's important and what it not. (2006, p. 102–103)

We might do something else entirely, but no longer something that would deserve being referred to as "ethics" or "practical reason," it seems, or that could be constructed as a continuation or revision of what Smith currently understands those terms to mean. If convergence fails, then our moral vocabulary would be broken beyond repair.

But given our distinction between P-reasons and NP-reasons, that just doesn't make sense. Even if we should become anti-realists about NP-reasons, then the type-III dispositionalist may still agree with the type-I dispositionalist that our moral practices provide us with knowledge about our P-reasons, which would show that those practices were up to something all along even in the absence of NP-reasons.

In response, Smith might argue that it is only on the assumption that NP-reasons exist that the type-III dispositionalist has to accommodate something like P-reasons in order to explain the degree to which NP-reasons involve ineliminable luck. If it turns out that there are no NP-reasons, then it might be that we cannot remove the nonrelationalist presuppositions from our moral language without crippling that language in such a way as to make knowledge of P-reasons impossible as well. The very idea of a P-reason might only make sense as that of a nonrelationalist NP-reason *modulo ineliminable armchair luck*. It may be that the concept of a

relationalist P-reason cannot make sense of its own accord, let alone fulfill the conceptual role of a normative reason for action.

However, without further argument in support of this idea, I see no reason to accept it. Moreover, to some extent the proof will simply be in the eating: if the type-I dispositionalist can come up with a plausible and coherent theory that makes sense of our moral practices, then the impossibility of nonrelationalist value will not threaten the possibility of relationalist value. This is exactly how I mean to argue for type-I dispositionalism in the rest of this thesis: by showing it to be plausible on its own terms, given our background knowledge of the problems that the alternative views are facing.

III Practical Disconfirmation : A New Perspective

7 *The Affective Response View*

In part II I have discussed three reconciliatory solutions to the Facts Problem: type-I, -II, and -III dispositionalism. We have seen that each of these views gave rise to its own set of further problems. Hence, the *status quo* appears to be a kind of dialectical stalemate. It should be noted, though, that my discussion of the problems for the latter two accounts has been the most extensive. In chapters 5 and 6, I have not only identified their problems, but also discussed various ways in which they might be dealt with, only to conclude that none of these seemed satisfactory. By contrast, in the much shorter chapter 4 I have mostly done constructive work, articulating the two central remaining problems for the type-I dispositionalist in section 4.4 without yet going into much detail concerning their possible solutions. This is because these problems are much more likely to receive plausible solutions, in my view—solutions which I set out to develop in the chapters to come.

However, as I have already explained in the introduction of chapter 4, even though the Facts Problem seems to represent the fundamental nexus around which the various positions in the meta-ethical literature are organized, in my own view we should shift our focus to the Disconfirmation Problem as the proper dialectical starting point of our discussion. The first reason is that we have already seen how questions concerning disconfirmation have played a crucial role at almost every juncture during our investigations of how to construct a convincing dispositional account.

But there is a second reason, which is that I think I can provide an account of practical disconfirmation that seems rather plausible on its own right, without depending so much on metaphysical assumptions about normative facts, truths about reasons, or the nature of moral properties.[1] According to the "Affective Response View," as I shall call it, we disconfirm our practical judgments on the basis of *affective experiences* that we *did not*

[1] I have defended this account in "On the Disconfirmation of Practical Judgements" (2011). The present chapter is an adaptation of that article.

expect ourselves to have. This may be understood as an alternative to the "Principles of Reason View," the view that we disconfirm practical judgments by showing that they violate *a priori* principles of reason, which is not only implied by type-II and type-III dispositionalism, but also widely held amongst moral philosophers more generally.

In section 7.1 I will argue that the Principles of Reason View is deeply problematic, and that it is hard to see how the Disconfirmation Problem can be solved on the basis of this view. Then, in section 7.2, I introduce the Affective Response View. In section 7.3 I will argue that this view implies relationalism, and more specifically, that if we accept the Principles from chapter 1, it implies type-I dispositionalism. I continue to develop the view in section 7.4 by articulating a concept of "volitional interpretation," as I call it, which will play a crucial role when we return to the Facts Question in the next chapters in order to address the remaining issues that the type-I dispositionalist must deal with. Finally, in section 7.5 I will argue how the Affective Response View solves the Disconfirmation Problem.

7.1 Problems for the Principles of Reason View

As I explained in section 1.5.2, the Disconfirmation Problem is the problem of answering the Disconfirmation Question in non-instrumental cases. If, as the Distinctness Principle implies, motivation requires intrinsic desires that are entirely non-cognitive attitudes, which are not subject to matters of belief, then how could there be any X such that X would *both* disconfirm a belief, as the Disconfirmation Principle requires, *and* diminish an intrinsic desire of a self-governing agent, as the Authority Principle requires?

I have already noted that the Facts and Disconfirmation Problems are really two sides of the same coin. Nevertheless, it is not obvious from the dispositional solution to the Facts Problem how we might solve the Disconfirmation Problem as well. Dispositionalism makes a claim about agents under *ideal conditions* of agency, conditions that are never actually fulfilled, which gives the dispositionalist a lot of room for speculation about what might be true under those conditions. However, in order to provide a 'reconciliatory' solution to the Disconfirmation Problem (i.e., one that accepts the Disconfirmation, Authority, and Distinctness Principles), we must explain how the progress of our understanding of our normative reasons for action could be connected to changes in our motivations *as a matter of actual fact*.

This challenge has implications for the answer to the Disconfirmation Question that type-II and type-III dispositionalists are committed to: the view that practical beliefs are disconfirmed by showing that they violate *a priori* principles of reason. Let us now call this the "Principles of Reason View." We have already discussed this view at great length in chapter 5. However, in our earlier discussion we have focused on the question whether disconfirmation on the basis of *a priori* principles would lead to a convergence of all agents in their intrinsic desires as they approach their ideal selves. Even on the assumption that we *do* disconfirm our practical judgments on the basis of such principles and that under conditions of sustained self-government such disconfirmation *would* change our intrinsic desires, it turned out to be very difficult, if not impossible, to explain why this would make the desires of different agents converge. In this chapter, we shift our attention to that assumption itself: do we disconfirm our practical beliefs in this way, and does that make our desires change?

To be sure, the two questions, whether principles of reason yield intrinsic desire convergence under ideal conditions, and whether they explain intrinsic desire change in actual practice, are closely related. But whereas the former question applies specifically to nonrelationalist dispositionalists, the latter question applies more widely, to all moral philosophers who think practical deliberation is essentially a matter of applying reasonable principles to one's attitudes. Furthermore, whereas a positive answer to the first question is highly controversial in the literature—recall that even the champion of this view, Michael Smith, has become somewhat skeptical about its plausibility—, a positive answer to the latter question is fairly common, I think, and often assumed implicitly.

In particular, note that type-I dispositionalists may also adopt the Principles of Reason View. The resulting account would basically be the relationalist counterpart to type-II dispositionalism: on both views, normative reasons for action are determined by applying *a priori* principles of reason to the subjective motivational set of the agent insofar those principles may be justified proceduralistically from the perspective of that set. The only difference would be that according to the type-I proponent of the Principles of Reason View, this would not yield convergence amongst all conceptually possible agents.

Note also that a defender of such an account need not claim that this is the only way in which the process of deliberation may change our intrinsic desires. Because, as Williams argued, deliberation may also have a

constitutive role, through the employment of the imagination, for example, generating new or more specified intrinsic desires where there were none or only non-specific desires before. But it is important to realize that such developments of the subjective motivational set are not *disconfirmatory*. In such cases, nothing is *rejected* from the initial set, and none of the agent's prior beliefs turned out false.

Many philosophers feel that such a view is too relativistic, however, because it does not seem to allow us to settle any substantial moral disputes. After all, if only internal incoherence is a reason for rejection, and the standards of coherence are such that they do not guarantee convergence, then it seems that on any substantial moral issue, both opponents can rationally keep their views as long as they keep them coherently. This is another reason why I prefer the label "relationalism" to "relativism." Because whereas this implication is associated with relativism, the type-I dispositionalist can avoid it by rejecting the Principles of Reason View and adopting the Affective Response View instead, as I will argue in the chapters to come. For now, note that the Principles of Reason View has already led us to an unattractive dilemma: if we combine it with relationalism, then we get a view that is too relativistic, but if we combine it with nonrelationalism, we run into the problems of convergence that we discussed in chapters 5 and 6.

7.1.1 Disconfirmation in Actual Practice

We have already seen that the substance of principles of reason poses dilemmas for both type-II and type-III dispositionalists. The dilemma for the former is that the more substantial we make these principles, the less likely they are to be justifiable in the procedural sense to all conceptually possible agents, while the less substantial we make them, the less likely they are to make all those agents converge upon ethically interesting conclusions. The dilemma for the latter is similar, but based on the idea that increased substance also increases the cost of postulating *a priori* knowledge of that substance as a matter of ineliminable luck.

To these we may now add a third dilemma, which all proponents of the Principles of Reason View must face: while less substantial principles are less likely to yield ethically significant disconfirmations, more substantial principles are less likely to explain changes in the agent's intrinsic desires *in terms of the agent's insight into conceptual necessity*. Of course, each dilemma

exploits the same basic conceptual tension, but I think it is helpful to approach the problem from all these different perspectives in order to appreciate the vulnerability of the delicate idea that there would be some intermediate level of substantiality that is substantial enough to be ethically significant while at the same time remaining 'formal' enough to qualify as *a priori*.

Now in the case of less substantial principles of reason, it should be clear why the application of such a principle may change our intrinsic desires under conditions of sustained self-government. When I have two goals and I suddenly would realize that they contradict each other, for example, this may certainly alter my motivations, and it seems reasonable to say that I have in such a case applied my understanding of a conceptual requirement. But what about the principle of universalization? Consider Mackie's general formulation of the principle as the requirement that anyone who judges in approval or disapproval of some action "is thereby committed to taking the same view about any other relevantly similar action" (1977, p. 83). Different interpretations of the phrase "relevantly similar" then lead to different universalizability principles. Thus, in the case of the ex-racist bus driver, it might be argued that allowing a white person to occupy a certain seat would be relevantly similar to allowing a black person to occupy that seat.

Considerations of universalizability play an important role in our thinking about poverty and distributive justice, about racism, gender equality, and gay rights, and about the treatment of animals of different kinds (which display similarities and dissimilarities to us in different respects). When the bus driver disconfirmed his prior judgments, this may have involved the making of the new judgment that allowing white passengers to sit on certain seats and allowing black passengers to sit on those seats were relevantly similar. In that sense, the bus driver may have disconfirmed his segregationist judgments on grounds of universalizability. But from this we might simply conclude that honoring this level of relevant similarity is among his substantial ends, and that he has disconfirmed his prior judgments by discovering the implications of this end, or perhaps even the very end itself, if we can come up with an account of disconfirmation that would explain the discovery of ends.

It is quite something else, however, to suppose that it is an *a priori truth* that differences in skin color are irrelevant in this context, and to claim that it was the discovery of this truth that led our bus driver to the

disconfirmation of his prior practical beliefs. From the common sense understanding of universality in terms of relevant similarity as a principle used in deliberation, it does not follow that this principle is conceptually necessary and that our understanding of it is *a priori*. On the contrary, it sounds rather implausible to me to say that racial discrimination and segregation are matters of poor conceptual intuition or insufficient logical skills.

7.1.2 *The Disconfirmation Problem Remains Unsolved*

Let us once more consider the parsimony principle that Smith has suggested and that we discussed in sections 5.4.1 and 6.2:

> Reason requires that . . . (If someone has an intrinsic desire that p, and an intrinsic desire that q, and an intrinsic desire that r, and if the objects of the desires that p and q and r cannot be distinguished from each other and from the object of the desire that s without making an arbitrary distinction, then she has an intrinsic desire that s).

Suppose that a self-governing agent A desires that p, q, and r, but not that s. Suppose furthermore that we could give an interpretation of "arbitrary distinction" that would make the principle *a priori*, and that A would discover that the difference between desiring that p, q, and r, on the one hand, and desiring that s, on the other, is arbitrary under that interpretation. If dispositionalism is true, then this means that A has a normative reason to bring it about that s. Therefore, dispositionalism removes the mystery about why normative reasons and motivations go hand in hand under the ideal conditions. But that does not explain why *in the actual world*, A would acquire an intrinsic desire that s. In the actual world, intrinsic desires do not just spontaneously burst into existence. Therefore, the defender of the Principles of Reason View still has to explain how reflection on a principle of reason would create a desire that s in the mind of A without presupposing that, contingently, A already happened to desire that his desires wouldn't be arbitrary in the relevant sense.

Once again, it seems that an intuitive difference between more and less substantial interpretations of the relevant principle are relevant here. As we have seen in sections 3.3.2 and 3.4.1, the concept of self-government will give us certain non-instrumental revisions 'for free,' so to speak. When

I am making plans which require me to revise some of my earlier plans, I might decide to shift some task from a Monday to a Tuesday in a manner that would perhaps reflect the fact that I am not Tuesday indifferent, for example, and we get this sort of revisions for free even if it would turn out that we cannot derive them from strictly means-end requirements.

Now, the point is that the concept of self-government makes sense of the idea that intrinsic desires would be updated accordingly, in such a case, *independent* from our commitment to the Authority Principle. Our understanding of the idea of self-government *explains* how the Authority Principle can be true in such cases. But our understanding of the concept of self-government does not, *prima facie*, explain how an intrinsic desire to treat white people as superior to black people could be diminished upon *a priori* reflection. Instead, it is part of the burden of defending the view that racism is an *a priori* violation to make it plausible that it is rational reflection under conditions of sustained self-government that has really changed intrinsic desires and thereby reduced racist behavior.

I am not sure that this burden of proof for the Principles of Reason View is often recognized or fully appreciated. Smith's notion of systematic justification, which we discussed in section 5.3, is in his words "the most important way in which we create new and destroy old underived desires" (1994, pp. 158-159). In his subsequent discussion of this idea, he talks freely about how agents may "add" an underived desire to their existing desires. It seems to me that with all this talk of "creating" or "adding" *underived* desires on *rational* grounds, Smith may be stretching the limits of the Humean theory of motivation too far.

It may be less obvious why this would be a problem in the case of type-III dispositionalism. As we have seen, this view is not committed to the idea that deliberation has to be proceduralistic, and may therefore allow that some rationally required desire revisions cannot and need not be explained in actual practice, because they will never actually happen. However, even though this sort of response might take care of the Martians, it does not explain our example of the ex-racist bus driver. Practical disconfirmation really does happen, even in morally substantial cases, and insofar it does, no ineliminable luck had apparently been involved.

7.1.3 Personal Choices Are Not Disconfirmed By Principles

Many of the decisions that we make about our personal lives concern options that would be equally permissible from any "principled" point of view, but that nevertheless may not be trivial to us. If we change our minds about such matters, then we may want to speak of disconfirmation. It is an unfortunate fact, for example, that many couples break up their relationships shortly after they made decisions towards further commitment. People come to realize that it's not going to work just *after* they bought the new house, or went on their honeymoon, or had gotten pregnant.

If a woman spends seven years of her life with a man, and after much deliberation decides to marry him and to sell her apartment in order to move in with him, then it seems plausible to say that she *got it wrong* if after three months she decides to go looking for her own apartment again. It would be odd to argue that *at the time* of her initial decision it was actually best for her to agree to the marriage, but that as a matter of unlikely and unpredictable misfortune, her *preferences* concerning husbands and housing, which had been stable for the past seven years, suddenly changed shortly after she made her decision. Nevertheless, it also seems implausible to insist that she must have violated some principle of reason in her deliberations when she was thinking about marriage and living together. What would be the *conceptual mistake* in such a decision? Principles of reason may constrain our thinking, but they cannot dictate the choices that we should make in our personal lives. Therefore, the Principles of Reason View cannot explain how these kind of judgments can get it wrong.

To be fair, it may be that even though the Principles of Reason View fails to handle these cases, this may be accounted for by nonrelationalist dispositionalists if they would subscribe to a 'mixed' or 'hybrid' account of practical disconfirmation, according to which moral disconfirmations are based on reasonable principles, while disconfirmations in other practical matters are disconfirmed in some other way. As we have seen in sections 1.3.2 and 5.1.2, the distinction between relationalism and nonrelationalism does not apply in matters of taste. However, this does mean that we do need to look for a new account of disconfirmation in order to handle the cognitivism of personal choice and preference, so to speak. Furthermore, once we would have such an account, we may wonder whether it would not outperform the Principles of Reason View in explaining moral disconfirmations as well.

7.2 The Affective Response View

According to the view that I want to propose, our practical beliefs can be disconfirmed by our own *affective responses* to our self-governed actions, or to the intended consequences of those actions, *insofar as we did not expect ourselves to experience those responses*. If a thief judges that he has a good reason to steal from someone, and does not expect himself to feel very guilty about it, then he may come to doubt his initial judgment if after the theft he gets overwhelmed by feelings of guilt. On my proposal, this is because such feelings have the power to disconfirm practical beliefs. I call this the "Affective Response View."

First, let me explicate the notion of an "affective response." By this I mean any affective attitude, experience or sensation that can be understood as a response to an event that preceded it. This might be any sort of event, but our discussion concerns affective responses to actions or the consequences of actions. Whether an affective experience should be understood as a response to a certain event may be a subject of interpretation. If the affective experience is an intentional attitude of remorse about having murdered someone then the experience is clearly a response to the murder, but if the affective experience is a general feeling of joy without any specific content, then it may not be clear whether or not this is a response to a certain previous act or event. I will say more about the interpretation of responses in section 7.4 below.

Since affective responses are backward-looking attitudes, as it were, they may be thought of as a kind of counterparts to desires, which are typically forward-looking. The feeling of satisfaction as a result of an action is such a counterpart to the feeling of desire that motivated the action. However, sometimes forward-looking desires may also be understood as affective responses themselves. For example, suppose one decides to become a vegetarian. If, subsequently, one's desire for meat increases, this may be interpreted as a response to the (consequences of the) decision. Furthermore, every affective response to an event may be understood as a desire in a very loose sense—as the desire that P, where P is the proposition that the event occurred (in the case of a positive response) or the proposition that the event did not occur (in the case of a negative response). Finally, every affective response contributes to the *resultant desire* of the agent at the time of the response in the sense defined in section 1.4. For example, suppose that I have eaten seven slices of pizza and I am

wondering whether or not to eat the eighth slice. My affective attitudes might be mixed. On the one hand, I desire to eat it because I want to taste some more. On the other hand, the way my stomach feels tells me that I have already eaten too much. This affective response may outweigh the desire to eat the last slice and make a decisive contribution to my resultant desire not to finish the pizza. Hence, affective responses can be efficacious motivational states.

Let us now turn to the role that affective responses play in practical disconfirmation. Common examples of affective responses that sometimes make us rethink our prior judgments are feelings of regret, remorse, guilt, shame, embarrassment, jealousy and boredom. Nevertheless, disconfirmation is not intrinsic to these responses. Rather, whether an affective response disconfirms a prior judgment depends on how that response is related to other affective states. The general idea behind the Affective Response View is that self-governing agents will expect a kind of "match" between the affective states that motivate their actions on the one hand, and their overall affective responses to the consequences of those actions on the other hand. If John desires to see Rome, and judges that he should spend his money on a holiday to Italy's remarkable capital, then he will expect his visit to Rome to be a pleasurable and rewarding experience. If the holiday would fail to meet these expectations, then John might start to think that his money would have been better spent differently.

However, we will rarely expect our responses to completely match the desires that we decided to act upon. If Carol deliberates about whether or not to quit her job and accept another one, then she will probably both have desires in favour of quitting and desires in favour of staying. Should she decide to make the change, then in the light of her multitude of desires, she may expect both positive and negative responses. In the short term, she might even expect the negative responses to be stronger because of the stress and the various difficulties of adjustment. Nevertheless, it seems plausible that if she decides to quit, her expectation will be for her overall response to be more positive *in the long run* than if she would have stayed. Should her actual responses, after a while, give her reason to believe that she would have been happier if she had kept her old job, then her responses may disconfirm her prior judgment.

This does not mean that the Affective Response View commits us to a hedonistic egoism about maximizing one's own happiness—at least not under any shallow interpretation of the terms "hedonism," "egoism" or

"happiness." Suppose that Jack is in a hurry, and decides not to help an injured person on the street. If Jack would feel ashamed of himself afterwards, and if he were to conclude that he should have helped the injured person, then a defender of the Affective Response View might argue that the feeling of shame disconfirmed Jack's prior practical belief and made him adopt the practical belief that he had a normative reason to help the injured person. But that does not mean that Jack merely had a normative reason to do so in order to prevent himself from feeling bad about himself, which would have been a purely instrumental consideration. Rather, it means that his feeling of shame informed him of the fact that helping the injured person was more important to him than he initially thought.

7.3 Relationalism and the Normative Will

Although the Affective Responsive View does not commit us to shallow egoism or hedonism, it does imply relationalism. The practical beliefs of agent *A* about what she has normative reason to do are subject to disconfirmation by *her* affective responses, which may tell her something about what is important to *her*. Nonrelationalist dispositionalists might want to object that affective responses could be intuitions about substantial *a priori* principles of reason, which would carry us back to the Principles of Reason View. But it is not clear how this suggestion would make the problems for the Principles of Reason View any easier. Our affective responses are a result of contingent psychological mechanisms, and without a proper explanation of why their content would involve *a priori* truths about reasons for action, we have no reason to think that they constitute anything other than a matter of empirical fact.

Moreover, even though the experience of an unexpected affective response must be understood within the context of deliberative activity on the part of the agent, as I will argue in the next section, the deliberative *modus operandi* is nevertheless hardly that of the formal application of conceptually necessary principles to prior ends in order to revise those ends. Perhaps the nonrelationalist could simply accept this, and opt for an account of moral perception. But what is unexpected about the affective response is precisely the affective, non-cognitive part of the response. If we accept the Distinctness Principle, then what is perceived is the fact that the agent turns out to have that particular non-cognitive attitude under

these circumstances, but this attitude is not itself a percept of some further substantial fact or principle existing independently of the agent. *Prima facie* then, for any agent A, the content of A's affective responses seem to give empirical information about A, rather than *a priori* information about all conceptually possible agents.

Thus, suppose that Sharon is a vegetarian. She becomes friends with Marc and David, who are both used to eating meat and never felt bad about it. However, once Marc gets to know Sharon better, and starts to consider things from her perspective, he discovers that he begins to experience negative feelings about eating meat. He starts to feel guilty about the idea that animals were killed in order for him to enjoy a particular eating habit, even though that habit is not necessary in order for him to live a healthy life. On the basis of these feelings, Marc starts to disapprove of the killing of animals by humans for food, thereby disconfirming his prior belief that it was okay to do so. Should it be the case, in virtue of *empirical facts about Marc*, that this is also how Marc would feel under ideal conditions of rational agency, then it does not follow that *any conceptually possible agent* would have to feel the same under those conditions. Thus, suppose that David does not develop negative feelings about killing animals for food at all, not even after extensive discussion with Marc and Sharon. Under ideal conditions of rational agency, David may still have a resultant desire to eat meat, even though Marc and Sharon may have the resultant desire under those conditions that nobody would eat meat.

However, it may well be an empirical fact about human psychology *in general* that there are certain actions that all of us do have normative reason to disapprove of. For example, it may well be an empirical fact that every human being would under ideal conditions of rational agency have the resultant desire that no sentient being ever be tortured. Therefore, the Affective Response View does not prevent us from arguing, say, that the Nazis got their practical judgments wrong. Psychologically, the Nazis had so much in common with us that it seems plausible that under different circumstances, they would have had the same feelings of horror about the Holocaust that we do. We may ascribe the fact that they did not actually feel this way to a type of upbringing and training that, effectively, removed them further away from the ideal conditions of rational agency, and thereby made it impossible for them to fully understand their own affective dispositions. In other words: every SS officer who believed that he had a normative reason to torture and murder his victims may have

gotten *himself* wrong.[2]

What the Affective Response View does rule out is that every conceptually possible agent would get it wrong when judging in approval of torture and genocide. If, as a matter of empirical fact, the Martian invaders would have no disposition whatsoever to sympathize with us, then it will be impossible, on the Affective Response View, to disconfirm their practical beliefs. What this means is that if the Nazis got their practical judgments wrong, they got it wrong precisely because they were *not* alien monsters: they got it wrong *as human beings*.

Allow me to introduce some additional terminology at this point. According to the type of relationalism that we have been discussing, every agent has his own source of normative reasons for action, which consists in certain empirical facts about his psychology. Let us call this source the "normative will": let us say that A wants to ϕ under circumstances C in the "normative sense," or that ϕ-ing under C is "part of the normative will" of A, if and only if under the ideal conditions of rational, self-governing agency, A would desire in the resultant sense that under the circumstances C, he would ϕ. One may think of the normative will as a complex of the agent's deepest attitudes of caring and love, which establish what is most important to him. In this respect, the proposal bears similarities to recent work by Harry Frankfurt (2004; 2006), which I will explore in the next chapter. For now, however, note that these attitudes may be phenomenologically *opaque*. The normative mode of wanting is a mode of wanting that we may ourselves be ignorant of: at the time, Jack did not know that he wanted, in the normative sense, to help the injured person, and SS officers did not know that in the normative sense, they did not want to torture and kill their victims.

Thus, the view that we have now arrived at is a form of type-I dispositionalism: normative reasons for action are constituted by our desires under ideal conditions, and what our desires would be under such conditions is 'strongly dependent' (see section 5.1.2) on contingent empirical facts about our actual selves. However, note also that the view is clearly different from the combination of type-I dispositionalism with the Principles of Reason View. On that account, the facts about normative reasons were merely

[2]The idea that his training prevented the SS officer from understanding his ideal self raises the question of whether it would have been possible to make that training undone. If not, then we may wonder what sort of counterfactual life histories we should consider in order to construe his ideal self. At what point would it become the ideal self of a *different* person? I return to this matter in section 9.1.4.

facts about what the coherent version of one's actual views would look like. Instead, on the type-I account that I am proposing, the facts about our normative reasons include contingent facts about ourselves that are not directly accessible through introspection, but require our interaction with the outside world in order for us to discover them. Thus, by interacting with the outside world, we not only learn empirical information about the world that may be relevant to our instrumental deliberations, but we also learn empirical information about *ourselves* that allows us to deliberate on our ends.

7.4 Volitional Interpretation

I have claimed that a self-governing agent will expect the intended consequences of her actions to generate the most positive overall affective response, in the long run, compared to the alternatives that she might have chosen. The underlying intuition is that in the long run, our overall responses to our actions tell us something about what we want in the normative sense: that they will approach the resultant desires of our ideal selves, so to speak. However, the notions of "overall response" and "in the long run" are of course totally vague and abstract. In practice, our responses change from moment to moment and from situation to situation, and it is often hard to determine which of our affective experiences are responses to which consequences of our actions. Therefore, whether an affective experience disconfirms a practical belief is always a matter of *interpretation*: we must judge what the experience *means* to us.

Suppose, for example, that a student feels an unexpected embarrassment after having asked a question during a course meeting (perhaps it turned out that it was not a very intelligent question). Does that mean that he shouldn't have asked it? Perhaps it does, but perhaps it doesn't. The student might also conclude that this merely reveals that his questions can be as unintelligent as those of anybody else, and that perhaps he may be more easily embarrassed about this than he thought he would be. But since he won't get any smarter by not asking such questions, he might reason for himself, it was still a good idea to ask the question anyway. In other words: the negative response of embarrassment about his action does not *intrinsically* have higher normative authority than the positive affect of curiosity that initially motivated the action.

In fact, the judgment that an agent must make in order to determine

whether a response disconfirms an action is of exactly the same kind as the judgment that he had to make before the action in order to determine whether he wanted to act upon the desire that motivated the action: it is just another practical judgment, a judgment about whether the affective experience is an expression of his normative will. Our concern to get our practical judgments right brings with it a concern to know whether our affective responses are appropriate or not. The point of the Affective Response View, however, is that it is only on the basis of *other* affective responses that an agent could be justified in judging that his current affective response is *in*appropriate.[3]

Let me illustrate. Suppose that the student does conclude that he shouldn't have asked the question. The next time that his curiosity arises, he remembers the unpleasant embarrassment, which is itself an unpleasant experience that counteracts his motivation to ask another question. Suppose that he decides not to ask the question this time, and that he is self-governing—i.e., the unpleasantness of his memory of the previous time is stronger than his desire to ask another question. By not asking the question, he might save himself another embarrassment, but when the course is finished, his curiosity is unsatisfied, which makes him feel frustrated. Perhaps, again, more than he would have expected. And again, this is an experience that requires interpretation. What does it mean? Perhaps it means that he should visit the *Wikipedia* and try to find the answer to his question for himself. But it might also mean that he does not want his fear of embarrassment to prevent himself from asking what he really wants to know, and that he should have asked the question after all. In that case, the judgment that the embarrassment was a disconfirmation would itself be disconfirmed, and the initial judgment which led him to ask the first question would be *confirmed*. This shows that deliberation

[3]Some people might also believe that *most* of their affective responses are inappropriate, that their *overall* affective response is inappropriate, or perhaps even that their overall response would still be inappropriate under ideal conditions of rational agency. An unmarried man might have an overall affective response in support of his promiscuous life, for example, and still believe that he should not be living such a life, because his religion teaches that sex is only allowed within marriage. Such an agent would have to reject the Affective Response View, but that does not mean the Affective Response View is false. Instead, it means that if the Affective Response View is true, then given that his overall response is in favour of his lifestyle, his practical judgment must be evaluated in terms of a pluralistic semantics that explains its falsehood in terms of a misunderstanding on the agent's part about what it means to make a practical judgment. I have briefly discussed this idea in section 4.3, and will return to it at the end of this thesis in section 10.5.3.

is an ongoing process of what I shall call "volitional interpretation": the interpretation of our affective experiences in order to determine which of them express our normative will.

Let me now make a number of brief additional remarks about the notion of volitional interpretation, in order to give a better overall idea of what I am driving at. I will elaborate on each of these remarks in the chapters to come.

7.4.1 The Normative Will as an Explanatory Pattern

First of all, note that we are now dealing with two directions of explanation. The moral philosopher who wants to answer the Facts Question is interested in knowing whether we can explain the existence of normative reasons for action in terms of facts about the affective attitudes of agents. But the idea of volitional interpretation is that *as deliberating agents*, we often reason in the opposite direction: we want to know whether a certain reason for action would explain the affective attitudes we are experiencing. If I have a normative reason to ϕ, after all, then insofar I am rational, well-informed and in control of my own agency, I should expect to find a certain pattern in my actual experiences in support of ϕ. Thus, the hypothesis that I have such a reason may be explanatory relevant, in a structural sense, to my actual motivation to ϕ. Metaphysically, then, the nature of normative reasons for action will be something like *patterns* or *structures* of affective dispositions.

7.4.2 Volitional Interpretation as a Social Practice

A second point that I want to highlight is that the analysis of deliberation as volitional interpretation allows us to understand deliberation as a *social practice*, even though the core of the analysis is individualistic. For one thing, different people often display similar responses in similar situations as a result of underlying psychological structures that we all have in common due to our shared environment and biological ancestry. Therefore, we can learn from each other's mistakes, and we can search for "human values"—things that all human beings would want under the ideal conditions of rational agency as a result of the empirical facts about human psychology. This allows us to argue that the Nazis got their practical judgments wrong, for example, as I have outlined in the previous section. Furthermore, once you come to believe that you have even more

in common with a specific group of people, or with a particular person, then it becomes even more plausible to expect similar responses in the situations which pertain to those commonalities. I will develop this idea later on in the thesis, when we return to the problem of squaring type-1 dispositionalism with the Intersubjectivity Principle.

Note that we may also improve our interpretation of what we really want on the basis of discussion with those who want something different. Such discussion often forces us to articulate more precisely the reasons that we have for our practical beliefs, which may lead us to revise those beliefs in subtle ways and to make them more sophisticated and precise. Furthermore, there may be cases where someone close to me understands what I really want before I understand it myself, even though it need not be something that *she* really wants.

In fact, we often do not take alternatives seriously until they are being demonstrated or suggested to us by certain individuals, groups, or media. In the absence of any social pressure to change their views, people generally stick with their initial gut feelings, and self-confirmation bias is everywhere in our psychology (Haidt, 2001). That is why we rarely disconfirm our beliefs about what we care about most. The problem is not just that we protect our self-image by ignoring evidence, or by giving heavily biased interpretations of unexpected affective responses. The problem is also that we rarely experience unexpected responses concerning our most cherished practical beliefs in the first place, because our responses do not arise independently of those beliefs.[4] Thus, it is possible that an agent experiences no affective responses against ϕ-ing, and that there is no doubt in his mind that he has a normative reason to ϕ ("he knows what he wants"), while his normative will is actually opposed to ϕ-ing, in virtue of the fact that he would eventually experience massively adverse responses with regard to ϕ-ing, once he would start taking the possible reasons not to ϕ more seriously.

Therefore, volitional interpretation may benefit from attempts to break out of dogmatic self-assumptions, and social influence can be a way of making people consider new alternatives. Of course, in reality, social practice often only makes things worse, because people will prevent each other

[4] An analogy may be drawn with issues concerning the theory-ladenness of observation in the philosophy of science. This analogy between science and ethics is pursued, to some extent, in Churchland (1995, ch. 6, esp. p. 146–147). See also his (1989, pp. 188–196) on the theory-ladenness of observation in general.

from starting to doubt the views that constitute the identity of their group. Nevertheless, many of our most fundamental changes in our practical beliefs have occurred in the context of social developments. The case of the Montgomery Bus Boycott and the rise of the civil rights movement would be an example of such a development. Thus, we might argue that as a result of this development, people like our fictitious bus driver did not just start taking a different point of view seriously, but they also started experiencing different affective responses to established practices, including their own actions, which eventually led to the disconfirmation of their segregationist practical beliefs.

7.4.3 The Hypothetical Nature of Volitional Interpretation

A related point about volitional interpretation is that it never reaches *final* results. Practical beliefs, on this view, are always *hypothetical*: they are forever subject to revision in the light of new experience. This does not rule out that an agent may have good reason to be "fully resolved" in some of his practical judgments, as Frankfurt has put it, in the sense that the agent may have the "belief that no further accurate inquiry would require him to change his mind" (1987/1988d, p. 169). Sometimes we *do* know what we want. For example, nowadays we may well believe that no inquiry will ever disconfirm our practical belief that people should be treated equally regardless of the color of their skin or their sexual orientation, say. Nevertheless, in the light of the theory of volitional interpretation, it would probably be wise for most of us to keep an open mind and to take alternatives to *most* of our present practical beliefs seriously. Furthermore, the hypothetical nature of volitional interpretation *does* rule out Frankfurt's notion that any particular affective experience could reveal a "volitional necessity," a directly experienced constraint, imposed by the normative will of a person, on what he can and cannot bring himself to do (2004, pp. 46–49; 2006, pp. 33–34). Instead, on my account, if we cannot bring ourselves to do what we had judged that we should do, then it is always a matter of interpretation whether we are experiencing disconfirmation or merely an impairment in our self-government. I will develop this criticism in further detail in the next chapter.[5].

[5]See section 8.3.1

7.4.4 *Interpretative Strategies and Normative Ethics*

Another point to note about the concept of volitional interpretation is that it is compatible with different methods and approaches to deliberation, and that it also allows us to combine those approaches. Thus, it might be that in certain areas of ethics and political theory, it will be very useful to try to formulate individual or shared practical beliefs using *principles*, such as principles of universalizability. Note that in order to make sense of these principles, we can make our interpretations of notions like "relevant similarity" or "arbitrary difference" as substantial as we want, without having to worry about keeping them formal enough in the way that nonrelationalist dispositionalists have to do so. Note also that we might want to adopt certain principles but admit that we want to allow certain exceptions to them. On other moral issues, however, we might have better luck if we try to describe *values*, or if we reflect on how we might embody certain *virtues*. Perhaps that in certain domains of our personal lives, it would help if we took a *narrative* approach, by trying to articulate what kind of story we would want to tell about ourselves. From the point of view of will interpretation, these are merely different *interpretative strategies*, and every strategy is valid as long as it yields disconfirmable expectations about our future affective responses.

7.5 The Disconfirmation Problem Solved

Let us now return to the Disconfirmation Problem. According to the Affective Response View, our practical beliefs may be disconfirmed by unexpected affective responses. So why should that result in a non-instrumental change in our motivations? The answer is that the unexpected affective responses *are* the changes in our motivations. After all, affective responses are themselves motivational experiences, which influence our future behavior. If they come unexpected, then it is likely that we were not used to having them, which may signify a change, and alter the balance of affective attitudes in such a way as to lead to a new resultant desire. If we judge that they nonetheless express what we really want, then the motivational change and the belief change may correspond to each other in such a way that we maintain the same level of self-government.

Recall the example of the student who got embarrassed after having asked the unintelligent question. Suppose that the student judged that he should not have asked the question, and decides that he won't ask

such a question again. Then the embarrassment, through his unpleasant memory thereof, will have changed his motivational disposition, and—if it is stronger than the curiosity—will prevent him from asking another question.

Note that this solution is very much in the spirit of the dispositional solution to the Facts Problem as I introduced it in section 4.1. We remove the mystery about why the disconfirmation would imply a motivational change, under conditions of sustained self-government, by claiming that under those conditions, the disconfirmation simply *is* the motivational change.

8 A Normative Reality Within Ourselves

The argument in the previous chapter lead me to the notion of a *normative will*, which I briefly characterized as a complex of the agent's deepest attitudes of caring and love, which establish what is most important to him. The idea that caring and love are constitutive of practical normativity has been defended by Harry Frankfurt in some of his recent work (2004; 2006). Frankfurt rejects what he calls "normative realism," the view that some things are inherently important regardless of whether we care about them (2006, p. 33), which I have called nonrelationalist realism in section 1.3.2, and of which type-II and type-III dispositionalism are varieties. In contrast, Frankfurt proposes to understand mistaken practical judgments in terms of mistakes about what we actually care about: that there are contingent empirical facts about our own attitudes that our practical judgments get right or wrong. Thus, there is a normative reality, but this reality is "within ourselves," as he puts it:

> In matters concerning practical normativity, the demanding objective reality that requires us to keep an eye out for possible correction of our views is a reality that is within ourselves (2006, p. 34).

Let us call this the "Inner Reality Thesis." It is a version of relationalist cognitivism, and as we shall see in this chapter and the next, my brand of type-I dispositionalism is a version of the Inner Reality Thesis. Frankfurt's version of the thesis is different, however. In this chapter I will raise a number of problems for his account and sketch how my own proposal can solve them. I develop my own account in further detail in the next chapter.

8.1 Inner Reality Theories

In section 7.3 I have argued that the Affective Response View commits me to a version of type-I dispositionalism, according to which every practical reasoner has a *normative will*, which is constituted by empirical facts about his actual self. I have not said much about the nature of these facts so far, but I did note an important difference in comparison to the version of type-I dispositionalism that one might defend in combination with the Principles of Reason View. According to the latter, 'principle driven' type-I dispositionalism, the facts about our normative reasons would simply be facts about what we get when we apply certain principles to our actual motivations. As we have seen, this is just type-II dispositionalism *minus* the convergence, so to speak. But without convergence, such a view seems too subjectivist or relativist: whatever motivations you happen to have provide you with normative reasons as long as you have them coherently, and the standards of coherence are too weak to settle moral disputes.

By contrast, the 'response driven' type-I dispositionalism that I am proposing assumes that there are facts about *opaque* attitudes that we may not be actually experiencing yet, or be motivated by so far, but that may reveal themselves on future occasions, so to speak, through our unexpected responses to the consequences of our own actions. These facts belong to our contingent psychological make-up, and so they can only satisfy relationalist truth-conditions: the Martian invaders will not share them. Nevertheless, because the attitudes in question are opaque, making judgments about them is not so subjective any more: our responses may be unexpected in ways that we could not have derived by *a priori* means from the prior attitudes that we were *aware* of so far. Instead, these responses may constitute truly empirical discoveries about ourselves.

Furthermore, because the opaque attitudes may be shared among agents whose experienced attitudes are more opposed, the account is not so relativistic any more either: if A and B make conflicting, yet internally coherent, sets of judgments on the basis of the attitudes that they are subjectively aware of, it may turn out that their opaque attitudes are nevertheless shared in such a way that A may disconfirm his judgment and come to agree with B. Thus, even if conflicts between Martian and Earthling cannot be resolved through deliberation, conflicts between different human beings might.

The idea that a relationalist may appeal to such empirical, 'inner' facts

in order to account for practical disconfirmation is precisely what the Inner Reality Thesis claims. Thus, it seems to me that Frankfurt is looking for the same middle-ground between realism and relativism that I am aiming to achieve, even though he does not argue for it on the basis of a commitment to a dispositional theory of value. Nevertheless, his defense for the thesis is driven by concerns that are similar to the intuitions behind my own line of argument. In particular, Frankfurt argues that the substance of our reasons can never be derived from pure rationality alone, and that a more substantial source of normativity cannot be external to our attitudes as that would fail to explain the feature of normative truths that they "require that we submit to them" (2006, p. 34). Furthermore, Frankfurt has spent a great deal of his writing trying to explicate the relations between higher-order attitudes and desires, while his concept of desire seems to reflect a broadly Humean independence from matters of belief, even though this latter point is never made very explicit. Finally, Frankfurt wants to defend the account against the charge of being "unacceptably noncognitive and relativistic" (2006, p. 26), despite its strong dependence on contingent attitudes, by arguing that those attitudes may be common features of human psychology:

> People care about many of the same things because the natures of human beings, and the basic conditions of human life, are grounded in biological, psychological and environmental facts that are not subject to very much variation or change (2004, p. 27).

Thus, the view allows that all human beings have normative reason to promote equal rights for same-sex couples in comparison to heterosexual couples, for example, even though many people, especially in non-western cultures, currently believe otherwise. In its general form, then, the Inner Reality Thesis is completely neutral about how much of our moral disagreement reflects differences between our respective inner realities, and how much reflects a misunderstanding of those realities.

Although the thesis is meant to avoid the "queerness," to use Mackie's term, of an external normative reality, the idea of an inner reality of course raises metaphysical and epistemological questions of its own. How is this normative reality constituted, exactly, by the empirical facts about our attitudes? How do we access this reality within ourselves, how do we "keep an eye" on it so as to correct mistakes in our practical views? And

why should we submit to this reality, i.e., how does this reality make it required for us to act in accordance with its normative implications? Let us call comprehensive attempts to answer these questions "inner reality theories."

In Frankfurt's own inner reality theory, the attitude of *caring* does much of the explanatory work. Caring is a matter of contingent empirical fact, yet on Frankfurt's view, it may resist a reductive analysis in terms of the content and motivational weight of desires. Caring is neither affective nor cognitive: rather, it is a *volitional attitude*, as Frankfurt puts it. This volitional attitude involves treating certain things as reasons, yet it also establishes the objective normativity of reasons. I will argue that a single type of attitude cannot fulfill all these roles at the same time. Furthermore, I will argue that volitional attitudes cannot be accounted for empirically if they resist analysis in terms of either cognitive or affective attitudes. In order to resolve these problems, I will propose to distinguish two types of volitional attitude. The first, which I shall call the "cognitive will," is to be analyzed in terms of what the agent believes. The second type is the normative will, which is to be accounted for in terms of empirical facts about the agent's affective dispositions. Which of the two is referred to when we say that someone cares about something depends on the context of the utterance.

8.2 Frankfurt's Volitional Inner Reality Theory

Since 1971, when he published "Freedom of the will and the concept of a person," Frankfurt has been developing an insightful theory about the different ways in which an agent can be said to "want" something (1971/1988b; 1999b). One of the goals of this theory is to account for the intuition that a human being can want something "as a person," and that this mode of wanting is richer in its psychological structure than that of merely desiring something in the sense of being attracted to it. Frankfurt called this richer mode of wanting "volitional," and set himself the task of analyzing its structure—first in terms of higher order desires, later in terms of ideas about wholeheartedness, caring, and love. Moreover, the theory was meant to explain freedom of the will, since Frankfurt argued that a person acts of his own free will whenever he manages to act upon his volitions. Thus, Frankfurt's work has often been discussed in the context of the free will debate.

In Frankfurt's more recent publications, the analysandum of the theory seems to have shifted. Although a certain notion of free will is still accounted for (2004, p. 20; 2006, pp. 14–16), the main purpose of the theory has become to explain practical normativity. According to Frankfurt, the question of how we should live is, properly understood, a question about what it is that we care about:

> The totality of the various things that a person cares about—together with his orderings of how important to him they are—effectively specifies his answer to the question of how to live (2004, p. 23).

This means there are only two ways in which practical judgments can get it wrong: they may involve instrumental errors about the world and how to achieve one's goals in it, and they may involve errors of self-knowledge in the form of misunderstandings of what we care about:

> Once we have learned as much as possible about the natural characteristics of the things we care about, and as much as possible about ourselves, there are no substantive corrections to be made. There is really nothing else to look for so far as the normativity of final ends is concerned. There is nothing else to get right (2006, p. 50).

Hence, the notion of caring plays the central role in Frankfurt's inner reality theory. We can therefore understand the theory as consisting of two parts: first, the account of normative reasons in terms of attitudes of caring; and second, an account of the nature of those attitudes of caring themselves. These two accounts may be evaluated more or less independently, because the notion of caring may be thought of as a philosophically interesting analysandum in its own right. It is philosophically interesting because "caring about something is essential to our being creatures of the kind that human beings are" (Frankfurt, 2004, p. 17). Thus, one may agree that Frankfurt's analysis of caring correctly captures this essential feature, but disagree with his analysis of normative reasons in terms of caring, or even disagree that normative reasons are to be analyzed in terms of caring at all. Conversely, one may agree that Frankfurt's intuitive concept of caring does explain normative reasons for action, but reject the way in which he analyzes the nature of caring itself. Let us first take a look at the latter part, Frankfurt's account of caring.

8.2.1 The Meaning and Nature of Caring

The first thing to note about caring is how it differs from mere desiring. One may desire something without caring about it at all. In fact, an agent may not care about anything and yet desire all sorts of things. One may disagree about which entities belong to this class—perhaps mice and rats, perhaps spiders and insects, or maybe thermostats and chess computers—but the point is that in an important sense, such agents are not *persons*. They are "wantons," who do not care about what it is they want (Frankfurt, 1971/1988b, p. 16). This gives us a clue about the analysandum for a theory of caring: it must be a theory that accounts for "strong" agency—the type of agency that is distinctive of personhood.

A second thing to note is that even agents who are persons may not always be motivated to act in accordance with what they care about. Frankfurt's example of the unwilling addict illustrates this: he cares about being healthy, but he keeps using drugs because he is driven by a desire that is too strong for him to resist, even though he does not want to be driven by this desire at all (1971/1988b, p. 17). The example shows that the affective strength of a desire does not by itself determine whether that desire speaks for the agent—whether it carries "agential authority," as Michael Bratman calls it (2009, p. 430). Various phrases have been employed by Frankfurt and his commentators to articulate the meaning of this form of authority: the agent "identifies with" the desire, he is "fully behind" the desire, or the desire expresses what the agent "really wants" (Frankfurt); the desire is "endorsed" by the agent (Watson, 2002/2004d, p. 112); or it is a desire "of his own" (Bratman, 2003). In contrast, desires such as the craving of the unwilling addict, which do not carry agential authority, are "external to the person," in Frankfurt's words (1976/1988c). We "regard them as disconnected from us or as alien intruders by which we are helplessly beset" (2006, p. 8).

In his recent work, Frankfurt distinguishes between a weaker and a stronger sense of agential authority. The weaker sense is that of "accepting" a desire and "consenting" to being driven by that desire. Accepting desires in this way means that we have "taken responsibility for them as authentic expressions of ourselves." Any desire that is authoritative in this sense is therefore not regarded as external or alien. If one acts upon desires that one accepts in this sense, one acts out of free will, in Frankfurt's view.

We may wonder whether the kind of responsibility that this notion of free will delivers is equivalent to the sort of responsibility that would

justify blame or retribution, however. It might not be necessary for blame, as we may want to blame people for many acts, and the dispositions behind those acts, that they weren't too happy about themselves. On the other hand, it might not be sufficient either, as certain conditions may undermine blameworthiness without upsetting the psychological structure of agential authority. To borrow Watson's terminology, the concept of responsibility that Frankfurt's account may deliver is that of "attributability," whereas the one that underlies blame and retribution is that of "accountability" (Watson, 1996/2004c).[1]

Nevertheless, attributability is an important notion of responsibility in its own right, which can be found elsewhere in philosophy—it is what Charles Taylor has called "responsibility for self" (1976/1982) and one might argue that this notion is also closer to responsibility in Sartre and other continental authors than the type of accountability on which analytic philosophers tend to focus. Furthermore, the idea that an agent freely wants what she wants when she can fully accept the fact that she wants it is certainly intuitive, and may be traced back to the work of Spinoza (Frankfurt, 2006, p. 16–17).

Consenting to a motivation does not imply, however, that one also has to care about being so motivated. For example, suppose you usually prefer a certain type of dessert when you have dinner in a restaurant—tiramisu, say. As long as you're not too much of a health freak you're probably not going to have any problems accepting the preference for tiramisu as your own, but at the same time you're probably also not going to care should you start to prefer a different type of dessert. So caring seems intuitively to involve a stronger, more obligatory form of agential authority than consent.

8.2.2 The Hierarchical Account

Let us now see how Frankfurt attempts to account for these two different forms of agential authority in terms of what are, in his view, matters of empirical fact. From his intuition that acting out of free will involves really wanting to be driven by the desire that one is motivated by, Frankfurt has argued from the beginning that it must be a necessary condition for freedom of the will that the agent has a *second-order* desire for the relevant

[1] Frankfurt himself does not seem to make this distinction. Thus, his famous criticism of the "Principle of Alternative Possibilities" (1969/1988a) clearly addresses what Watson would call "accountability," paving the way for a new type of compatibilism, and his discussion of responsibility elsewhere suggest that his positive account of free will is meant to fit this bill.

first-order desire to be *efficient* (1971/1988b, p. 16). This condition is not sufficient even for the weaker mode of authority, though, because the second-order desire may itself be lacking in authority. This raises two issues, one technical and one fundamental. The technical issue is that one might imagine another, conflicting second-order desire that would challenge the authority of the first one. Frankfurt argued that this problem would be resolved if the agent were to have a third-order desire for one of the second-order desires to be effective, and no third-order desires to back up the other second-order desire. As long as the agent is coherent at the highest reflexive order of desiring, freedom of the will would be established. Thus, a generalized formulation suggests itself: in order for a first-order desire to express what the agent really wants, there must be a desire of the n-th order to back it up, such that no other desires of orders higher than or equal to n are in conflict with it.

But this solution does not address the fundamental issue, which is whether a higher- or highest-order desire explains agential authority any better than a first-order desire does (Watson, 1975/2004b). Perhaps Frankfurt is right that in order to really want to do ϕ one must reflect on the desire to do ϕ and then really want that desire to be effective. But if we cannot reduce "really wanting to ϕ" to "desiring to ϕ," then it seems neither can we reduce "really wanting the desire to ϕ to be effective" to "desiring the desire to ϕ to be effective." I call this issue fundamental because it forces a choice between fundamentally different approaches to agential authority. One approach, taken by Gary Watson and Charles Taylor, is to reject the naturalistic, Humean framework according to which all motivation, even in cases of strong agency, must be explained in terms of desires, i.e. in terms of non-cognitive, affective states which are "given" as a matter of empirical, natural fact. But Frankfurt's background metaphysics is clearly naturalistic, and he seems to want to keep his account as close to the Humean theory of motivation as he can.

The approach that Frankfurt initially adopted instead was to retain the hierarchical desire structure, but with an added element of *decision* by the agent about which desires to identify with (1976/1988c, pp. 67–68; 1987/1988d, pp. 170–176). Thus, whereas the desire hierarchy is an essentially passive structure, the relation of identification between the agent and that structure is now understood as something actively performed by the agent himself. When the agent identifies with the higher-order desire to be motivated by a certain first-order desire, he makes a "decisive

commitment" (1987/1988d, p. 167) to act upon that first-order desire. Let us call such an act of identification a "volitional judgment." In his early writings, Frankfurt seems to be suggesting that such a judgment *constitutes* the agential authority of the desire that the judgment endorses. Thus, whether a desire expresses what the agent really wants is entirely up to the agent himself—a way of thinking that seems to reflect a certain existentialist intuition along the lines of the early Sartre. Or perhaps we should call this a *noncognitivist* view of volitional judgment, as it understands the judgment to express the authority directly, without the possibility for the judgment to get it wrong. In any case, Frankfurt soon concluded that this approach was misguided:

> To be sure, a person may attempt to resolve this ambivalence by deciding to adhere unequivocally to one of his alternatives rather than to the other; and he may believe that in thus making up his mind he has eliminated the division in his will and become wholehearted. Whether such changes have actually occurred, however, is another matter. When the chips are down he may discover that he is not, after all, decisively moved by the preference or motive he supposed he had adopted. (1992/1999a, p. 101)

As critics of Sartre have pointed out before (e.g. Taylor, 1976/1982), sometimes it is not up to us at all to decide to care about something and not care about something else. We *cannot help* caring about certain things, and neither are we infallible in our volitional judgments: sometimes we do not understand very well what it is that we really want. We can get the volitional reality within ourselves wrong. Thus, Frankfurt is forced to adopt a different approach, an approach that is cognitivist about volitional judgment. The challenge for this approach is now threefold: first, it must explain agential authority in a way that does not make it as arbitrary and contingent as first-, second-, or higher-order desires; second, it must do this within a naturalistic framework that preserves a broadly Humean intuition about motivation; and third, it must do this in such a way that we can explain what it means for an agent to be mistaken in his volitional judgments.

8.2.3 The Theory of Love

So what does the new account look like? Hierarchical desire structures still play a role as necessary conditions, and Frankfurt now argues that whereas free will merely requires a *synchronic* coherence of desires, caring requires a *diachronic* coherence. Thus, in order for an agent to care about P, he must not merely desire that he be motivated by a desire for P, but also that he will remain so motivated in the future. However, Frankfurt does not claim that the synchronic and diachronic structural requirements which he sketches are sufficient conditions for free will or caring. In fact, he now remains officially *neutral* about the question of whether any structural requirement might be sufficient to reduce agential authority to a complex of desires:

> There are significant relationships, of course, between wanting things and caring about them. Indeed, the notion of caring is *in large part* constructed out of the notion of desire. Caring about something *may be*, in the end, nothing but a complex mode of wanting it. (2004, p. 11, my italics).

Note that this is quite a substantial matter for a theory to remain neutral about. If caring is a complex mode of desiring, then the first part of the challenge becomes pressing: why are the relevant desires more authoritative than any other desires? On the other hand, if caring cannot be reduced to a complex of desires, then it is the second part of the challenge that we must worry about: how caring can remain a matter of empirical fact within a naturalistic, broadly Humean framework. Although Frankfurt speculates that an answer to the first lemma of this dilemma may be possible, his own proposal is best understood as a response to the second lemma. Rather than offering a sufficient explanation in terms of desires, Frankfurt attempts to explain the nature of caring by accounting for a special type of caring, of which other forms of caring are derived. This is the disinterested notion of caring about something for its own sake, which Frankfurt calls *love*.

Now love is probably a very good candidate for Wittgenstein's idea of a family resemblance concept: there are many different varieties of love, and each variety will resemble the others in certain respects, but there may be no central features common to all varieties that explain our correct usage of the word "love" in ordinary language. Not every form of love is a form of caring, for example. If someone says that he "loves ice cream," then

that does not necessarily mean that he cares about it. And when a student tells you that she is "in love" with her teacher, she might be simply talking about a crush which is not very important to her at all. So what sort of love does, essentially, involve caring? Frankfurt's paradigm examples of facts about love that constitute agential authority are the fact that almost all parents love their children and the fact that most of us "love living." These examples are also meant to illustrate the idea that facts about love are natural facts: "The basis for our confidence in caring about our children and our lives is that, in virtue of necessities that are biologically embedded in our nature, we love our children and we love living" (2004, p. 29–30).

But what is it about their embeddedness in our nature that gives these attitudes of love their authority? After all, the dispositions to be jealous, or to seek revenge, or to humiliate others in certain situations, may also be biologically embedded in the nature of our species, but it is not obvious that that gives them agential authority. In many cases we might not want to assign any authority to these dispositions at all. In fact, our disposition to *fall in love* probably has an important species-wide biological basis, but as mentioned above, being in love does not always involve the sort of love that carries agential authority. Furthermore, not everything that we love in the sense that does carry authority needs to have a species-wide basis. It may instead be rooted in the particular nature of the individual: I really love being a philosopher, but there are others who would hate it. That I love philosophy says first of all something about me, not about human nature. And Frankfurt does not deny this:

> Needless to say, many of our volitional necessities and final ends are far from universal. The fact that I care about various specific individuals, groups, and ways of doing things is not a function simply of generic human nature. It arises from my particular makeup and experience. Some of the things that I happen to love are also loved by others; but some of my loves are shared only by, at most, a small number of people. (2006, p. 48)

By saying this, I do not mean to retract my sympathy for the idea that there might be features common to all human beings in virtue of which we will ultimately turn out to care about the same things in a manner that could resolve our moral disputes. My point is rather that in order to make this idea plausible, one must *first* come up with an inner reality theory

that explains which type of features of an individual agent determine what that agent really cares about. Once we have such a theory, we can *then* attempt to show that the features of different human individuals will exhibit similarities in such a way that in moral cases, we will turn out to care about the same things. If we would skip the first step and merely claim that feature *F* explains what we all love, in the authoritative sense, because it is so deeply entrenched in our biological nature, then we would be guilty of a kind of naturalistic fallacy.

That does not mean that parental love is not a good paradigm case of the authoritative type of love. A proper conceptual analysis of the relevant sense of authority will probably have to account for the fact that parental love is, at least in most cases, authoritative. Furthermore, the case of loving philosophy is not meant as a counterexample to the idea that the type of love that underlies practical normativity may exhibit similarity among all human beings. After all, the view that the ideal selves of human beings shall end up having similar desires will have the same conceptual resources at its disposal to handle individual differences in taste that *nonrelationalist* dispositionalists have at their disposal, which we discussed in sections 1.3.2 and 5.1.2. Thus, all human beings may turn out to 'love' the state of affairs in which those of us who love philosophy may practice philosophy. But what the example does show is that not all matters of 'taste' in this sense—i.e., the sense of there being no practical conflict when tastes differ—are like the preference for tiramisu which only exhibits authority in the weaker sense. Instead, some of our individual preferences really seem to have authority in the strong sense: I really hope I would not suddenly lose my desire to do philosophy, for example. Therefore, we must begin with a theory of authoritative attitudes at the individual level.

Now, the most important criterion that Frankfurt seems to give in order to distinguish, at the individual level, the love that carries authority from other types of love, is that the authoritative type of love involves treating certain things as reasons for action:

> Loving someone or something essentially means or consists in, among other things, taking its interests as reasons for acting to serve those interests. Love is itself, for the lover, a source of reasons. It creates the reasons by which his acts of loving concern and devotion are inspired. (2004, p. 37)

However, this passage seems hard to reconcile with the whole cognitivist notion that we are trying to account for: that of being mistaken about one's reasons. In fact, from the cognitivist perspective, the second and third sentence of this passage would seem to contradict the first sentence. If what we happen to treat as reasons might not really be our reasons, and our real reasons reflect what we love, then what we love cannot be explained solely in terms of what we happen to treat as reasons. A similar criticism was formulated by Michael Bratman:

> Love is a source of reasons, according to Frankfurt. Frankfurt associates this claim with the idea that love necessarily involves counting or treating certain things as reasons for action. I agree with this connection between love and treating as a reason. But this does raise the question of whether what we love could be so bad—indeed, in other work Frankfurt has specifically noted the possibility of wholeheartedly loving "what is bad, or what is evil"—that, though we are thereby set to *treat* it as a reason, it is not a reason. (Bratman, 2006, p. 81–82)

Bratman's criticism already incorporates his own preferred way out of this predicament: to understand the notion of love in terms of what we treat as reasons, while understanding the idea of something really being a reason with reference to what is good or bad in a manner that is at least to some extent independent from what we love, such that we can love the bad. Actually, his view is more complex, as Bratman allows volitional judgments to go beyond, and sometimes even against, the agent's own judgments of what is valuable or normative, in a manner that is incompatible with a volitional account of practical normativity.[2] Let us however first see whether Frankfurt's account might offer ways to counter the aforementioned critique. In the end I will argue that his account cannot overcome this problem, but I will also disagree with Bratman, because I think a different volitional inner reality theory can be devised that will not fall prey to his argument.

8.3 Getting It Wrong

Despite the remark about love involving treating things as reasons, Frankfurt remains clearly committed to the idea of the opacity of love:

[2] I return to these claims in section 10.5.5.

> In addition to the fact that our understanding of the things we love may require correction, there is also the fact that we often do not understand ourselves very well. It is not easy for people to know what they really care about or what they truly love. Our motives and dispositions are notoriously uncertain and opaque, and we often get ourselves wrong. (Frankfurt, 2006, p. 49)

In the light of this passage, the alternative interpretation that we might give of the statement about treating things as reasons would be as follows: what is defining of the type of love that carries authority, as opposed to other types of love, is that such love involves our treating certain things as reasons *when we get it right*. Suppose that someone realizes that he is in love with his best friend's wife. Depending on what sort of love it is, he may or may not treat it as a reason to act in certain ways. If he knows that he merely has a crush on her, then he may not attach much significance to it, and simply enjoy or suffer the experience (depending on how bad a case it is) without any further action. Instead, if he knows that this is the real deal, then he may treat it as a reason to confess his feelings, perhaps even propose a relationship, and face the drama.

But this interpretation would make the theory of love completely empty. If an attitude L towards P is authoritative, then upon coming to believe that he has attitude L towards P, an agent will thereby acquire a self-adopted reason to promote, support, or approve of P. To say that authoritative love differs from other sorts of love in the sense that knowing what he loves in this sense will make the agent treat things as reasons is simply to repeat what it means for an attitude to be authoritative. It does not give us a substantial account of the attitude of love that would fulfill this conceptual role. In particular, even if it is consistent with the opacity of love, then it does not yet explain it: what sort of facts does this love consist in, what does it mean to get them wrong, why are they hard to get right, and how is it nevertheless possible to get them right in the end?

At some points Frankfurt seems to take a coherentist line in order to address these worries: given that we care about certain things, there are certain other things that we should also care about in virtue of their relations to the things we already cared about. This gives us means-end rationality and the sort of non-instrumental considerations that come 'for free' with the attribution of self-governing agency, as discussed in sections 3.3.2, 3.4.1, and 7.1.2. However, to suppose that this is the *only* source of

disconfirmation would simply take us back to the Principles of Reason View. It does not truly give us an inner opacity as a matter of empirical, natural fact about the individual. To be sure, Frankfurt does not advocate a 'full' opacity, so to speak: we can never be completely and utterly mistaken about what we love, or it would never be possible to improve our knowledge about it:

> If he [the agent] attempts to suspend all of his convictions, and to adopt a stance that is conscientiously neutral and uncommitted, he cannot even begin to inquire methodically into what it would be reasonable for him to care about. No one can pull himself up by his own bootstraps. (2006, p. 23–24)

But this insight, by itself, establishes not much more than that which many philosophers hold to be true for knowledge generally: that we cannot justify or improve our beliefs without assuming that, as Davidson put it, 'most of our beliefs are true.' In his commentary on Frankfurt, Michael Bratman compares this point to Otto Neurath's famous metaphor of scientific inquiry as the making of improvements to a boat while at sea: one can replace certain planks while others are left in place, but one cannot replace them all at once. Nevertheless, this does not mean that all disconfirmations in empirical science are made by testing previous beliefs against principles of reason, or it would not be empirical science. Therefore, the Neurath/Davidson intuition about belief revision does not preclude us from looking for disconfirmations that reveal a deeper opacity of our attitudes than considerations of mere coherence would establish.

In section 3.2.1 I have already discussed a similar comparison between continuity in empirical science and continuity in matters of practical deliberation. But there I argued that the Internal Reasons View actually left *less* room for discontinuity in matters of practical reason compared to theoretical empirical reason. However, if the Inner Reality Thesis is true, then this comparison may involve another dis-analogy, depending on how we understand Williams's notion of the "subjective motivational set." Because if that set includes the *opaque* attitudes, then there is no such discontinuity when efficient motivation changes upon discovery of those attitudes. I will return to this point in the next chapter.[3]

For now, let us simply note that if Frankfurt is to make his inner reality theory work, then he must explain what the opacity of love consists in,

[3]See section 9.4.

and how disconfirmation regarding what we love is therefore possible, *beyond* considerations of internal coherence. In his essay "Getting It Right" (2006, ch. 2), Frankfurt makes two suggestions: the first is an appeal to his notion of "volitional necessity," the second involves a concept of the "unthinkable." I will now discuss these in turn.

8.3.1 Volitional Necessity

The paragraph quoted above, in which Frankfurt affirms the opacity of love, continues as follows:

> [. . .] Our motives and our dispositions are notoriously uncertain and opaque, and we often get ourselves wrong. It is hard to be sure what we can bring ourselves to do, or how we will behave when the chips are down. The will is a thing as real as any reality outside us. The truth about it does not depend on what we think it is, or upon what we wish it were. (2006, pp. 49–50)

This passage reveals an important similarity to the Affective Response View that I have proposed, but also an important difference. The similarity is that Frankfurt, too, wants to explain disconfirmation in terms of unexpected future affective experiences: experiences that surprise us in the light of what we intended, given our deliberation upon our awareness of our attitudes so far. Furthermore, it seems that Frankfurt agrees that the possibility of being surprised by ourselves in this manner is what we need in order to account for an inner reality that is "as real as any reality outside us."

The big difference, however, is that on Frankfurt's suggestion we are going to have this experience just *before* we were going to act, with the result that we cannot bring ourselves to act as we had planned to:

> Someone who is bound by volitional necessity is unable to form a determined and effective intention—regardless of what motives and reasons he may have for doing so—to perform (or to refrain from performing) that action that is at issue. If he undertakes an attempt to perform it, he discovers that he simply cannot bring himself to carry the attempt all the way through. (2004, p. 46)

By contrast, on the Affective Response View it is possible that we *do* act contrary to our normative will, only to find out about this *afterward*, in response to the intended consequences of the act. This is an important difference, because it means that the opacity that Frankfurt can deliver on the basis of the suggestion above is merely the opacity of our disposition to act "when the chips are down." Thus, no self-adopted reasons that motivated any of the murders, rapes, or acts of enslavement and racial discrimination that *actually* happened in human history could be disconfirmed in this manner. But that makes the proposal entirely unsatisfactory. Consider again our example of the ex-racist bus driver: he came to change his mind *after* he had been discriminating against black passengers for years. Or consider the example of the woman who disconfirmed her wish to spend her life with her husband just *after* she married him and gave up her own house.

Furthermore, as I already argued in section 7.4.3, it is not even clear why, when it *does* happen that we cannot bring ourselves to do something, this experience could tell us, by itself, whether it is volitional necessity that's preventing us to act, or rather weakness of will. We may be able to categorize familiar examples, such as that of Luther proclaiming that he could do no other, as cases where strength rather than weakness of will is constraining the possibilities for action, but we do this on the basis of all sorts of background information that we have about such cases—for example, that Luther's act reflected his elaborate deliberations on the subject, and not just his experience of will when the 'chips were down.' By contrast, consider Frankfurt's example of the mother who has decided to give up her child, but then finds herself unable to do so when the time comes. It may be that her deliberations were indeed misguided, in which case her experience of being unable to go through with her plan would indeed reflect her inner reality. However, a similar example is possible in which the deliberations were sound and the failure to comply was due to weakness of will, and Frankfurt simply assumes that volitional necessity would allow us to discriminate between these cases introspectively at the time before the act.[4] But why should we believe this?

[4]Gary Watson discusses yet another variation on the example, in which the mother does succeed in parting with her child, but only after "severe volitional difficulty" (2002/2004d, pp. 119–121). In Watson's view, it is rather the *presence* of such a struggle, and not so much it's *outcome*, that may tell us something about how much the mother cares about keeping the child independently of her having endorsed a plan to give it up for adoption. Therefore, if we insist that her success in executing the plan does imply that keeping the child was not a volitional

It might be that the 'typical' phenomenology of weakness of will differs from the 'typical' experience of volitional necessity, but even then these sort of experiences would be hardly infallible. Of course, the same goes for the experiences that may disconfirm our practical judgments according to the Affective Response View, but as I argued in the previous chapter, the whole approach behind that view is based on the idea that, in principle, disconfirmations are never final and certain. But again, this sort of infallibility requires that we are able, in principle, to explain how further disconfirmation, after the fact of the act, would be possible. Which brings us back at the first, and most important, point of criticism: that volitional necessity would fail to explain disconfirmation of judgments in favor of acts that we already performed.

8.3.2 Unthinkability

So what are we to say about rapists, murderers and mass-murderers? Frankfurt does not want to argue that their error lies in a violation of formal logic. Furthermore, he cannot argue that their self-adopted reasons are disconfirmed by volitional necessity, because murderers *can* bring themselves to kill. However, Frankfurt does seem inclined to say that something is wrong with them, at least in extreme cases. Before we move to real-life cases, let us first consider a thought experiment from David Hume that Frankfurt discusses: that "'tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger." Here is what Frankfurt has to say about it:

> Now it is true that this preference involves no purely logical mistake. So far as logic alone is concerned, it is unobjectionable. Someone who chooses to protect his finger from a trivial injury at the cost of unlimited destruction elsewhere is not thereby guilty of any contradiction or faulty inference. In this purely formal sense of rationality, his choice is not at all irrational.
>
> But what would we say of someone who made that choice? We would say he must be *crazy*. In other words, despite the unassailability of his preference on logical grounds, we would consider both it and him to be wildly irrational. Caring more

necessity for her, then "we are taking her volitional activity to be fundamentally a matter of what she stands for or endorses, rather than what she cares about in some independent sense" (p. 120).

> about a scratched finger than about "destruction of the whole world" is not just an unappealing personal quirk. It is *lunatic*. Anybody who has that preference is *inhuman*. (2006, p. 29)

What does all this mean? Frankfurt goes on to explain that these qualifications are "literal denials that the person is a rational creature," but that this "mode of rationality" is not "exclusively defined by *a priori*, formal necessities" (p. 30). It is not entirely clear to me whether these notions of a "person" and a "rational creature" are meant to be applicable to human beings only, or whether the same verdict of irrationality would also apply to non-human agents, such as the Martian invaders. Perhaps Frankfurt means that our *concept* of volitional rationality is *a posteriori* because it is shaped by our experience of a world in which there are no Martians, and that if there had been Martians, our concept of volitional rationality would have been different. Perhaps the concept of volitional rationality cannot be *formal* because the entire conceptual network of relevant terms cannot be separated from considerations particular to our contingent nature and circumstances in a manner similar to Thomas's conception of the "relativized *a priori*" (see section 3.2.1). I am not sure, because much is left implicit in the text.

What Frankfurt does tell us, however, is how we arrive at the judgment that someone would be crazy or inhuman in this sense: by determining that *we* could not bring ourselves to do the same thing:

> An outcome from which we recoil in horror is, to him, positively attractive. The critical point has to do with possibilities: he is prepared to implement voluntarily a choice that we could not, under any circumstances, bring ourselves to make. (p. 30)

Thus, the idea is closely related to the concept of volitional necessity, except that the error in his motivation is not determined by the fact that it would be impossible for him to go through with it, but rather that it would be impossible for us to be like him. Furthermore, this impossibility for us is not revealed "when the chips are down," because we would not even *intend* to act in such a way. Rather, the volitional necessity in our case is revealed through the fact that such an act would be "unthinkable" for us (p. 31).

But why, we may now ask, would the fact that some act or preference is unthinkable for *us* be a reason to suppose that those for whom it is not

must suffer from a "defect of the will" (p. 30)? The only way in which this could be squared with the Inner Reality Thesis is if we assume that we, i.e. those of us who are not mass murderers, are *privileged* in the sense that our notion of what is unthinkable reveals something about an inner reality that we share with the mass murderers, while their notion of what is thinkable gets this reality wrong. However, without a further explanation of why we should think we are so privileged, we do not even have a reason to think that we share our inner reality with the mass murderers in the first place. But I find no such explanation in Frankfurt's work. In other words, even if these claims about volitional irrationality and the unthinkable make sense, then they do not serve to account for the opacity of the mass murderer's inner reality, but rather presuppose it.

Finally, even if we would have such a further explanation, then the account would still only handle the extreme cases: Hume's madman and the mass-murderers perhaps, but not our reasonable political opponents in disputes concerning patent law reform.[5] Nor does it handle personal cases, such as the case of the woman who discovers that she should not have married her husband. Surely, marrying him was *thinkable*, but that doesn't mean she got it right. Hence, in such cases the volitional irrationality approach does not fare any better than the volitional necessity approach. Both approaches are unable to explain the opaque nature of the facts that we got wrong when we disconfirm our practical judgments after we have acted upon them.

8.3.3 Frankfurt's Cartesian Preference for Immediacy

In view of these problems, we should wonder whether Frankfurt has really given us an *account* of the volitional opacity that his view presupposes. Rather than construing the metaphysics of an opaque attitude of love, he has tried to come up with conditions under which such an attitude turns transparent, so to speak. Thus, when the chips are down, what we truly love may reveal itself through an experience of volitional necessity. In similar fashion, the experience of the unthinkable is supposed to give us knowledge of the volitional irrationality of an action suggested or

[5] And hence, we have no reason to believe that our own political views in such 'reasonable disagreements' are volitional necessities for ourselves, an observation that has also been made by Bratman: "Although my condemnation of torturing children may well be volitionally necessary for me, my moral commitment to, say, a form of pacifism, or to political liberalism, may be wholehearted and settled without involving an incapacity to change" (2006, p. 81).

performed by someone else. Frankfurt, it seems, is still looking for those unique transparent attitudes or phenomenal experiences that, like the higher order volitions in his earlier work, could be blessed with a special authority that other attitudes and experiences lack.

Recall Watson's fundamental challenge for Frankfurt's earlier account: to explain why desires of a higher order would have a better claim to authority than first-order desires, as their occurrence may ultimately be just as contingent or arbitrary from the perspective of the self. In Frankfurt's later work, the experience of volitional necessity is supposed to have the required features that set it apart from other motivating experiences in a manner that can meet the challenge. By contrast, my own solution to this problem lies in the opposite direction. On my view, no phenomenal experience is more privileged than any other in terms of its likelihood to reflect the agent's normative will. I think that is crucially how my proposal differs from Frankfurt's account.

Why does Frankfurt keep looking for transparency? Perhaps he would be skeptical about the prospects of construing the normative will as a pattern across dispositions, as I mean to do, where none of the particular experiences that we have as a result of those dispositions offer privileged access to the will itself. However, I suspect there may be a deeper reason why Frankfurt wants to secure privileged experiences, which is that he is unwilling to give up on a type of *certitude* that such experiences would provide. Let me explain.

Frankfurt has written extensively about the epistemology of Descartes.[6] In an interview, Alex Voorhoeve asked him about how this interest might be related to his thinking about agency:

> Is Descartes' search for ideas that you can hold in the face of all attempts to doubt them paralleled in your work on the will? (Voorhoeve, 2003, p. 70)

And Frankfurt replied:

> It is in this respect: what Descartes was looking for were things that you couldn't help believing. Because it was his view that what you are in the midst of clearly and distinctly perceiving, you cannot help believing, that assent is constrained by clear and distinct perception. What I have become interested in is

[6]See esp. his (1970/2008) and various essays (e.g. 1999b, chapters 2, 3, and 4).

> not what one can't help believing but what one can't help but being moved to do; what constrains the will in action rather than what constrains the will in belief. (p. 70)

However, as Frankfurt is now defending an inner reality theory, according to which our practical views can be true or false depending on whether they get the facts about our will right, the "rather than" in this quotation seems slightly overstated. If the falsehood of our practical views can be demonstrated by an experience of volitional necessity when the chips are down, as Frankfurt argues, then clearly, volitional necessity not only constrains action, but also a special and corresponding type of belief, namely practical belief. To be sure, the mother could believe that she wanted to give up her baby, but only to the point were she tried to go through with it and found out that she couldn't. If her experience of volitional necessity is disconfirmatory, as Frankfurt wants it to be, then it constrains what she can keep believing in the light of that experience.

I am not sure that Frankfurt would want to commit himself to the view that volitional necessity is really a form of Cartesian clear and distinct introspection, which forces intimate self-knowledge upon the agent beyond the possibility of doubt. It sits uneasy with his remarks about us often not knowing ourselves very well, unless he would be willing to argue that amidst the opacity of everyday deliberation there are sometimes totally different moments of pure clarity where our transparent experiences reveal our innermost volitions with immediate certitude. Rather, I think these two opposites reveal an implicit tension in Frankfurt's thinking about agency. On the one hand, his official view postulates an opaque will. This view reflects his existential insights into the ambivalence of the human condition, and the imperfections in our means of figuring out what we're up to. But on the other hand, he has been working from the beginning upon the premise that such ambivalence and doubt are signs of unfreedom, and inspired by his philosophical heroes Descartes and Spinoza, the tendency in his work has always been to understand the resolution of doubt, and hence the existence of freedom, in terms of establishing clarity and tranquility in the mind of the agent.

Thus, in the hands of Frankfurt, notions like being "wholehearted," "fully resolved," or having "decisive commitment" all breathe a certain air of tranquility, which is in the end perhaps best understood as part of a *phenomenology* of freedom that Frankfurt finds appealing. However, the problems we have been discussing suggest that the epistemology of

immediacy behind this picture is fundamentally flawed. In the next section I will explain how it should be altered in order to build an inner reality theory that may avoid these difficulties. This will set the stage for the detailed formulation of my own proposal in the next chapter. I will also return to Frankfurt's work and discuss his concept of wholeheartedness in the next chapter, arguing that we should in fact distinguish two notions of wholeheartedness that can both be accounted for within my own framework in a manner that does not presuppose epistemic immediacy.[7]

8.4 How Should We Modify Frankfurt's Theory?

In section 8.2.3 above I have argued against Frankfurt that if he wants to be a cognitivist about reasons, then he cannot analyze the attitude of love—which is to be a source of reasons—in terms of what the agent treats as reasons. We have seen that Michael Bratman offered a similar criticism, allowing that love indeed involves treating certain things as reasons, but rejecting the idea that love could be a source of reasons at the same time. In contrast, I have followed Frankfurt in supposing that love, or at least some opaque attitude for which Frankfurt has chosen to use the term "love," must be on the 'source' side of things. But from that supposition, it does not follow that we cannot see the attitude of treating something as a reason as an attitude of love in a different sense of the term. In fact, if we want to take the inner reality theory seriously, then there are always going to be two types of attitudes involved: first, the inner attitudes that we can be correct or mistaken about, and second, the correct or mistaken attitudes that we would then have about those inner attitudes. We can think of the first as the love that is a source of reasons, and of the second as the love that involves treating things as reasons. In line with my terminology from the previous chapter, I will call the former the "normative sense" and the latter the "cognitive sense."

The same applies to *caring*, which in Frankfurt's terminology is the attitude towards either that which we love, or that which we care about for the sake of something else that we love. Be it caring instrumentally or caring for its own sake, we can now make the orthogonal distinction between caring in the normative and caring in the cognitive sense. Hence, what we care about in the normative sense is what we should care about in the cognitive sense, and what we care about in the cognitive sense is what

[7] See section 9.3.

we think we care about in the normative sense. However, instead of talking about caring, and especially about love in this manner, which is somewhat idiosyncratic on Frankfurt's part, I prefer to talk about "volitional attitudes," which is a terminology that he also uses and that does not invite confusion with the unauthoritative forms of love that we distinguished earlier on.

8.4.1 Normative vs. Cognitive Volitional Attitudes

Hence, my first proposed alteration of Frankfurt's view is that we should distinguish two types of volitional attitude: normative and cognitive volitional attitudes. The tensions in Frankfurt's approach that we discussed above are largely due to his having a single type of attitude play too many conceptual roles at the same time. Thus, whereas normative volitional attitudes are opaque, cognitive volitional attitudes are, at least by comparison, relatively transparent, since they involve our treating things as reasons, while the former provide us with the normative reasons that we actually have.

I have briefly introduced the idea of normative volitional attitudes in the previous chapter: they involve wanting something "in the normative sense," and together, all the normative volitional attitudes of an agent constitute what I have called his "normative will." Conversely, when an agent has a *cognitive* volitional attitude towards something, then we may say that he wants it "in the cognitive sense" or that it is part of his "cognitive will," which consists of all his cognitive volitional attitudes combined. The cognitive will, in other words, is the thing that has to get the normative will right: the thing that gets updated, tested, and revised through the process of volitional interpretation.

This may seem like terminological overkill: aren't cognitive volitional attitudes simply practical beliefs, on the view I am proposing? Well they are roughly the same, except that there are three reasons to nevertheless distinguish the two notions. First of all, I would prefer to say that a cognitive volitional attitude towards ϕ-ing is a *complex* attitude, which consists of both the agent's belief that he wants to ϕ in the normative sense, and whatever affective dispositions that may motivate him to ϕ that accompany that belief. This preserves the idea that attitudes like caring and love, in the sense of attitudes that direct our self-governed actions and involve our treating things as reasons, are *emotional* attitudes: they involve both thought and feeling. Of course, given our Distinctness Principle, we

should say that the affective constituents are logically contingent upon the belief: it should at least be conceptually possible for an agent to hold that same belief without the same, and perhaps with utterly different, affective dispositions.[8] Furthermore, in practice, these beliefs may represent very strong convictions and yet be accompanied by very weak motivations, or vice-versa, depending on the psychological circumstances.

The second reason is that even the concept of the belief-part of the attitude is not strictly identical to my concept of a practical belief. For as I have briefly argued in section 4.3, and will argue more extensively in chapter 10,[9] we should adopt a *pluralistic* semantics of practical belief, according to which practical beliefs are to be understood as relationalist beliefs about the agent's inner reality in some cases, and as incoherent beliefs about nonrelationalist values or standards that do not really exist in other cases, depending on the agent's 'folk meta-ethical' views (so to speak). To clarify the implication of this pluralism, let us refer to the belief-part of a cognitive volitional attitude as a "volitional belief" (this is consistent with the notion of a "volitional judgment" introduced in section 8.2.2 above: a volitional belief is adopted by making a volitional judgment). Then it follows not only that some practical beliefs are not volitional beliefs, such as the belief in a nonrelationalist commandment against homosexual acts, but also that some volitional beliefs may not be practical beliefs, such as when a homosexual man believes that even though some of the sexual acts that he decided to commit did correspond to his *volitionally normative* inner reality, they were in violation of what is *practically normative* in the light of the nonrelationalist moral commandments that he erroneously and tragically believes in.

Finally, the third reason is that we may also want to distinguish practical beliefs from volitional beliefs in order to better explain how the proposed view differs from views held by other philosophers. Recall that I divided Frankfurt's inner reality theory into two parts: the first part is the account of the nature of caring itself; the second part is the account of practical normativity in terms of the notion of caring, provided that some account of its nature can be given. A philosopher who does not agree that caring explains practical normativity may nevertheless agree that we can be

[8] Recall that motivational Humeans do not deny the existence of attitudes that have intrinsically motivational as well as cognitive implications. We merely claim that such attitudes are always complex in a manner that allows for a "two-factor analysis" (see sections 1.5 and 3.5.1).

[9] See section 10.5.3.

mistaken about what it is that we care about, and therefore adopt the framework of volitional beliefs, while denying that they are also practical beliefs.

Nevertheless, if my view is correct, then in practice our volitional and practical beliefs usually amount to the same thing. For even though belief in nonrelationalist practical normativity is widespread, the content of people's nonrelationalist practical beliefs usually corresponds to that of their volitional beliefs in terms of the thing being approved or disapproved of, for reasons I will explain in chapter 10.[10] The "perverse cases," to use Watson's term, in which they come apart, are the exception. Such cases must not be confused with the two far more common types of discrepancy that our theory is meant to account for: (a) getting the normative will wrong, where the volitional and practical beliefs are at odds with the agent's normative volitional attitudes, and (b) lacking in self-government, where the volitional and practical beliefs are at odds with the agent's resultant desires. In order to deal with these matters without over-complicating things, I will from now on talk about practical and volitional beliefs interchangeably as if they were the same thing except when it is necessary to distinguish them.

8.4.2 Varieties of Authority, Freedom, and Responsibility

Now that we have distinguished two types of volitional attitude, which of the two explains our original analysandum of agential authority, of an attitude that speaks for the agent? I think the answer is that we must also distinguish two types of agential authority. A political metaphor may be helpful here. In a democracy, an elected government has the authority to speak for the people. By analogy, that is the sort of authority that the cognitive will has. I have already been using the term "self-government" for the condition that an agent acts upon her practical beliefs, and hence (setting perverse cases aside), upon her cognitive will. Even though she may get her practical beliefs wrong, in which case what she treats as her reasons are not really her normative reasons, they would still be her "self-adopted reasons," in our terminology from chapter 1. Recall our example of the woman who decided to marry a man and move in with him, only to find out afterward that this was a mistake. Even though her decision did

[10] See sections 10.5.3–5.

not reflect her normative will, it was nevertheless *her* decision, in a manner that did reflect her reasoning and her ideas.

However, returning to our metaphor, the authority of any elected government is always derived, as it is a fallible means of the people to exercise authority over itself and achieve freedom. In a different sense, the will of the people itself is the deeper, underlying authority, but an authority that can never be expressed directly, not even during elections. The will of the people is opaque, just like the normative will, but it is nevertheless with reference to the authority of the people's will that a government must justify its own authority amongst the various movements and forces operating in a society, and in the same manner it is only as an attempt to express the normative will that the cognitive will has a claim to authority amongst an agent's various motivations.

When the people perform badly in their attempt to express their will during an election, say, and when the newly elected government proceeds to act in ways that will turn out to be wholly against what the will of the people had really been, then there is a manner in which the people were free and a manner in which they weren't. They were free in the sense that they managed to rule themselves, and were hence not oppressed. They weren't free in the sense that the end result was not what they had really wanted in the normative sense: the will of the people was not realized. In similar respects, the unhappily married woman is both free and unfree.

Note that the distinction between these two types of authority cuts across Frankfurt's distinction between agential authority in the weak sense of 'accepting' a motivation and the stronger sense of caring about remaining so motivated. As we have seen, in Frankfurt's view and terminology, it is the weaker kind that establishes "freedom of the will." In our modified account, this notion of freedom may now be disambiguated into the two types of freedom distinguished above. Nevertheless, one might also wish to say that a more substantial type of freedom is achieved only once we manage to implement the things we really care about, and there too, we can distinguish between the cognitive and the normative sense. In my own account, Frankfurt's distinction between accepting and caring will not play an important role.

What we should keep in mind, however, is our earlier observation that what Frankfurt calls "freedom of the will" in this context may not be a very plausible candidate for the sort of free will that would help justify praise and blame. As I noted in section 8.2.1 above, if an agent would only be

blameworthy for self-governed acts, then we could never blame people for acting against their better judgments. Furthermore, even philosophers who are skeptical about blameworthiness and the associated concepts of moral responsibility and free will (e.g. Pereboom, 2001, 2007; Strawson, 2010) usually subscribe to a weaker notion of accountability that does apply to some of our lackings in self-government and yet requires the agent to have willed the action in some sense that exceeds mere desiring (see esp. Pereboom, 2002, section 1). Let us call the former, blame-including notion "strong accountability" and the latter "weak accountability."

As I mentioned in section 8.2.1, Gary Watson distinguished the concept of accountability from what he called "attributability," arguing that both notions are important varieties of responsibility. The freedom that gives rise to *attributability* is what Watson calls "self-disclosure" and it is this notion that my concept of volitional interpretation is meant to make sense of. In particular, the distinction that I am proposing between the cognitive and the normative will should be understood as a further refinement of the self-disclosure side of Watson's taxonomy, by explaining how acting upon the cognitive will discloses the *views and ideas* that the agent really endorses as her own, while it is only when those views correspond to her opaque normative will that they disclose, in a 'deeper' sense, the normative source within herself that those ideas are supposed to represent. Let us say that the former type of self-disclose gives rise to "weak attributability" whereas the latter, 'deeper' variety constitutes "strong attributability."

By contrast, the freedom that gives rise to *accountability* is often understood as a kind of *control* that an agent can have over his actions. Let us refer to the psychological structure exerting this control as the "executive will." Thus, when an agent acts upon his executive will, he may or may not be acting in accordance with his cognitive will, but even if he is not, then at least he is still controlling his action in a manner that allows us to hold him accountable for it. The account of the cognitive and normative will that I am developing here can remain pretty neutral about what the nature of the executive will might be. For example, John Martin Fischer's concept of a "moderately reasons-responsive mechanism" would be one account of the executive will with which my theory about the cognitive and normative will is perfectly compatible.

Furthermore, the account of the cognitive and normative will can also remain neutral about the question whether the executive will gives rise merely to weak accountability, or whether it also justifies strong

accountability. In fact, this allows us to capture the disagreement between Fischer and Pereboom nicely, as Pereboom has claimed that he finds Fischer's notion of reasons-responsiveness a plausible candidate for the justification of weak accountability. In our terminology, we can therefore say that Fischer and Pereboom *agree* about the nature of the executive will, but *disagree* about whether the executive will allows us to justify merely weak accountability, or also strong accountability.

Our taxonomy of responsibility concepts is now as follows: we have Watson's distinction between accountability and attributability, Pereboom's further distinction between strong and weak accountability, and my further distinction between strong and weak attributability. We have seen that the normative and cognitive will explain these two forms of attributability, that the executive will is whatever control-structure that allows us to justify weak accountability, and that it is an open question whether such a structure allows us also to justify strong accountability.

For the sake of completeness, however, we should also recognize that the phrase "free will" is sometimes used in ways that do not presuppose it to be a justifier of any sort of responsibility attribution in the first place. For example, a lot has been written about whether "conscious will" might be an "illusion" (Wegner, 2002), but it may be argued that conscious will is neither necessary nor sufficient for responsibility attribution, and that it does not correspond to either the executive, cognitive, or normative will in our framework. Nevertheless, some researchers have used the term "free will" to refer to ideas involving the immediate, conscious direction of bodily movement (e.g. Libet, 1999), suggesting that at least for some people, conscious will is part of their idea of freedom.[11]

Moreover, some philosophers seem to regard certain metaphysical notions that play an important role in the debate on responsibility as meaningful concepts of free will in their own right, regardless of whether they do explain responsibility or not. Fischer, for example, argues that the ability to have done otherwise is not necessary for moral responsibility, but nevertheless a meaningful concept of control that explains what people in everyday life mean when they use the term "free will," even though Fischer

[11] In our textbook on free will, Tjeerd van de Laar and I have divided the analysandum of free will into the following three concepts (Van de Laar & Voerman, 2011): as a requirement for moral responsibility (which is another term for strong accountability and thus concerns the executive will); as a form of self-realization (which concerns the normative and cognitive will); and as the conscious direction of action (which concerns conscious will in the Wegner/Libet sense).

believes that we may not have this ability and that it is incompatible with determinism.

Summarizing, my proposal preserves Frankfurt's idea that volitional attitudes explain something that we call "freedom of the will," but only in a suitably qualified sense of the term, which should be distinguished from a myriad of other, and very prominent, free will analysanda that the theory remains neutral about.

8.4.3 The Affective Analysis of the Normative Will

As we have seen in section 8.2.3, Frankfurt has stated that "the notion of caring is in large part constructed out of the notion of desire" and that it "may be, in the end, nothing but a complex mode of wanting" (2004, p. 11). He also writes, concerning love, that:

> It is not essential to love that it be accompanied by any particular feelings or thoughts. The heart of the matter is not affective or cognitive, but strictly volitional. The necessities of love, which drive our conduct and which circumscribe our options, are necessities of the will. Their grip means that there are certain considerations by which we cannot help being moved to act, and which we cannot help counting as reasons for action. What is essential to love is just these constrained dispositions to reason and to act out of concern for the beloved. (Frankfurt, 2006, pp. 42–43)

At first glance, this passage seems to suggest the view that love cannot be analyzed as, or reduced to, a complex of desires. Despite the ways in which love is commonly associated with various desires, Frankfurt seems to be saying, the "heart of the matter" or the "essence" of love is neither cognitive nor affective, but "strictly" volitional in a *sui generis* fashion. However, he also claims that this essence consists in *dispositions to act* in certain ways. But note that in the broad sense of desiring explicated in sections 1.4 and 1.5, any disposition to perform an action ϕ under circumstances C, which is not cognitive, counts as a disposition to have a desire to ϕ under C. Hence, any analysis of volitional attitudes into dispositions that are not cognitive must be an analysis of the volitional in terms of such desires by definition.

Perhaps we should conclude that when Frankfurt talks about the "affective" and "feelings" in this passage, the notion he has in mind must be

narrower than our desires in the broad sense. His reference to "feelings" suggests that his purpose is rather to distinguish the volitional from a certain type of *phenomenally* affective state, not from a motivational attitude in a dispositional sense. Perhaps he merely means to express the claim that love need not be associated with a very "passionate" or so-called "hot" phenomenology, but that it may also exclusively involve rather "dispassionate" or "cool" manners of going about one's business that reflect one's loving something strongly or firmly. In that case, the claim seems true, but also somewhat trivial, and rather uninteresting in the light of the prospects of analyzing the volitional as a complex structure of more basic states or attitudes that caring organisms share with the wanton organisms that act exclusively upon first-order desires. After all, there is no reason why all the first-order desires of the wanton should be accompanied by a 'hot' or 'passionate' phenomenology.

Maybe Frankfurt is claiming that love resists analysis in terms of phenomenal states generally, not just the passionate states. In that case, his view might be that whereas love cannot be analyzed as a complex of feelings, it may perhaps still be analyzed as a complex of desires, on the assumption that desires need not be phenomenal states themselves. Such a conception of desires has been defended by Smith, who argues that actions are sometimes explained by attributing desires with certain propositional content to agents who do not phenomenologically experience themselves to be desiring that content while acting the way they do (Smith, 1994, pp. 104–111). Instead, Smith favors an account according to which desires are themselves dispositions (pp. 111–116). However, on his view, at least, these include dispositions to have certain feelings under certain situations, not just behaviors. Thus, certain *types* of desires are essentially phenomenological, or essentially do have certain phenomenological implications. And we may wonder whether it makes sense to speak of love, caring, or really wanting something, without implying that the agent will at least be *disposed* to experience certain feelings about what he loves under the appropriate conditions.

To some extent, this issue depends on the analysis of phenomenal consciousness itself. For example, consider the view that phenomenal experience is an *epiphenomenon*, and that it is possible to conceive, without incoherence, a world without experience whatsoever in which my "zombie twin" is physically and functionally identical to me (Chalmers, 1996, 2003). Although this is certainly not the majority view amongst philosophers of

mind, it may be interesting to consider what it might mean for a zombie to love something. To be sure, my zombie twin would have a similar inner psychological structure that he might be mistaken about, albeit in a non-phenomenological sense, compared to the structure of my normative will that I might be mistaken about. However, given that zombie psychology is just as functionally rich as ours, we might therefore also attribute various 'affective' states to my zombie twin, such as jealousy, anger, lust, or curiosity, in order to explain his inner workings and behavioral dispositions, so long as we keep these states functional. Therefore, epiphenomenalism would not prevent us from wondering whether the volitional implies affective dispositions understood as functional roles. At the same time, even if a zombie were possible with a psychological structure identical to the functional structure that constitutes my will, it seems that without phenomenal experience, the things that my zombie twin and I both want would not really matter to him as they do to me. They could not really be important to him because there would not be a "him" in the relevant sense. Thus, if the phenomenal does not reduce to the functional, then we can give a weak interpretation of the affective that is purely functional and a strong interpretation of the volitional that implies phenomenality. This hardly shows that the volitional would a be less phenomenological concept than the affective. Rather, it seems that both the volitional and the affective can be approached in functional and phenomenological terms.

Furthermore, if we on the other hand consider *analytic functionalism*, the view that an analysis of phenomenal consciousness in terms of its functional role must be possible and that zombies are therefore logically impossible (e.g. Lewis, 1990; Jackson, 1994; Dennett, 1991a, 2001), then presumably an agent could not have the sort of dispositions underlying love or caring without being disposed to have phenomenal states as well. Again, it does not seem plausible to think that volitional states do not have phenomenal implications when affective states do.

Another problem for this phenomenological interpretation of Frankfurt's claims about the irreducibility of the volitional to the affective is that it seems to imply that Frankfurt's concept of volitional necessity cannot have a phenomenology either. However, without an experience of volitional necessity from the inside, as it were, distinguishing a disconfirmatory event of volitional necessity from a non-disconfirmatory event of incontinence would no longer just be difficult, as I have argued in section 8.3.1, but outright impossible to do on the basis of the event itself. This

leaves Frankfurt once again with the problem of explaining the opacity of the normative will.

Finally, yet another interpretation of the above passage would be that Frankfurt merely meant to say that the structure of love never reduces to any *particular* affective type of attitude. Thus, no single type of affect can by itself explain or establish the fact that an agent loves something. Being jealous does not entail love, feeling romantic does not entail love, and so on. But whoever said it would? As I see it, the whole idea behind analyzing the volitional as a complex affective structure is that this would involve *multiple* affective attitudes, and it seems plausible that these will also be of different affective types. Hence, the complex structure of love may lead to jealousy on one occasion, and to peace of mind on another. Strangely enough, it seems Frankfurt himself who comes closest to the view that a single type of experience might entail volitional love. As we have seen in the previous section, the type of immediacy that Frankfurt seems to associate with volitional necessity threatens to turn it into such a privileged experience, and implausibly so.

In order to leave some of the confusion behind, let us call "affective" any state or attitude that (a) may be understood as a motivational disposition by attributing a desire in the broad sense and (b) is at least *potentially* phenomenal in the sense that its motivational force may be experienced phenomenally under certain conditions. Note that this does not mean that its propositional content must be introspectively accessible. It merely means that under the appropriate conditions, there will be something that it is like for the agent to be motivated by this state. This is consistent with Smith's dispositional understanding of desires. It follows that thermostats and chess computers do not have affective states even if we may attribute desires to them, but also that all desires that enter into our practical deliberations can be safely assumed to involve affective states in this sense.

Given this conception, I think the opacity of the normative will must be explained by some analysis of normative volitional attitudes in terms of complex affective dispositions. For let us assume the opposite. Assume, for a second, that some normative volitional attitude could not, in principle, be explained in terms of the various ways in which the agent might be motivated, or might experience motivational pressure phenomenally, under various conditions. Furthermore, since the normative will is opaque, the agent is also unable to experience the content of this attitude introspectively. Then how could the reality of this attitude possibly be a matter of empirical

fact? What empirical reasons could there be to attribute it to him as part of his normative will? What sort of biological or psychological properties could imply that it is a fact that he wants this in the normative sense, even though this will never translate into an effective and/or experienced desire? Given our broad understanding of the notion of desire, any non-cognitive mode of 'wanting' something that could not be analyzed in terms of desires would amount to unwarranted mysterianism.

Or assume that the normative will could be analyzed as a structure of dispositions that had no phenomenal implications whatsoever. Even if the agent has phenomenal consciousness, there seems no reason why such a structure should mean anything positive to him, in the same sense in which it could not really mean anything to the unconscious zombie or chess computer. On the contrary, from the perspective of his conscious self, such a structure of behavioral dispositions whose motivational force he is not able to experience phenomenally, if such a thing would even be remotely possible, could only appear as something alien that is getting in his way. In order for his will to mean anything to him as a conscious agent, we must understand it in terms of dispositions that are affective. I will attempt to provide such an analysis, and one that I believe explains the opacity of the normative will, in the next chapter.

8.5 Modes of Normativity

In section 8.2 I divided Frankfurt's inner reality theory into two parts: the first part was his account of the nature of volitional attitudes, the second his analysis of practical normativity in terms of such attitudes. In section 8.3 I have criticized the first part, and in section 8.4 I have explained how it should be modified, setting the stage for my own account of the nature of the normative will in chapter 9. Let us now take a brief critical look at the second part of Frankfurt's theory, in preparation for my own analysis of practical judgments in chapter 10.

8.5.1 Moral, Volitional, and Practical Normativity

Consider the following example. John believes that animals are being treated unethically in the factory farming industry. He also believes that this gives him a reason not to consume meat that was produced by this industry. Instead, he tries to buy meat that comes from so-called 'organic'

farms when he can, or to opt for vegetarian alternatives when they are available. However, John is also not a fundamentalist about this and his policy allows considerable leeway. He is happy to eat whatever meat is served when he is a guest somewhere, and he allows himself to buy the 'wrong' meat when there is no 'organic' alternative available and he really feels like eating that type of meat. Furthermore, let me stipulate that these occasions do not involve weakness of will, or at least not all of them. In other words, John has made the *volitional judgment* that he really wants to allow himself this sort of leeway.

Suppose he has gotten this judgment right, would it then follow that he also has a *normative* reason to buy the 'wrong' meat on the occasions where he really wants to do so? Does the mere fact that he wants it in this volitional sense make it right for him to do so? Several people, I believe, would be inclined to feel otherwise. Whether we really want something is one thing, they will say, and whether it is right or wrong is something else. There are many ways to unpack this intuition, but what they all seem to have in common is that they contrast the personal and perhaps partial ends of the *agent* with the impartial and perhaps impersonal ends of *morality*.

For example, as we have seen in section 1.4.1, according to the type of externalist who would reject the Authority Principle, morality is the object of our practical judgments, but self-government does not require that we act upon these judgments. A proponent of this view might argue that John is self-governing when he follows his volitional judgment to eat the wrong meat on occasion, while even John himself might at the same time, and without contradiction, hold the practical belief that he should never eat that kind of meat.

Or consider the external reasons view, which we discussed in chapter 2. If an external reasons theorist were to believe that there is a resultant reason, in the external sense, for all of us never to eat any meat produced by the unethical treatment of animals, then he may still allow that the taste for meat could be so dominant in John's subjective motivational set that even after extensive reflection upon its unethical production, John might remain motivated to buy the 'unethical' meat when the 'ethical' meat is not available, which would constitute a resultant reason in the *internal* sense. The external reasons theorist might say that the truth about John's internal reason establishes what he really wants, whereas the truth about his external reason establishes what is right.

But even internalists who accept the Authority Principle and subscribe

to a form of dispositionalism contrary to the external reasons view may still distinguish between morality and volition, and hold that when the two are in conflict, we should do what morality requires, rather than what we personally want. The type-III dispositionalist, for example, could say that the volitional is the source of our P-reasons while morality establishes our NP-reasons (see section 6.3.4). And the type-II dispositionalist might want to argue that, whereas morality is nonrelationalist and *a priori*, there nevertheless also is a relationalist and empirical inner reality of the sort that Frankfurt and I attempt to describe, in such a way that the content of the latter could conflict with the former. Finally, even a proponent of the Inner Reality Thesis might argue that the volitional is not practically normative, by claiming that the inner reality that gives us normative reasons should not be analyzed in volitional terms like "wanting," "loving," or "caring."

Now I think that all these views are implausible, for reasons that I have, to a large extent, already discussed. However, the reason why I have summarized these views here is that Frankfurt also makes a distinction between the moral and the volitional. In fact, he makes two distinctions, one between volitional judgments and moral judgments, and another between volitional judgments and value judgments, that may both seem similar to the sort of distinctions mentioned above:

> [Y]our most fundamental problem is not to understand how to identify what is valuable. Nor is it to discover what the principles of morality demand, forbid, and permit. You are concerned with how to make specific concrete decisions about what to aim at and how to behave. Neither judgments of value in general nor moral judgments in particular can settle this for you. (Frankfurt, 2006, p. 27)

However, there is a big difference: whereas the aforementioned ways of unpacking the division all took morality to be the practically normative phenomenon, giving it a kind of precedence over the volitional, Frankfurt instead argues that the volitional is practically normative in the most fundamental sense, *rather than* the requirements of morality or the facts about value. Let us consider morality first:

> It is often presumed that the demands of morality are inherently preemptive—in other words, that they must always be accorded an overriding precedence over all other interests and

> claims. This strikes me as implausible. [...] Morality is most particularly concerned with how our attitudes and our actions should take into account the needs, the desires, and the entitlements of other people. Now why must *that* be regarded as being, without exception, the most compelling thing in our lives? (Frankfurt, 2004, p. 7)

Does this mean that on Frankfurt's view we might also claim that even though John gets his volitional judgment right that he wants to be able to eat the 'wrong' meat now and then, it would still be morally wrong for him to do so? To be sure, Frankfurt only talks about the interests of other people, not other animals, but presumably animals count as well (and otherwise we could construct an example involving the unethical treatment of workers in a factory, say). So I suppose this distinction between what John wants and what is moral can be maintained on Frankfurt's view.

Note that Frankfurt uses the term "morality" in a narrow sense, which only involves the reasons we have for taking other people's interests into account. Therefore, morality is only one "mode of normativity" amongst many others, including those involving "nonmoral ideals" such as "aesthetic, cultural or religious ideals" (p. 8). By itself, that is pretty much in line with my own usage of the term in section 1.3.2. But the way in which Frankfurt seems to be construing these modes of normativity is that each of these only provides reasons *in the non-resultant sense*. Hence, after each mode has contributed its reasons, there is still the general normative question of how to balance them in order to determine one's reasons in the resultant, all-things-considered sense. And this is, apparently, not a moral question even when moral considerations are among those at stake. Thus, if John gets his volitional judgment right then Frankfurt only allows you to say that John's policy is not moral because it is an all-things-considered policy and Frankfurt simply does not wish to call the all-things-considered stuff "moral."

Frankfurt's terminology seems highly idiosyncratic: most philosophers would say that morality is not only about what sort of things you might do in order to care for others, but also about how much time and resources you should spend on that, and how you should balance these considerations against your other ends. Frankfurt does seem to admit that morality is about balancing the needs of others against your self-interest, but he argues that many of the other modes of normativity, such as religious or aesthetic ideals, are not self-interested either. However, insofar these ideals need to

be balanced against moral ideals, I think most philosophers would still call that balancing a moral issue. And insofar various nonmoral ideals need to be balanced against each other, morality is not being outweighed, but simply neutral or irrelevant.

Hence, because of how he has construed the term, morality is not an external practical normativity independent of our volitional attitudes, but merely a non-resultant subdivision of practical normativity, which as a whole is still understood in terms of the volitional according to the Inner Reality Thesis.

8.5.2 Value Judgements and Volitional Judgements

Let us now take a look at the second distinction that Frankfurt makes, between value judgments and volitional judgments:

> Caring about something differs not only from wanting it, and from wanting it more than other things. It differs also from taking it to be intrinsically valuable. Even if a person believes that something has considerable intrinsic value, he may not regard it as important to himself. (Frankfurt, 2004, p. 12)

And the point is not that the agent simply judges that the thing of value would be better pursued by someone else. Rather, the agent may be perfectly happy to leave the value in question unresponded to:

> Something that we recognize as having intrinsic value (a life devoted to profound meditation, perhaps, or to courageous feats of knight errantry) may nevertheless fail to attract us. Moreover, it may be a matter of complete indifference to us whether anyone at all is interested in promoting or achieving it. We can easily think of many things that might well be worth having or worth doing for their own sakes, but with regard to which we consider it entirely acceptable that no one is especially drawn to them and that they are never actually pursued. (Frankfurt, 2004, p. 13)

I am puzzled about what these intrinsic values are supposed to be, on Frankfurt's view, or what the purpose would be for us to go about recognizing them. Frankfurt seems to have some sort of modal claim in mind: a value is something that it is possible for a reasonable, self-governing agent

to love, even if nobody actually does, or has to. But if this would be mere conceptual possibility, then making such value judgments in everyday life seems otiose (except perhaps for philosophers who are trying to figure out what is constitutive of self-governing agency). And if it is a more restrictive mode of possibility, then Frankfurt should say something about that, which he doesn't.

Perhaps it is a kind of dialectical move, where he wants to grant his opponents that they may have good reasons for believing in the existence of mind-independent values, while restricting his own defense to the claim that such values would not imply that we should care about them. But I think this is not a prudent move, dialectically speaking. On the contrary, the fact that it does *not* involve a metaphysics of mind-independent values should be a major selling point of the Inner Reality Thesis. As Korsgaard put it, "it seems a shame to go to all the trouble to deny normative realism about values and then espouse a kind of nonnormative realism about them after all" (2006, p. 71).

Nevertheless, the most important thing at this juncture is our observation that Frankfurt does not consider morality or intrinsic value, insofar as these may come apart from the volitional, as practically normative. In fact, his analysis of practical normativity in terms of our volitional inner reality seems that of straightforward identification: practical judgments are volitional judgments. I have already indicated that this claim is too quick, because we need to account for erroneous nonrelationalist judgments as well. We will return to this matter in chapter 10.[12]

[12]See section 10.5.

9 *The Nature of the Normative Will*

In section 4.4 I formulated two problems for the type-1 dispositionalist. The first was to account for the idea that some of our contingent desires have the authority to discredit others. The second was to give a plausible relationalist account of the intuitions behind the Intersubjectivity Principle. In chapter 8 we have seen how a volitional inner reality theory might solve the first problem. On such a theory, some of my desires are in accordance to my volitional attitudes—the attitudes I have towards the things I really want as a person—whereas others are not. However, I have argued that Frankfurt's own volitional theory fails to explain the opacity of *normative* volitional attitudes: first, because he does not properly distinguish them from *cognitive* volitional attitudes; and second, because he ultimately fails to provide an analysis of them as structures of more basic and *unprivileged* affective attitudes.

In this chapter I shall attempt to formulate a volitional inner reality theory that does satisfy these requirements: an account of the nature of the normative will, in accordance with the Affective Response View of disconfirmation from chapter 7, that will address the first problem of type-1 dispositionalism. I return to the second problem in the next chapter.

9.1 The Affective Pattern View

According to the view I want to propose, volitional interpretation is essentially a form of *pattern recognition*, and the normative will is the pattern that the deliberating agent is trying to recognize: a pattern across her affective dispositions. Let us call this the "Affective Pattern View." I am borrowing the general idea of a pattern as an ontological notion from Daniel Dennett, who has argued that beliefs and desires are present as patterns in our behavior (1991b). However, my purpose is to apply the notion at a different level, which does not presuppose Dennett's purely behavior-based ontology of intentional states. Let me explain.

As an example, Dennett discusses visual images, represented as bitmaps of square pixels, that depict a 'bar code' of adjacent larger black and white squares, except that there are also some white pixels within the black squares and black pixels within the white ones (1991b, p. 31, figure 1). What is important is that despite the fact that such an image is strictly nonidentical to what a 'pure' bar code image would have looked like, we can nevertheless *recognize* the bar code and *reject* the deviant pixels as *noise*. Furthermore, it seems that we would have *missed* something about the 'impure' image if we had not recognized the bar-code pattern in it, and therefore, that the presence of this pattern is a *matter of fact* about the 'impure' image.

Now recall Harry Frankfurt's terminology of desires that are "external to the person." In a nutshell, my suggestion is that such desires are like the 'noise' in the image. When we get it right, we *reject* such desires as not being part of our will, just like we 'reject' the deviant pixels in the image as noise. Conversely, I propose that we may recognize how the other desires that we have, the ones that are "internal to the person," jointly constitute a pattern of what we really want, just like the pixels in the image that contribute to the bar code pattern. We try to form a picture of our normative will by considering the totality of affective experiences from different situations together. Even though our desires pull us into different directions at different times, and sometimes even at the same time, we search for recurring themes and for ways in which different emotions might support each other and jointly point in some direction.

Instead, Dennett pursues a different analogy: he wants to compare the individual pixels to our individual *behaviors*, and explain our *beliefs and desires* as patterns across these behaviors, such that we can recognize the behaviors that contribute to these patterns as *actions*, while once again classifying the behaviors that do not fit into these patterns as a form of noise. This is not a perfect analogy, however, because of some issues with regard to attitudes and behavior that do not arise in the case of images and pixels. Therefore, let us first discuss how Dennett's application differs from his bar code example, and see whether these differences also apply to the application that I have in mind. After that, I will say more about how my application of this analogy differs from Dennett's.

9.1.1 Limited Data and Predictive Success

In the case of the bar code image, all the relevant data—i.e., all the pixels that make up the image—are presented to us *at the same time* and processed by our visual cortex in parallel, unlike the data pertaining to the analysis of propositional attitudes. Instead, we attribute beliefs and desires to an agent, according to Dennett's view, on the basis of our experience of that agent's behavior on different occasions through time. Rather than seeing the pattern directly, we have to *interpret* the behavior that we are currently observing in the light of our memory of the agent's past behavior, by attributing beliefs and desires from the "intentional stance" (Dennett, 1987). Adopting this stance towards an agent involves the assumption that her behavior can be made sense of in terms of rationally interrelated beliefs and desires.

Below I will say a bit more about how Dennett's intentional stance theory squares with my own views concerning belief–desire psychology and with the things I have said about motivational Humeanism earlier on in this thesis. But first let us note that at least with respect to my account of volitional interpretation, something similar applies to my own view as applies to Dennett's. As I have argued in chapter 7, we deliberate upon our ends by considering the relations between our different affective experiences on various occasions. One might say that on my view, the deliberator adopts a 'volitional stance' towards *himself*, by assuming that he has a normative will.

Furthermore, in contrast to the bar code example, Dennett's interpreter must always work with a limited set of data in order to make judgments about an underlying reality. Thus, when we look at the bar code image, we are in a sense fully informed: because we are able to see all the pixels, nothing further about the image is hidden from us. Or, to put it differently, the pixels are not just data *about* the image; they *are* the image. Instead, the observed behaviors of an agent are never fully constitutive of his mental states: they are merely the various *manifestations* of those states under the particular circumstances that the agent has so far encountered. Even in so far as intentional attitudes can be analyzed as behavioral dispositions, they would still be constituted by all the behaviors that the agent *would* have displayed under an infinite range of counterfactual circumstances that we will never actually observe.

Hence, the behaviors that Dennett's interpreter *does* observe are rather like a statistical *sample* of the reality that the interpreter is trying to get

right. The bigger the sample, the larger the chance that a pattern in that sample reflects a real underlying pattern in the dispositions of the agent. Because the intentional stance must, of course, lack any scientific rigor or statistical method, Dennett gives a somewhat informal explanation of the idea that an agent really has certain beliefs and desires in terms of our success in predicting his future behavior by attributing those beliefs and desires. If we achieve 'above chance' results, then we can explain our predictive success by assuming that we recognized a "real pattern."

Once again, a similar thing can be said with respect to my views about practical deliberation. In chapter 7 I argued that it follows from the Affective Response View of disconfirmation that we must analyze our practical judgments in terms of predictions about our future affective responses. The higher their predictive success, the more reason one has to believe that the reality of one's normative will must explain this success. Of course, there are also uncritical ways of avoiding disconfirmation, which do not count as predictive success—more on this later on. For now, however, note that we can argue for the reality of the normative will on the basis of the Affective Response View in a manner similar to Dennett's line of argument: if we were able to correctly predict our affective responses on the basis of our practical judgments, and if our success in doing so was not due to chance or to circumstances that undermined the possibility for disconfirmation in uncritical ways, then it seems we *would have missed something* about ourselves if we had failed to make those predictions.

However, we shall see below that the relation between predictive success and correct interpretation is more complex in the case of my Affective Pattern View, and therein lies an important dis-analogy between Dennett's application of the pattern concept and mine. But in order to explain this, let me first say a bit more about how the two accounts are different even insofar they *are* analogous.

9.1.2 Levels of Organization and Varieties of Attributivism

To begin with, the most obvious difference is the level of organization at which the two theories operate: whereas Dennett analyzes beliefs and desires as patterns of behavior, I propose to analyze the normative will as a pattern of desires. Hence, while desires are the *analysanda* in Dennett's account, they are the *analysans* in mine. The two views may very well be independent of each other: one might subscribe to only one of them, or

neither, or both. If one subscribes to both, then one might think of the normative will as a 'second order' pattern: a pattern of desires, which are themselves patterns of behavior.

And the theories are even further apart, because as I argued in the previous chapter, I do not think the normative will should be analyzed in terms of the desires that even chess computers and thermostats might have, which are amongst the desires that Dennett is presenting as real patterns. Instead, I mean to analyze the normative will in terms of *affective* attitudes, which I have defined as dispositions that, in addition to their motivational impact on behavior, also imply a phenomenal experience of their motivational 'pull' under at least certain circumstances (although without thereby necessarily providing introspective access to the propositional content of the attitude).

To be sure, this constraint on the normative will is not incompatible with Dennett's philosophy of mind. After all, Dennett also accepts a distinction between conscious and unconscious mental states. He thinks this distinction can be made from the intentional stance using the method of "heterophenomenology" (2003), which basically consists in taking verbal reports of phenomenal experience at face value. And he argues that the experience of consciousness that people report can be analyzed functionally from the "design stance" (1991a; 2001). If this is correct, then the Affective Pattern View of the volitional is fully compatible with Dennett's pattern view of the mental, and followers of Dennett who subscribe to both his explanations of intentionality and consciousness may view the "volitional stance" as a further extension to the "design stance" and the "intentional stance."

Nevertheless, I want to be clear that the Affective Pattern View does not *depend* on such an approach towards consciousness and intentionality. On the contrary, I myself think Dennett's theory of consciousness is wholly unsatisfactory, though for reasons that are beyond the scope of this thesis.[1] What matters for now is that the Affective Pattern View analyzes the normative will as a pattern of potentially phenomenal states, regardless of how phenomenal experience should itself be analyzed.

[1]Briefly, my view is that although phenomenal states can only be understood in the context of a functional, cognitive architecture implemented by the physical workings of the brain, no physical description of the supervenience base of phenomenal states can capture the full nature of that base on which the phenomenal character of our experiences supervenes logically. Similar views have been defended by Peirce (1891/1992a, pp. 292–293) and Russell (1927), and more recently by Strawson (2006) and Nagel (2000).

Furthermore, I am also not convinced by Dennett's account of intentionality, although I am more sympathetic to it, in particular to the idea that the *propositional content* of beliefs and desires should be understood in an attributivist manner. However, unlike Dennett, I do not think that attribution of beliefs and desires is what we usually do when we interpret each other's behavior in everyday life. As I noted in section 3.4.2, human behavior may be easier to explain in terms of attitudes that do not represent separately our understanding of the world and our goals for acting in it. Instead, attitudes which are both cognitive and motivational may not only be more psychologically realistic, but also more commonly attributed in everyday life. By contrast, I have briefly alluded to an argument that the purpose of belief–desire psychology is rather to reconstruct our attitudes in relation to our 'disinterested' concept of truth. However, an extensive defense of this idea is beyond the scope of this thesis.

Finally, let me note that even someone who would completely reject the attributivist approach towards beliefs and desires, and who would for example opt for a computationalist understanding of belief-desire psychology and a much more traditional defense of the Humean theory of motivation, may still subscribe to my analysis of the normative will as a pattern in the desires that the agent will or would have under various actual and counterfactual circumstances. Basically, my aim is to maximize the plausibility of the views I put forward in this thesis by on the one hand *defending* them in a manner informed by ideas and insights from philosophy of mind and cognitive science, while on the other hand keeping their *formulations* fairly neutral with respect to many of the disputes in those fields.

9.1.3 Motivational Character vs. Normative Will

What does it mean to say that something is a pattern across our desires? An important feature of desires is that we have many of them: some are only incidental or present during a short period of time, others are recurrent or stable during a long phase in the agent's life. We can experience multiple desires at the same time, even if they conflict, and we may note even more conflicts between the desires, or affective responses, that motivate us under different circumstances, or during different times of the day, week, or year. Therefore, one way in which we might construe a pattern *across* this multitude of attitudes towards conflicting propositions would be to

identify the 'largest' *coherent subset* of those attitudes, so to speak, where the 'size' of this subset would be measured in terms of the motivational strength of its elements.

Thus, in a greatly oversimplified picture, suppose an agent has desires d_1–d_5 towards propositions P_1–P_5 with motivational strengths m_1–m_5. Suppose that P_1 and P_2 are compatible, but that P_1 rules out P_3, P_4, and P_5. Suppose furthermore that m_1 plus m_2 make up more motivational impetus than any other combination of these desires that do not feature incompatible contents. Then $\{d_1, d_2\}$ would be the 'largest' coherent subset of the agent's desires. However, the problem is of course that we cannot individuate our affective states or dispositions in such a discrete and finite manner. Nevertheless, we might be able to imagine a structurally similar idea of the degree to which we find ourselves motivated, across different circumstances, towards our various competing and even conflicting goals. Then, we might think of the coherent subset of desire attributions which collects the largest overall amount of motivational potential as the motivational *character* of the person. Think of the person with the lazy character who usually decides to act in ways that minimize effort, even though he will sometimes act otherwise.

But this idea raises many questions. For one thing, does it not depend heavily on the type of circumstances this person will find himself in how he will be motivated most of the time? Maybe if such a person became a recruit with the Marine Corps, then he wouldn't be so lazy after all. Some philosophers[2] argue that "situationist" social psychology[3] has shown that little, if any, of our behavior is explained by "global" (i.e. situation independent) character traits, because the impact of situational differences generally outstrips that of differences between individuals. And even if we follow the more recent, and in my view more plausible, approach in social psychology which focuses on the interaction of personal and situational factors,[4] then it still seems our dispositional construal of an agent's motivational character is going to depend heavily on how wide a range of counterfactual situations we are going to include in our analysis of the relevant dispositions. Add to that the formal complexities concerning the ways in which we would have to individuate and weigh the myriad of counterfactual situations in order to determine which of his dispositions

[2]E.g. Harman (1999); Doris (2002).

[3]E.g. Isen & Levin (1972); Milgram (1974).

[4]E.g. Van Zomeren et al. (2011); Skitka et al. (2005); Mullen & Skitka (2006); De Kwaadsteniet et al. (2007).

would be 'strongest overall' and it seems we can attribute to an agent any motivational character we like, depending on how we set up our framework.

Roughly, I think there are two ways out of this, leading to two different concepts of a coherent "overall" dispositional pattern. The first is the closest analogue to Dennett's notion, for which I shall continue to use the phrase "motivational character." The second is the normative will.

Let us begin with the first notion. Recall that the evidence we have about our dispositions is always a limited and contingent sample of the underlying reality. Now as we have seen, in order to circumvent this limitation, Dennett infers the reality of the underlying pattern from the *predictive success* of the attribution. Furthermore, we have seen that when we attribute beliefs and desires, we try to find a theory that would correctly predict most acts of the agent by attributing *conflicting* desires in order to account for his conflicting behaviors under different circumstances and, we might add, for his verbal reports of experiencing conflicting desires pulling him into different directions even in a single situation.[5] But when we attribute a motivational character to the agent, we abstract away from those ambiguities and try to represent the dominant tendency in the agent that retains as much predictive power as possible. In order to do so, we aim for a model that works in the sort of situations in which the agent already tends to find himself. Furthermore, we understand the notion of motivational character in such a way that, should the agent's situation change radically (as in the case of the lazy person who becomes a Marine recruit) then we would say that his motivational character will change as well.[6] Hence, the motivational character is a 'conservative' pattern attribution aimed at explaining the agent's behavior in his *actual* life in

[5]I am assuming that Dennett's intentional stance does allow the attribution of conflicting individual desires to a single agent, although it is not always clear from his writings that this is what he intends. Nevertheless, without this possibility the intentional stance would seem wholly unsuitable to explain the notion of desire in common parlance, as the possibility of conflicting desires is so deeply entrenched in the human condition.

[6]If situationism is true, then the agent's behavioral traits will be different 'immediately,' as it were, from the moment he sets foot in the military base. And if situationism is false, then his behavior will still already be different when he starts his training on the basis of his old dispositions, since the punishments and rewards of his choices will now be different. But most likely, the discipline of military life is also something that recruits will *get better at* as their training progresses, which means that the change in motivational character will also involve a substantial *physical* development of the agent's brain and body. Hence, the change in motivational character is not just a conceptual flip, but also a real, causal process.

terms of coherence.

Note that this coherence has nothing to do with practical normativity: if we want to attribute a coherent desire set to the unwilling addict that maximizes our predictive success, then that set may include his desire to take drugs rather than his desire to resist it if, on average, the addict tends to take the drugs when he has the opportunity to do so. Furthermore, note also that we may expand this notion of coherence in ways that go beyond strict logical consistency, which we have discussed earlier on in the context of practical normativity.[7] Thus, if we know that someone likes to eat T-bone and rib-eye steaks, we may infer that she will probably like sirloin steaks as well. On the other hand, if we know that she tends to respond politely at work, she may well be aggressive at home, as the situationists keep telling us. Since the notion of motivational character is construed in terms of actual predictive success, all these things are allowed in so far as they work: the proof of the pudding is in the eating.

By contrast, the normative will is only supposed to be predictive insofar the actual circumstances do not deviate from the ideal conditions of rational, self-governing agency, in line with the type-1 dispositionalist approach. Thus, suppose that the lazy person does not want to be a lazy person. Every time he takes the lazy route, he had actually made a volitional judgment to do something else. In order to make it possible that such a judgment could get it right, we must cast the net of possible circumstances much wider into the realm of the counterfactual. Thus, if the lazy man believes that he *would* lose his lazy character if he were to join the Marine Corps, and if not being lazy allows him to realize many other desires that he has, then this might actually justify the volitional judgment that he does not want to be lazy. It might even justify joining the Corps, if being a Marine does not conflict with the desires that he has.

Volitional interpretation thus involves a survey of "alternatives of oneself," to borrow a phrase from Jan Bransen (2002). However, this also

[7] It may be argued that a notion of coherence beyond mere consistency is already presupposed at the level of desire attribution. Without it, any set of behavioral data might allow an infinite amount of attribution sets that are logically consistent with those data, making inductive reasoning in order to predict future behavior impossible. Hence, like any predictive empirical theory, a theory about the mental states of an agent must assume empirical regularities that go beyond the data themselves. Furthermore, it is only in the light of such regularities that the intuitive idea of conflicting motivations in different situations can be made sense of. Otherwise, we could simply attribute desires whose content would be tailored to unique situations and therefore conflict with no motivations in different situations at all.

implies a limit on the range of relevant counterfactuals. Replacing my brain would no longer yield an alternative of *myself*. Neither would a possible world in which my environment and upbringing had been radically different since my birth. Therefore, the notion of the normative will involves a concept of identity: it is about recognizing a pattern in what I could be, given the empirical facts about who I am. But note that the implicated identity relation is not the often discussed notion of "personal identity through time," which is presupposed by our social relationships and institutions of accountability. Rather, it is an identity relation that holds between modal alternatives to my current, actual self at this point or period in time, which captures the fact that all these alternatives are manifestations of the same dispositions that are grounded in the empirical facts about my actual self.[8]

This concept of identity has already been presupposed in our discussions of "ideal selves." If we think that the notion of an 'ideal self' refers to a possible alternative of myself, then this identity relation must hold between me and my ideal self, in order to establish that it is *my* ideal self and not just 'an' ideal self. By contrast, if we think that the 'ideal self' does not depict a possibility, as I have suggested, then we might say that the relevant identity relation must hold between successive alternatives of myself that *approach* my ideal self in the limit, such that they approach, once again, *my* ideal self, and not just 'an' ideal self.

9.1.4 Identity and Proceduralism

This last observation, that the dispositions in which the normative will is present as a pattern must be determined in the light of a modality that allows us to identify with our ideal selves, gives us a further clue about the scope of this modality. If the ideal self is the limit of deliberative improvement in the light of a proceduralist understanding of disconfirmation, as I have argued, then it makes sense to think of my normative will as ranging over dispositions that are constrained by proceduralism as well. Let me explain.

[8] I am not sure whether this is simply identity across possible worlds. One might say that the Sander Voerman who would have been given a radically different education from the beginning would still be identical to me in some basic sense of identity across possible worlds, even though this person would only share my genetic dispositions, which are too shallow, and also too arbitrary, in order to establish the inner reality that we are trying to analyze.

Suppose that three SS officers in Nazi Germany, who committed hideous acts during the Holocaust when they were, say, 30 years old, had had sufficiently similar genes and early upbringing that their overall dispositions were more or less the same when they were about 15 years old. However, after that, their further development into 30 year old war criminals progressed in different ways. In the case of the first officer, despite his atrocities, there remained a flexibility in his thoughts and motivations that would lead him, long after the war, to regret his acts and disconfirm his prior judgments. At the age of 60, let us say, officer 1 has arrived at the volitional belief that his crimes were horrible mistakes.

By contrast, the other two officers never renounced their earlier actions, and kept defending the Holocaust until their deaths. However, let us imagine that in the case of officer 2, this was partly due to the circumstances of his life after the war. If those circumstances had been different, then in principle he could have been made to understand that he did not have good reasons to act the way he did. We might say that his failure to appreciate these facts was a matter of *eliminable* epistemic bad luck.

Instead, let us suppose that for officer 3, given the way in which his psychology had developed between his 15th and his 30th year, there was no longer any proceduralistically possible route that would have lead him to disconfirm his support for Nazi Germany. If he would have been given the medical treatment to overcome the biological limitations that make us susceptible to diseases of old age, let us say, and if he would have continued to live through the centuries, then still no requirement would ever be demonstrable to him that would have showed him an error in his ways. *Ex hypothesi*, then, officer 3 is quite like the Martians in this respect.

The question is, did these officers act against their normative will during the Holocaust? Now even in the case of officer 1, a skeptic might wish to deny that there were any such deep facts about the will of the officer at that time, despite the judgments of his later self. Instead it might be simpler to say that the will of the officer merely changed after the war had been lost. But we have already seen how the Affective Response View allows us to say more than that: if the practical beliefs of the 30-year-old officer implied predictions that turned out to be false, then we have a reason to think that the officer already did not want, in the normative sense, to be a Nazi at the time. And it does seem plausible that there would be such predictions. After all, Nazi education was not exactly a celebration of critical thinking. SS trainees were not instructed to seriously attempt

to love and cherish their Jewish or homosexual fellow human beings in order to critically examine the Nazi theory that such a thing would be pointless. If you would have told officer 1 that his life would have been richer and more rewarding if he could have loved the people that were now his victims, he would've laughed at you (or reported you to his superiors). But he never *tested* those predictions, and his later self might well conclude that his earlier self kept his views exactly because he never tested them.

Of course, it does not follow that life must have been more pleasant for him as he reached his 60s. On the contrary, the immense feelings of guilt that officer 1 must have gone through would have made his life absolutely miserable. So the guilt-ridden 60 year old had no way of *directly* testing how a different life for his past self could have been. But as we have seen, volitional interpretation offers various roundabout ways. Perhaps the old man recognizes how he was as a 15 year old teenager when he watches his grandson grow up. And perhaps he can see how life is so much better now for his grandson in a liberal democracy—something he was taught to abhor during his indoctrination. Furthermore, he can now see how various elements of the Nazi belief system that were blatantly false—concerning the biology of race, for example—were designed to make it easier for him not to feel horrible about torturing other people. None of these elements were sufficient in their own right, but combined with the authoritative pressure and the group thinking of the Nazi society, we now know how our disposition to feel horrible about torture can be silenced. But that doesn't mean the disposition is no longer there. What it means is that the relevant difference between the 60 year old ex-officer and his 30 year old former self lies in their beliefs and circumstances. Hence, it seems justified for the 60 year old to postulate an alternative of his 30 year old self that would have recoiled in horror from his acts, and would have lived a better life as a tolerant liberal.

So the point is not merely that his dispositions enabled a happy and flourishing alternative of his 30 year old self in a different society, showing that he has a disposition to favor such a life. The point is also that to some extent, the affective manifestation of his 30 year old *actual* self *does not even count* because if how strongly biased it is by, and dependent upon, his ignorance. Rather, if we want to compare his disposition towards living in a liberal democracy with his disposition to being an SS officer in Nazi Germany we must project how he *would* have felt as an SS officer if he had all the relevant knowledge about the psychological mechanisms mentioned

above and if he had fully understood his alternatives. His normative will is that pattern in his dispositions that will manifest itself across different situations *as the agent comes to understand* himself.

This entails a certain non-straightforward dialectic of self-knowledge: it is knowledge about how we would be if we were to have that knowledge. One might object that this knowledge is therefore empty, or indeterminate, or in principle inaccessible: if the state to be known is a state that does not obtain as long as we do not know, then we can never gain that knowledge. But that does not follow, because gaining knowledge is a gradual phenomenon, and we may assume that we already have *some* knowledge about the nature behind our own feelings, which means that the feelings we have can at least tell us *something* about the feelings we would have if we knew even more about them. Therefore, if by understanding our feelings, we subsequently change our feelings, then the object of our knowledge does seem to be a moving target, but the 'curve' of how that target will progress may slowly approach a fixed point where the gap between the normative and the cognitive will would finally be closed: the limit of the ideal self. Of course, if the ideal self is not a real possibility, as I have suggested, then this point is always infinitely far away, so to speak, and the ideal would be like an asymptote to which the curve approaches ever nearer.

We can now see how the second officer also may have acted against his normative will. As I briefly discussed in section 7.4.2, mechanisms of self-confirmation bias are common. It may require special circumstances to turn people into Nazis, but once we have settled on a belief set, we are intrinsically inclined to stick with it, especially if rejecting it would imply harsh criticism of our own actions, as in the case of officer 1. Therefore, it is not surprising that many war criminals have never come to renounce their acts. But from that it does not follow that they *could* not have disconfirmed their views in principle. Because psychological mechanisms of bias should be classified, I submit, as matters of *eliminable* epistemic bad luck. And we can argue for this with respect to theoretical matters of fact that are not response-dependent in the way that practical reason is. Many people are so strongly biased that they fail to take the evidence for the theory of evolution seriously, for example. Some people even insist, against all odds, that the massive evidence for the idea that the earth is billions of years old is merely staged by God to test our faith in a literal interpretation of the Bible. Clearly, these failures are not matters of ineliminable epistemic luck: they depend on the contingent limitations of people's cognitive

abilities. Our ideal selves do not have unlucky twins that fall prey to these types of mistakes. Furthermore, the things that these unlucky twins might falsely believe in, such as nominalism in metaphysics, would not really be matters of bias, because in such a case there was no *evidence* to be assessed that would have made a difference: there was just no way they could have known. Hence, bias is a matter of bad procedure; it is in principle eliminable. But given that psychological mechanisms of bias introduce this sort of eliminable luck, we should expect the same thing in the field of practical reason.

Thus, it seems plausible that many war criminals follow the course of officer 2: even though they closed their eyes for it until their deaths, the evidence about their inner dispositions was out there, just like it was for officer 1, and would in principle have made it possible for them to disconfirm their views. Once again, let me stress how this distinguishes relationalism from superficial relativism: even with respect to Nazi war criminals who never showed any signs of regret we can have reason to believe that they were acting against their inner normative reality.

To be sure, the theory does not *guarantee* that all of them are like officer 2 either. It is at least conceptually possible that some remorseless criminals are rather like officer 3. But can we say that officer 3 acted against his normative will? Here I think the answer must be no. Of course, like officers 1 and 2, officer 3 is a product of his upbringing, and insofar this involved indoctrination and uncritical education, we must understand the normative will of officer 3 in terms of how his dispositions would manifest themselves under conditions of critical thinking and self understanding. However, our upbringing does not only teach us how to think, it also shapes our emotional nature. Therefore, if the nature of officer 3 has been shaped in such a way that no procedure of inquiry could in principle make him experience the sort of unexpected affective responses that would disconfirm his views, then it simply no longer contains the 'dormant' or 'latent' sort of pattern in favor of tolerance that we attribute to officer 2.

It is an empirical question whether remorseless Nazis were mostly like officer 2 or officer 3, and therefore, whether they acted against their normative reasons. Were they so strongly and deeply influenced by years of upbringing that their inner nature had become robustly in support of torture and genocide? Or was their behavior better explained by mechanisms that do not presuppose such drastic psychological changes—and therefore differences in comparison to us—at all? Interestingly, this is exactly the

question that motivated the empirical research of social psychologists in the 60s and 70s which I mentioned above. Insofar their situationist conclusions prevail, it seems that many of us could be persuaded to commit atrocities in a manner of days, given the sort of circumstances that have been demonstrated to elicit such behavior. Now one of the problems with situationist experiments is that they tend to 'hide' individual differences by randomizing their subjects. Of course, if we cannot randomize then any causal attributions become problematic, which is why personality traits are so hard to investigate scientifically. Thus, situationist experiments do not show that individual characteristics cannot have strong causal influences, but what they do show is that we *can* construct circumstances under which situational factors become clearly dominant. It is in this respect that I believe situationist results can still teach us something about the Holocaust.

To be sure, experimental subjects in the Milgram experiments, for example, reported a lot of emotional difficulty and negative affective responses afterward, in sharp contrast to those real life war criminals who remained remorseless. However, given the apparent ease with which people are led to violence and intolerance, we may speculate that the difficulty for people in admitting wrongdoing—even to themselves—is better explained by the mechanisms of bias alluded to above, than by a deep transformation of their inner psychological nature as they came to commit their crimes. It is one thing to be bothered by your own behavior during an experiment and another to accept responsibility for years of horrible criminal behavior. Furthermore, like the research participants, many soldiers did admit that their course of action was wrong while explaining their reasons for taking that course of action in terms of following orders and a transfer of responsibility to the authorities whose commands they were following. I realize that while this explains why those who where themselves in places of authority may have been less likely to admit wrongdoing, it also makes their situation different from the participants in the Milgram experiments. However, not all situationist research is focused on the role of authority or the circumstance of following orders.

Nevertheless, I do not mean to defend the claim that as a matter of empirical fact, exactly 100% of all human beings, with no single exception, share normative reasons against torture and mass murder. Psychopaths, for example, may be so emotionally different from us that the affective dispositions to which our practical judgments are attuned, are simply

absent in their case. However, the empirical evidence in these cases seems to suggest that the development of such persons already deviates from ours in their early childhood, and often as a result of severe abuse and neglect. Now it is conceivable that Nazi Germany was a suitable habitat for some individuals with extreme antisocial personality traits to lead successful careers in the military and the SS. Perhaps the top Nazi leaders were among them. But as I understand the science, the common soldier in the SS would have been psychologically more similar to you and me than to Charles Manson or Richard Kuklinski.

9.2 Mild Realism

Suppose that Diane has finished high school and is wondering what she wants to study at university. She has managed to narrow down her list to two options: philosophy and chemistry. Both would involve many things she would enjoy, or find rewarding, or that could help her find an enjoyable career afterward (the latter maybe less so in the case of philosophy). Both also involve things she might dislike, as well as risks that could lead her to drop out. As Diane surveys these two alternatives she tries to figure out which of the two would manifest most of the dispositions for positive affect, and least of the dispositions for negative affect, that she has known herself to have on the basis of her responses to all sorts of things in the past. Perhaps she found chemistry interesting in high school, but nevertheless hated to have to learn so many facts by heart. Perhaps she enjoyed discussing and arguing on philosophical questions in various online communities, but was sometimes frustrated by the seeming impossibility to ever really settle a major philosophical question. And the pattern that she's looking for may extend to various other situations: maybe she always finds herself reading the lists of ingredients on the package labels of various products, intrigued by the chemistry behind them. Or maybe she is always amazed when she sees pictures of giant oil refineries and offshore platforms.

It may well be that on balance, she would like chemistry more than she would dislike it, while she would *also* like philosophy more than she would dislike it. But which of the two would she like *most*? Her attempt to determine this is not exactly hard science. We do seem to have preferences regarding the different things that we like, but in the above example, without any exact methods to quantify, measure, and compare

one's attitudes towards the two options, the correct answer to her question, if any, does not really seem to be a very 'hard' matter of fact.

But are there such things as 'soft' facts instead? Interestingly, in his discussion of real patterns, Dennett does find ways to construe patterns as matters of fact that are less determinate than the things in which these patterns manifest themselves. First of all, the presence of a pattern is a matter of *degree*. Second, sometimes rival patterns can co-exist in the same set of data, which can render the correct attribution indeterminate. In the light of these insights, Dennett calls himself a "mild realist" about intentional attitudes. As I will argue below, similar observations apply to the normative will. But in addition to that, we will see that there is an interesting further phenomenon at the volitional level: when the facts are less determinate at the time of choice, the choice itself can render them *more* determinate.

9.2.1 Gradual Presence

Let us once again consider Dennett's bar code images (1991b, p. 31, figure 1). He describes how they are generated: each started out as a 'pure' bar code, and then a certain percentage of pixels were randomly selected and inverted to generate the noise. Image *A* has a noise ratio of 25%, for example, while image *D* has only 1% noise. Finally, Dennett also added an image *F* with a 50% noise ratio, just to demonstrate that in such a case the pattern is lost completely. Note that these percentages correspond to our success in predicting the color of a pixel on the basis of the bar code pattern if we do not know how the noise has been distributed. In the case of image *F* these would come out 50/50, which is the rate of chance for binary options. In the case of *A* we'd stand a 75% of being right, and in the case of *D*, 99%.

Now in image *F*, there simply is no real pattern. There no longer is a fact about the image being an image of a bar code, because it isn't. It is just noise. On the other hand, even though *A* and *D* do seem to contain real patterns, the fact about these images is not the 'hard' fact that they are copies of the bar code image, because they are not. Instead, the facts about them are facts about the *degree* in which the bar code is present in these images, and this degree varies between *A* and *D*. Now suppose that Diane really wants to study chemistry, and that even though she dislikes some things about it, there are definitely many more things that she likes

about it. From the point of view of the attribution that she wants to study chemistry, the things she dislikes about doing that are the noise, and their magnitude the noise ratio. In the absence of any specific knowledge about these matters, on the basis of the claim that she wants to study chemistry, we might guess for each aspect of chemistry that she would like it, and fail to predict her affects when we hit upon the noise.

Now suppose that Diane got her volitional judgment right (which would mean that any other choice would have raised the noise ratio, as I will argue below). And suppose that Kevin also decided to study chemistry, and that he also got that right. In that case, for both students, it would be a fact that they really wanted this. Nevertheless, the noise ratios might be different: Diane might dislike much more about chemistry than Kevin. In that case, we should say that the pattern is stronger, or present to a greater degree, in Kevin's dispositions when compared to Diane's.

9.2.2 *Conflicting Patterns Can Co-Exist*

The fact that real patterns admit of noise ratios implies that conflicting overall patterns—conflicting in the sense that pattern A leads to predictions that contradict those based on pattern B—may *both* be present in the same set of data. There are roughly two different ways in which this might happen. The first case is where pattern B is really just a more sophisticated version of pattern A, such that many elements that appear to be noise from the simple description of A can be accounted for as more complex regularities in the light of B. Dennett argues that when one man (Jones) bets on pattern A while another (Brown) bets on B, they are not really in disagreement, but merely adopt different strategies: while Brown can budget for a lower error rate, Jones requires less time or resources to do his calculations, which might for example mean that he can make more bets (p. 35–36). Still, one might be tempted to say that B would be the 'more real' pattern in this case. However, suppose that one day, Brown does not have the time or resources to employ his complex betting strategy. He might still make money if on that day, he would simply bet on A (provided that he can take the hit from the higher error margin). And thus, we might still say that Brown would have missed a fact about his business if he had overlooked the real presence of A.

Something analogous may apply to our analysis of the normative will. We might think of Diane's judgment that she wants to study chemistry

as the attribution of a relatively simple pattern *A* with a fairly high noise ratio. At least to some extent, Diane will be able to predict in advance which parts of chemistry she is going to dislike. Perhaps she would say, "What I would *really* want to do is chemistry, but without courses *X* and *Y*, and with the addition of course *Z* that they teach in philosophy." And perhaps the university will allow her to do just that, in which case acting upon this more sophisticated pattern *B* seems the right thing to do. But even then, there will be an upper limit on how fine-grained she can make her plans and policies. Therefore, the trade-off between precision and simplification that Dennett wants us to acknowledge seems to apply to volitional interpretation as well.

Nevertheless, in order to simplify my *own* theory, I will continue to speak of the normative will as a single pattern in such scenarios, on the grounds that if the different patterns are, despite their conflicts, not really in disagreement but rather at different levels of simplification, then the level with the lowest noise ratio would still be the most accurate depiction of the coherent set of goals that the agent wants to pursue *most*. If it would be in the interest of the agent to spend less time and energy on his deliberations, then paradoxically one of the normative volitional attitudes *at* this complex level of description will contain the prescription that the agent deliberate on a *less* complex level of description. I think this can be perfectly consistent, but I will not dwell on it any further at this point.

Now the second manner in which multiple patterns can co-exist is far more interesting for our present purposes. Suppose again that there is a pattern *A* in some set of data with a fairly high noise ratio of, say, 40%. Then there would be enough room for a pattern *B* that would only partially overlap with *A* in the 60% area that *A* is getting right, as long as *B* would cover a substantial amount of the 40% that *A* is getting wrong. Suppose that 30% is covered by both *A* and *B*, 30% only by *A*, 30% only by *B*, and 10% by neither. Then *B* would have a noise ratio of 40%, just like *A*, while *A* and *B* would contradict each other in a majority of 60% of the data. But if such a thing could apply to our affective dispositions, then how do we determine the normative will?

For example, suppose that *A* captures what Diane would like about studying chemistry while *B* captures what she would like about philosophy. Then what would she 'really' want, if anything? First of all, I don't think patterns that conflict in the majority of situations would apply to this particular example. After all, studying chemistry or studying philosophy

are options that have a lot in common, and therefore the underlying patterns would overlap substantially. Once Diane has narrowed down her choice to these two options, it seems plausible that a coherent set of dispositions that would be served by either option must already account for a majority of her affective responses regarding her occupation, in contrast to, say, working at a supermarket or joining the Marine Corps.

I realize that it does not make much sense to put numbers on these things without any formal model to specify what those numbers would mean, exactly. And although I would very much like to build a model of the normative will in formal language, such a project is beyond the scope of this thesis. For now, we can talk about noise ratios of 10% or 40% in the case of visual images and things like that, and merely treat these as metaphors when applied to the normative will. Nevertheless, a disagreement of 60% seems like a bad metaphor in the philosophy–chemistry example. Rather, what I think the example illustrates is that sometimes, when two options seem to receive equal amounts of support from our dispositions, this merely indicates that we have already achieved substantial deliberative success by narrowing our choice down to these options. There must come a point where one's normative will no longer discriminates between different alternatives, but that hardly shows that there are no facts to be known. In the example, Diane does know that she wants to study at the university, that she does not want to join the Marine Corps, and that she probably will be happy regardless of whether she picks chemistry or philosophy.

Still, it does not follow that she has to toss a coin. Given the enormous complexity of our psychological natures, there seems no reason to think that the noise ratios of the two alternatives would ever be *exactly* the same, in the same sense in which no two human beings are ever going to be *exactly* of the same height or weight. Thus, it will probably still be true that either philosophy or chemistry is what Diane wants *most*, and she might still try to make an informed guess as to which one it is. However, if the two patterns have nearly equal strength, then we should realize that while the presence of the pattern in support of the disjunction is of a high degree (because of their large intersection), the presence of her normative preference of the one option over the other is of a very low degree (because of the small difference between their noise ratios).

Of course, things start to look different when an agent has not narrowed down his options very much, or when the options that remain are still

so substantially different that the example of the patterns which disagree on 60% of the data suddenly does become an applicable metaphor. For example, suppose that Diane is wondering whether she would rather join the Marine Corps than study philosophy. In that case she does contemplate two very different alternatives of herself. If her dispositions seem divided with respect to such alternatives, and if the noise ratios of the two patterns happen to be comparable, then the presence of her normative will with respect to the choice of her occupation would be rather indeterminate. Technically, we might still say that the alternative with the slightly lower noise ratio captures what she wants *most*, but if the difference in content is so massive while the difference in support is so small then the big fact to be appreciated is the absence of a dominant majority, not the fact about which option slightly outweighs the other one.

Compare this to the outcome of an election where two candidates are radically opposed to each other's proposals, and one wins by a tenth of a percent. Now suppose that, contrary to real life, the people actually managed to vote exactly in accordance to their normative reasons. Furthermore, the candidates managed exactly to formulate the proposals for which their supporters wanted to vote. If the elected candidate gets to implement his proposal in a 'winner takes all' sort of way, without any further need to compromise, then we would hardly say that this policy reflected the reality of the will of the people. Rather, the reality would be that the people had been hopelessly divided on a question that apparently did not admit of a middle ground.

Our analysis implies that this can also happen to a single individual. But this is not a problem for our theory, even if it may be a big problem for the agent in question (as it would be for the society in the example of the election). The Facts Principle does not imply that there will be facts to settle practical questions in *all* cases. It merely implies that there will be such facts in *some* cases. Furthermore, this feature of the theory allows us to account for a phenomenon that many moral philosophers have had trouble dealing with: that of *genuine moral dilemmas*. The idea is that in the case of a genuine dilemma, when the agent has to decide between two options, both are morally required (or both are morally wrong) and so no matter what the agent does, he must fail.

Thus, in Sartre's famous example of the student who is torn between joining the Free French forces and caring for his ailing mother, one might suspect that which ever option he chose, remorse was the predictable

consequence. And according to our theory it is now possible that in each case, the remorse would reflect a pattern supported by a majority of his dispositions. Hence, in each case the remorse would seem internal to him as a person, to use Frankfurt's phrase again. He could not but go against himself. I will say more about this predicament when we turn to the matter of wholeheartedness below.

9.2.3 The Snowball Effect

In my discussion of Williams's remarks on deliberation about ends in section 2.2.2, I noted his idea of "constitutive solutions, such as deciding what would make for an interesting evening" (Williams, 1980/1981a, p. 104) as an example of indeterminacy in our ends. Of course, Williams's point is that once a solution has been adopted, the indeterminacy may be resolved: by having settled on a concrete intention or plan A, it may have become unreasonable to subsequently switch to an alternative plan B, even if it would have been reasonable to have initially settled for B rather than A. The reasons in favor of option A gain weight, so to speak, as a result of our commitments to and investments in that option. Following Michael Bratman (1987, p. 82), I shall call this the "snowball effect."

There are two complementary mechanisms behind this effect. First of all, by directing our thoughts and ideas to a course of action that we have chosen, and by getting accustomed to its implications or results, our emotional lives may become 'shaped' in its favor. And secondly, once we have invested a certain amount of effort in option A, there will be a *cost of transition* associated with a switch to option B in terms of abandoning the investment in A and making a similar effort for a second time in order to get started with B, that may put B at a disadvantage.

These mechanisms operate both in the short and the long term. Thus, in the short-term it is usually pointless to keep deliberating or switching plans concerning what restaurant to pick, or whether to order pizza rather than Chinese food for dinner. Instead, at some point you should just pick something and stick with it. In the long-term, the snowball effect may have a profound impact on the development of our normative will over time. Suppose that Pete desires to become an academic philosopher, but that he also desires to become a park ranger. As we have seen in the previous sections, two conflicting affective patterns may be equally present in his emotional life, such that the first desire is a constituent of pattern A and

a piece of noise with respect to pattern B, whereas the second desire is a constituent of pattern B and a piece of noise with respect to pattern A, while the total noise ratios of A and B are comparable.

Suppose that Pete decides to become a park ranger, and that he succeeds in becoming one. After years of working in Sequoia National Park, his motivational characteristics might have developed in such a way that his normative will has become determinate: he wouldn't want to become a philosopher anymore. But this is perfectly consistent with the possibility that, if Pete had become an academic philosopher, his motivational characteristics would have developed in such a way that he wouldn't want to become a park ranger anymore. His normative will might settle for whatever path he chooses.

Thus, the snowball effect illustrates that even though the cognitive will is meant to get the normative will right, it also creates or shapes the normative will over time. In other words, contrary to some of Harry Frankfurt's later remarks, the content of our normative will may sometimes really be "up to us." As Bransen put it, deliberation is both a matter of "making ourselves" and "finding ourselves" (2002). Whereas existentialist and decisionist approaches, such as the philosophy of the early Sartre, may have stressed the aspect of self-making too much at the expense of self-knowing, purely cognitive approaches to deliberation, which understand the metaphysics of normativity in a matter completely independent from the personal deliberative activity of the particular individual agent, run a risk of neglecting the self-constitutive element. I take it to be an advantage of my proposal that it clearly allows us to accommodate both aspects.

Sadly, however, the causal influence of the cognitive will upon the normative will need not always be self-fulfilling. It can also be self-defeating. Suppose that A has to decide between ϕ and ψ. It might be that ϕ would change his life in such a way that his normative will would become determined in favor of ψ, and *vice versa*. If there would be no possibility of reversing the decision, then the agent would have fallen victim to a kind of inverse snowball effect, or what we might call the 'grass is greener on the other side of the fence' effect. I think the snowball effect is usually stronger, but I suppose the self-defeating scenario does really occur now and then. This is no drawback of the theory, of course, but only another tragic fact about life.

What these scenarios have in common is that prior to deliberation, there really is no fact of the matter about which option the agent should choose.

But we have also discussed several examples in which it did seem plausible to say that there is a fact of the matter about what the agent had normative reason to do, even when the agent himself did not recognize it—both in moral and political contexts and within the context of the personal life of an agent. The attractive feature of being a 'mild realist' about the normative will is that we can *deny* that there is a fact of the matter with respect to examples which make the Facts Principle look implausible, while at the same time allowing that there *is* a fact of the matter with respect to the examples that make it look plausible.

9.3 Two Types of Wholeheartedness

As I discussed in section 8.2.2 of the previous chapter, when Frankfurt initially tried to solve the fundamental problem of explaining how any type of desire could have agential authority—including a "second-order volition" when understood simply as the desire that a certain first-order desire be effective—he turned to the idea of "decisive commitment." This approach found its culmination in his essay "Identification and Wholeheartedness" (1987/1988d), in which Frankfurt argues that an agent's commitment is decisive when it is "fully resolved," by which he meant that the agent identifies with his desire "in the belief that no further accurate inquiry would require him to change his mind" (pp. 168–169).

In the same essay, Frankfurt also introduces the concept of "wholeheartedness," which is the state of being *coherent* in one's authoritative desires. This notion is contrasted with the state of "ambivalence," which involves a "conflict within the authority itself," an incoherence between desires that both express what the agent really wants. Note that many conflicts of desire are not cases of ambivalence: if the agent is wholly behind 'one side' of the conflicting desires, so to speak, then she remains wholehearted. The conflict really has to be between desires with which the agent is identified. But if identification is to be understood in terms of being fully resolved, then we may wonder what it means to be fully resolved on both sides of a conflict.

As we have seen, in his later work Frankfurt gives up on the idea that decisive commitment really explains volitional authority. But he has kept the notion of wholeheartedness as a kind of further analysandum for which he wants his theory to be able to account. Thus, in his recent writings on the authoritative nature of love, Frankfurt allows that we

can either be wholehearted or divided in what we love, where what we love is understood in the opaque sense. Echoing Spinoza's notion that satisfaction with oneself is the "highest we could hope for," Frankfurt presents wholeheartedness as a kind of ideal configuration of agency (2004, pp. 97–98).

In this section, I want to accommodate a similar concept of wholeheartedness within my own framework, which I shall call "inner wholeheartedness." And of course, I am in agreement with the later Frankfurt concerning the inability to account for the volitional in terms of being fully resolved. Nevertheless, I do think that the notion of being fully resolved in one's volitional judgments is *also* an important concept, and one that may in fact be understood as *another kind* of wholeheartedness: it involves being wholehearted in one's *assessment of the evidence* in support of one's volitional judgment. Whereas inner wholeheartedness is a state of the normative will, being fully resolved is a matter of the cognitive will, and the two are independent phenomena. Let us now discuss each in turn.

9.3.1 Inner Wholeheartedness

We have already seen that conflicting patterns may be present in our affective dispositions, in such a way that our normative will may be indeterminate with respect to certain practical questions. The phenomenon of ambivalence (i.e. the lack of wholeheartedness) may be explained in a similar fashion, except that indeterminacy and ambivalence only partially overlap: some cases of indeterminacy are not cases of ambivalence, and vice versa. Sartre's example of the student who has to choose between fighting for his ideals and caring for his mother seems to be a case of both indeterminacy and ambivalence. Whatever he decides, he cannot abandon the other option wholeheartedly. However, it would be odd to say such a thing in the case of the choice between ordering pizza or Chinese food for dinner. Even when my normative will is indeterminate with respect to those two options, I need not be *divided* about them: when I choose pizza I am not going 'against myself' simply because I also would have liked Chinese food.

More importantly, we can lack wholeheartedness even when our normative will is determined. Suppose that Henry has been in a relationship with Janet for several years, and that he loves her very much. But he has also fallen in love with another woman, Sarah, and as time goes by, his love

for Sarah evolves to the point where it becomes impossible for him to make judgments about which of the two women he 'loves most.' Answering that question, for him, would be like a parent trying to answer the question which of her children she loves most. Of course, in ordinary life parents do not have to choose between their children: they can raise all of them. By analogy, some people opt for a 'polyamorous' lifestyle, which involves the agreement by all people involved that it is okay for someone to have multiple romantic relationships at the same time. And in our example, perhaps this is what Henry would want. However, like most people, Janet and Sarah may not be open to such a thing, and even if they are, then it might not work out, in which case Henry does have to choose between them.

But from the fact that he cannot answer which one of them he loves most, it does not follow that he may not have reasons for judging which of the relationships he wants most. If he has children with Janet, for example, then that may provide an additional reason to stay with her. But there might be other reasons involved. If a relationship with Sarah would simply work out better in the end, in his view, then Henry may have a resulting reason to leave Janet. In such cases, using our framework of pattern recognition, we might say that while the two options are supported by two conflicting patterns, one of these patterns does outweigh the other in terms of its lower noise ratio. Thus, suppose that the pattern in support of a relationship with Sarah is the stronger one, and that Henry has recognized this by making the volitional judgment that he wants to leave Janet. In such a case, it seems there would still be a lack of wholeheartedness: by making that choice, Henry would still be breaking his heart (not to mention Janet's).

What we should say, I think, is that there is a non-resultant sense of identification that allows Henry to remain identified with both alternatives even after he has made his decision and even when he is really convinced that he did the right thing. The pattern in support of his relationship with Janet harbors such a deep love, and such a robust structure of dispositions in his emotional being, that losing her really means that he loses part of himself. But by staying identified with that structure of dispositions, he also manages to retain that part of himself in some way. He could not honor, or do justice to, or perhaps mourn, the connection he had with Janet without some mode of identification. Recognizing how divided his heart is, even if one alternative does outweigh the other, is part of 'getting it right.'

Thus, inner wholeheartedness is a matter of fact about the normative will of an agent: it is the case when all that the agent should identify with in the above sense is captured by the normative will.

I suppose it is in some sense trivial that it is better to be wholehearted in this inner sense. Nobody hopes that life will often force one to tear oneself apart. In that respect, at least, I am in agreement with Frankfurt's appeal to the Spinozistic ideal of inner harmony and tranquility. However, on my analysis, that does not mean that inner wholeheartedness must be present in our ideal selves. On the contrary, it is only when multiple patterns retain their strength in one's ideal self that each of these patterns carries an amount of authority that demands one's identification with it. By contrast, consider the case of the unwilling addict: his addiction may be a very strong pattern in the 'motivational character' sense distinguished earlier on, but because it would be absent in his ideal self, the addict can be wholeheartedly against his addiction.

And from the fact that wholeheartedness need not be present in our ideal selves, it also follows that it is not an ideal to be *accomplished*. When the fact is that we are divided, I think we should simply recognize that fact. Over time, when we choose one course of action over the other, it may well be that the snowball effect, which we discussed earlier on, will reduce the ambivalence by expanding the pattern acted upon while diminishing one's regrets over the path not chosen. Nevertheless, time does not heal all wounds, of course. It may also be a sign of authenticity to be able to admit that we keep carrying the remnants of some of our decisions with us.

9.3.2 *Epistemic Resolvedness*

The idea of making a judgment "in the belief that no further accurate inquiry would require one to change his mind" does not only apply to practical judgments, of course. Instead, it depicts a general epistemic phenomenon which may in principle apply to all matters of belief. In the terminology of William James, it involves the distinction between "live" and "dead" hypotheses (1896/1979, p. 14): the latter are propositions that we cannot seriously consider as candidates for truth anymore in the light of our experience. Thus, when the proposition that not P is a dead hypothesis for us, we are resolved in our belief that P. In the terminology of C.S. Peirce, any attempts to doubt the truth of P would merely be "paper doubts" (1877/1992b, p. 115). We have many beliefs that are resolved in this sense.

For example, I do not think any future discovery is going to disconfirm the idea that humans and chimpanzees share common ancestors, or that the earth is much older than 7000 years. Of course, from the fact that *I* am resolved in these beliefs it does not follow that other people cannot believe otherwise, and in fact they do.

Frankfurt discusses the example of a mathematician who is checking and rechecking his calculation (1987/1988d, pp. 167–169). If he keeps getting the same answer, and cannot find any mistakes, then unless he wants to keep doing this for the rest of his life, he must either decide that his answer is correct, or abandon his activity without endorsing its result. In Frankfurt's view, the reason supporting the former option could be either that he has come to believe with "full confidence" that no further inquiry would turn out otherwise, or that the possibility of discovering a mistake has become so small that the cost of further inquiry outweighs its benefit.

However, in the case where the mathematician reaches full confidence, even if it is up to him to decide that he is not going to do the calculation another time, I do not think we should say that his confidence is itself established by that decision. Instead, becoming convinced is something that happens to him, while his decision is a response to that. Furthermore, if he has really become fully convinced, then this decision is straightforward: all the alternatives are dead. Let us therefore call this manner of reaching full confidence "passive resolution." It is the strongest sense in which we can be resolved in our beliefs, and it lies outside our deliberative control.

The second, and weaker, sense in which we may become resolved in our beliefs is when we judge between "live" hypotheses on the basis of what we find reasonable to assume, given the evidence that we have. Such a judgment is not straightforward: it is an act of reasoning, a jump to conclusions, for which we take upon ourselves a kind of epistemic responsibility to be able to justify the belief we have adopted. Beliefs arrived at in this matter may vary in degrees of confidence, although I shall not attempt to model or quantify that dimension. Instead, let me just assemble them under the heading of "active resolution." Together with the beliefs arrived at through passive resolution, these beliefs make up what we take ourselves to *know*. Of course, this self-attribution of knowledge is fallible: some people may be extremely confident in their false or unjustified beliefs.

Sometimes, however, we may be unable to settle our doubts in this

manner. Sometimes we know that we do not know, that we do not have a sufficient epistemic justification for any of the live hypotheses that we are entertaining, and that we cannot make that jump to a conclusion. In theoretical cases this may not be a problem: we simply confess to our ignorance. But in practical cases our choice may be "imminent," to borrow another term from James: we may have to decide on a course of action such that doing nothing would be making a choice as well. In such a case, we may have to guess what is true on the basis of our subjective probabilities, and adopt our guess as a kind of *working hypothesis*.[9]

Suppose that I am wondering whether I should ϕ, and that this depends on whether P is the case, but I do not know whether P. If my subjective probability that P is slightly above chance, then I may adopt the volitional belief that *I want to act upon the assumption that P* (without adopting the belief that P). Aside from the fact that this allows me to respond to imminent dilemmas, it also allows me to make plans and act consistently over time, even if my subjective probabilities keep fluctuating during that time. Thus, if I keep flip-flopping between finding P most probable and finding not P most probable, I can adopt the intention to act upon the assumption that P as a matter of *policy* for a certain amount of time, after which the question of whether P may again come up for review.

And of course, when deliberating about such plans and policies, we should not only factor in our subjective probabilities—in the absence of knowledge—but also the relative costs of our different choices should we get our assumptions wrong. If I find P slightly more probable, but the cost of acting upon the mistaken assumption that P greatly exceeds the cost of acting upon a mistaken assumption that not P, then perhaps I should act upon the latter assumption. However, the details of this matter are beyond the scope of this discussion. What I merely want to illustrate is that agential authority does not always require epistemic resolvedness: it will often suffice to adopt working hypotheses about one's normative will

[9]The analogy to working hypotheses in science re-establishes the symmetry between theoretical and practical reason in this respect, which I think makes sense. An empirical scientist cannot just 'observe' his data in order to look for patterns, because the statistical significance of such exploratory research is always problematic. Instead, in order to establish matters of empirical fact, a scientist must do confirmatory research as well: he must adopt a hypothesis *before* he obtains his data in order to truly test its predictive success. And he even runs a risk if his prediction fails, because then he must not only adopt a new hypothesis, but also obtain new data in order to run another confirmatory test. In this respect, an honest scientist is simply a practical deliberator who has a normative reason to figure out the empirical facts in his domain of research.

in combination with policies concerning volitional interpretation that allow for revision and make practical disconfirmation possible. Note, however, that such a plan does involve a practical *commitment* to the assumption about what one really wants.

Nevertheless, in some cases this may not be enough, and we'll want to adopt volitional beliefs with the amount of confidence that warrants thinking of it as knowledge. For example, most of us would want to *know* that they want to marry someone before they make their vows, rather than merely having committed themselves to a plan to give it a shot. However, so long as the resolvedness is active, based on an epistemic judgment concerning live alternatives, it does seem reasonable to keep some policies in place that promote continued volitional interpretation and the possibility of practical disconfirmation.

Finally, some of our practical beliefs may be fully resolved in the passive sense. By definition, that means we won't take the possibility of their disconfirmation seriously. As I indicated earlier on, we may think of this as a kind of 'epistemic wholeheartedness': there is no longer any room for doubt in our cognitive will. Now in some cases there may be nothing wrong with that. There is no doubt in my mind, for example, that I have a normative reason to disapprove of the Holocaust, of slavery, racism, oppression, and terrorism. And although, as a moral philosopher, I am interested in the justification behind my disapproval of such atrocities—in order to figure out whether it is relationalist or non-relationalist, for example—I do not think that further inquiry could make me approve of slavery or racism, and I do not invest my limited cognitive resources into the possible disconfirmation of my disapproval in these matters.

This does not mean that I do not need a justification for my practical belief, of course. In fact, as I have argued above, I think the Nazis had normative reasons similar to mine, and that they got them wrong while I am getting them right. I have sketched my reasons for thinking this, though a full normative ethical discussion of racism and fascism is beyond the scope of this meta-ethical thesis, of course. But what it does mean is that my practical beliefs in these matters are on a par with my belief in the general theory of biological evolution. Thus, every evolutionary biologist should be able to give you the justification of this theory, but their research time is not devoted to the possible disconfirmation of this general theory. Instead, they are working on the details of evolutionary mechanisms in a manner that takes the general fact that apes and humans

share common ancestors, for example, as empirically established beyond reasonable doubt.

Some philosophers even hold that having beliefs beyond doubt in this sense is constitutive of knowledge and intentionality in the first place. The pragmatist tradition that started with C.S. Peirce, for example, is an attempt to solve (or dissolve) skeptical problems by understanding *why* certain doubts, such as the Cartesian doubt in the reality of the world around us, must be "paper doubts" if we explain intentionality from a practical point of view. We cannot 'really' doubt them in any sense that is ever relevant to action, and from an attributivist understanding of belief states, that may be the only sense that really makes sense.

In somewhat similar fashion, philosophers inspired by Wittgenstein have argued that reasoning and criticism can only be understood against a background of uncontested knowledge, an epistemic horizon of the life form within which things make sense to us. It may be that our web of volitional beliefs must have a similar structure, and that we are bound to take certain ideas about our normative volitional attitudes for granted. Consider once again Hume's example of preferring the destruction of the world to a scratch on one's finger (section 8.3.2). Contrary to the sort of view defended by Alan Thomas (see section 3.2.1) I do not think a Wittgenstinian approach to moral knowledge can establish nonrelationalism. But I am open to the idea that some part of our knowledge in the relationalist sense, i.e. the knowledge of our volitional inner reality, must be beyond our doubt in a way that is constitutive of our being practical deliberators in the first place. On such a view, it would always be presupposed that we are 'in touch with our inner selves,' so to speak, and that this must be reflected in an uncontested background of volitional belief.

It is in this sense, then, that some alternatives to our practical beliefs, such as the idea that we might sacrifice the rest of the world in order to prevent a scratch, may be "unthinkable" for us, as Frankfurt has suggested. Nevertheless, this does not refute my argument in section 8.3.2 that unthinkability cannot explain opacity and disconfirmation. Instead, it presupposes such an explanation when we judge that an agent is mistaken for thinking what we find unthinkable. But with the Affective Pattern View of opacity and the Affective Response View of disconfirmation in place, we now have a framework that allows us to accommodate the idea of unthinkability in terms of epistemic resolvedness.

9.3.3 The Hazards of Passive Resolution

However, passive resolution of our practical beliefs is not always a good thing. On the contrary, it can be a very bad thing. Recall that the Nazis did not disconfirm their beliefs in part *because* they were so resolved in them and didn't take the alternatives seriously. As another example, consider Aldous Huxley's *Brave New World*. It depicts a society in which people are engineered genetically, and conditioned socially, to never question the roles they were meant to fulfill. They are prevented from accidentally encountering the sort of situations that might provoke any remaining disposition to feel unhappiness or discomfort with respect to the lives they are supposed to be living. More importantly, they are not looking for such situations. They are happy and satisfied. But the mechanisms employed to achieve this state of affairs strike us as highly unethical: the people destined to fulfill the 'lower class' roles were purposefully limited in their cognitive development, while being grown as fetuses in biological factories that replaced ordinary pregnancy. The idea behind Huxley's thought experiment is that people might accept this sort of thing if you raise them to be used to it and prevent them from disconfirming their assumptions. But the intuition that Huxley wants to invoke, of course, is that even though these people may seem happy, they are not *free*. Not in the sense of knowing, through critical self-examination, what they themselves want in the normative sense, and acting upon that (recall that I identified acting upon knowledge of one's normative reasons with freedom in the sense of 'deep' self-disclosure in section 8.4.2).

The problem with passive resolution, of course, is that once you are resolved in this manner, then from your own point of view, there is no reason to worry about the beliefs that you are so resolved in. That is what it *means* to be resolved in this sense. Just like you cannot decide to become fully resolved, neither could you decide to become unresolved when every possibility of doubt has left your mind. So how can we prevent becoming too resolved in too many of our beliefs if passive resolution is something that basically happens to us?

The answer, I think, is twofold. On the one hand, there may be situations in which there is little that someone could have done to become more cognitively flexible, given their social environment. In a sense the people from *Brave New World* are helpless. I realize that it may be troubling to draw a similar conclusion in the case of Nazi officers, or contemporary religious fundamentalists, for example, but I do believe that a lot of evil is

accomplished by people for whom the better alternatives were not even on their radar anymore. Passive resolution can be a matter of epistemic bad luck. But it is not *ineliminable* bad luck. Recall the example of the second SS-officer, who did not disconfirm his views, but could have done so in a different social environment. Passive resolution is not irreversible. It only means that from one's own epistemic point of view, one is not bound to take the initiative for critical re-examination. It may still be true that under the proper circumstances, a person who had been resolved could be lead to open himself up for disconfirming experiences again.

On the other hand, even though passive resolution is something that happens to us without our deciding for it to happen when it does, we can of course make an effort to develop methods and habits of critical thinking *in advance*, as it were, that will prevent us from easing into a state of full resolvedness too quickly. As I see it, this is one of the primary goals of our educational system. Nevertheless, finding the proper balance between a necessary amount of trust in one's epistemic background and a healthy amount of doubt with respect to one's unjustified assumptions remains of course one of the deepest problems of practical reasoning. In other words, it is a matter of virtue: we should attempt to avoid the extremes, and find wisdom in the middle. Self-evident as this perhaps might sound, it seems an important departure from the spirit of Frankfurt's 1987 essay on identification and wholeheartedness: decisive commitment can be as much an obstacle to one's freedom as a requirement for it. Frankfurt's later remarks about "keeping an eye out for the possible correction of our views" reflect a similar development in his own thinking.

9.4 The Facts Problem Solved

In section 4.4.1 I argued that the type-1 dispositionalist solution to the Facts Problem involved a *promise* that we could make sense of the idea that some motivational states have the authority to discredit others. I also argued that Williams's defense of the Internal Reasons View did not substantiate that promise. His account of the imagination as an important resource for deliberation about ends makes sense with respect to the "constitutive solutions" that we discussed above, which resolve indeterminacies in the subjective motivational set of an agent. But the problem was to explain the cases in which the facts about our normative reasons are already determinate in the light of a motivational set in which some elements

overrule others.

Frankfurt's distinction between desires that are internal and those that are external to the person offered us an intuitive concept of why some elements of the subjective motivational set might take precedence, in a volitional sense, over others. But although Frankfurt's solution of a "reality within ourselves" seemed to carve out the right middle ground between realism and relativism in matters of practical normativity, his theory remained somewhat evasive with respect to the nature of this inner reality as a matter of empirical fact. Furthermore, his Cartesian insistence on privileged insight stands in the way of a true account of volitional opacity.

By contrast, the Affective Pattern View that I have proposed in this chapter requires no experience, ever, to be epistemically privileged. In principle, every instance of being moved by desire, and every individual emotional response, is as good as any other. It is for patterns in our dispositions to have these responses that we must look in order to understand how we should live. This view answers the Facts Question in a manner consistent with how the Affective Response View answers the Disconfirmation Question. Thus, it allows that significant parts of our actual dispositions have not yet revealed themselves in our experiences so far and may surprise us in the future.

This idea forces another disambiguation of the Internal Reasons View. If the subjective motivational set is construed as a set of *experienced* or motivationally *effective* attitudes, then the Affective Pattern View would violate the internal reasons requirement. Because with respect to such attitudes, the Affective Response View requires the possibility of a disconfirming *discontinuity* in relation to pre-existing motivations. Instead, if we understand Williams's notion of the subjective motivational set as a set of *opaque* motivational dispositions that explain our motivations in future and counterfactual, as well as past and actual situations, then the Affective Pattern View is an Internal Reasons View. This is so because, at that level, disconfirmation presupposes *continuity*. To be sure, the view allows dispositions to change over time at that level too. However, in cases where such underlying changes offer the best explanation of discontinuities at the transparent level, the proper volitional interpretation of the unexpected experiences would be as evidence that one's normative will has changed, rather than as a disconfirmation of one's prior judgments.

10 *Intersubjectivity and Moral Discourse*

Now that we have seen how the Affective Response View solves the Disconfirmation Problem (chapter 7), and how the Affective Pattern View (chapter 9) provides a type-1 dispositional solution (chapter 4) to the Facts Problem, the question that remains is whether we can square the relationalist implications of these views with our common sense understanding of moral discourse.

I have already argued (in section 7.4.2) that my account of volitional interpretation is not solipsistic at all, in view of the ways in which a social practice of deliberation contributes to its purpose. Furthermore, I have been stressing from the beginning that relationalism is perfectly compatible with the idea that humans share some of the values they would uphold under ideal conditions of rational agency as a result of their common psychological dispositions (most notably in sections 3.2.1, 7.3, 8.1, and 9.1.4). In fact, this was the reason why I chose the term "relationalism" instead of "relativism" in the first place: the latter is too often understood as involving the denial of that possibility. But what relationalism merely rules out, remember, is that all conceptually possible deliberators—like the alien invaders from *Mars Attacks!*—would have to converge upon the same values as us. Finally, I already explained (in section 4.3) that the view is not even committed to the idea that all moral judgments made in practice are to be evaluated in relationalist terms.

The purpose of this final chapter is to tie these different strands of argument together into a comprehensive account of how to understand moral discourse. Since I have developed my relationalism in contrast to Michael Smith's nonrelationalism, I will now articulate this account of moral discourse in response to objections that he has raised against the relationalist alternative. As we shall see, these objections are meant to reflect and support his claim that relationalism violates certain "platitudes"

about moral judgment, by which he means "prima facie" truths that capture the "inferential and judgmental dispositions" of those moral language users that have "mastery" of the moral terms they employ (1994, pp. 30–31). Amongst these platitudes is the claim that "When A says that ϕ-ing is right, and B says that ϕ-ing is not right, then at most one of A and B is correct" (1994, p. 39), which, provided that A and B quantify over conceptually possible deliberators, rules out relationalism from the get-go.

So what are Smith's reasons for thinking that those, who have *mastered* moral language, cannot follow relationalistic inference patterns? The first, which I will discuss in section 10.1, is the idea that relationalism fails to explain the purpose of moral discussion and argument: if relationalism is true, then we would have no reason to engage in such discussions. I respond to this objection in sections 10.2 and 10.3. The second objection, to be discussed in section 10.4, is that relationalism fails to explain what we mean when we use moral terms in real-life conversations. As I will argue in section 10.5, this objection is based on a 'conservative' understanding of actual discourse as already incorporating meta-ethically sound moral concepts, against which I shall defend a 'revisionist' alternative.

10.1 The No-Purpose Objection

In addition to their rejection of "relativism," nonrelationalist realists often employ the language of "subjectivity" versus "objectivity" to distinguish their view from, and defend it against, the relationalist alternative. According to Smith, for example, the aforementioned platitude that at most one of two agents can be correct if one of them says that something is right while the other says about the same thing that it is wrong, is a platitude about objectivity:

> There are platitudes that give support to our idea of the *objectivity* of moral judgment: 'When A says that ϕ-ing is right, and B says that ϕ-ing is not right, then at most one of A and B is correct'; 'Whether or not ϕ-ing is right can be discovered by engaging in rational argument'; 'Provided A and B are open-minded and thinking clearly, an argument between A and B about the rightness or wrongness of ϕ-ing should result in A and B coming to some agreement on the matter'; 'The rightness of someone's ϕ-ing is determined by the circumstances in

> which that person acts, circumstances that might be faced by another', and so we could go on. (1994, pp. 39–40)

Recall that this idea of objectivity is one of the three premises that in Smith's framework give rise to the "moral problem." His formulation of this premise was as follows:

> We may summarize this first feature of morality in the following terms: we seem to think moral questions have correct answers; that the correct answers are made correct by objective moral facts; that moral facts are wholly determined by circumstances; and that, by engaging in moral conversation and argument, we can discover what these objective moral facts determined by the circumstances are. The term "objective" here simply signifies the possibility of a convergence in moral views of the kind just mentioned. Let's call this the "objectivity of moral judgement." (1994, p. 6)

Thus, for Smith, the postulation of facts that make our practical judgments true, and the nonrelationalist understanding of the content of those judgments, are both part of the same idea of their objectivity. They are aspects of the same feature of morality, and therefore presented as a kind of package deal. By contrast, in order to represent the logical option of my relationalist cognitivism, I have separated these two notions. On the one hand, I have referred to the postulation of facts as the Facts Principle (with the related assumptions concerning revision and discovery captured in the Disconfirmation Principle). On the other hand, I have introduced nonrelationalism as a possible interpretation of a different premise, the Intersubjectivity Principle, with relationalism as a possible alternative interpretation of this Principle.

Nevertheless, I think that Smith's association of objectivity with nonrelationalism is quite common among moral philosophers. Furthermore, and perhaps as a result of this, the relationalist alternative is often portrayed, and dismissed, as being implausibly *subjectivistic*. Consider Smith's discussion of "subjective vs. non-subjective definitional naturalism" (1994, pp. 41–43). Smith distinguishes between two ways in which conceptual analysis might be employed in order to identify the truth makers of moral judgments with matters of natural fact: a reductive or "definitional" approach, and a non-reductive approach which he himself favors. With

regard to the first approach of "definitional naturalism," Smith presents us with a dilemma that was originally raised by Alfred Ayer between subjectivism and utilitarianism, which Smith generalizes to a choice between "subjective" and "non-subjective" varieties:

> [A]ccording to the subjective definitional naturalists, "x is right" means "x has the natural property that is approved by so-and-so." Non-subjective naturalists, by contrast, focus in on the natural properties of acts that subjective naturalists say we merely approve of and define rightness directly in terms of one of those properties. For example, non-subjective definitional naturalists who are utilitarians think that we can define "x is right" as "x is conducive to happiness." (1994, pp. 41–42)

Smith's discussion of this dilemma is a bit complicated. First he agrees with Ayer that both options are unsatisfactory, but then he suggests that they are not exhaustive of definitional naturalism in general, because there is a further option of a "network style" analysis, as defended by Frank Jackson, which is also reductive, and hence "definitional." However, Smith then proceeds to raise an objection against Jackson's approach as well, concluding that Ayer was right to reject definitional naturalism after all, though "for the wrong reasons."

From a terminological point of view this may seem a bit odd, since "non-subjective definitional naturalism" is already Smith's own generalization of the view considered by Ayer, and a term that suggests "any old definitional naturalism except subjective definitional naturalism," so the reader may find it confusing when this dilemma is set up to be non-exhaustive. Nevertheless, what matters for us at this point is that both Smith's statement of non-subjective definitional naturalism, as generalized from Ayer's original discussion, and Smith's treatment of the network-style alternative are formulated so as to honor his nonrelationalist conception of the objectivity of moral judgment. By contrast, his objection against subjective definitional naturalism is precisely that it violates this idea of objectivity, and that therefore, it cannot account for the purpose of discussion and argument:

> [S]ubjective definitional naturalism is completely unable to account for either the objectivity of moral judgment or the various procedures via which we come by moral knowledge.

> For if desires are beyond rational criticism, as Hume thought, then the idea of a moral argument—an argument about the rightness or wrongness of an action, as opposed to an argument about the other non-moral features that might be possessed by an act—simply doesn't make a great deal of sense. An agent either approves of some natural property of acts or she doesn't. Either way there is nothing much to argue about; nothing to argue about in the way, and to the extent that, we argue about the rightness or wrongness of actions. Moreover, if another agent disapproves, then it simply isn't true that they express their disagreement with each other when the one says "This act is right" and the other says "This act is wrong." Rather, each self-ascribes their different pro- and con-attitudes, a self-ascription that the other can and perhaps should agree to be correct. (1994, pp. 42–43)

Whether this criticism is valid depends on how precisely we interpret Smith's characterization of subjective definitional naturalism. Recall that it is the view that "x is right" means "S approves of x" where S is whoever utters the judgment. But what does that mean? Usually, when we say that someone approves of something we mean that she has made a judgment in support of it, but on that interpretation the view would be that judging x to be right means something like "judging that you are judging in support of x" which introduces a weird self-referential structure that doesn't tell us anything. Instead, the criticism suggests that it means "having a pro-attitude" which might be a desire or something similarly non-cognitive. However, on the basis of our discussion concerning opacity in chapters 8 and 9 we can already see that the first part of the criticism, that "there is nothing much to argue about," would only refute the view if the pro-attitudes in question are fully *transparent*.

Instead, if we understand practical judgments as self-ascriptions of opaque pro-attitudes that we may be deeply mistaken about, such as the normative volitional attitudes that I have postulated, then, as we have seen, a lot of argument and reasoning may be required as a matter of volitional interpretation. Now perhaps Smith might be the first to admit this much, as his criticism might have been intended solely against the view that practical judgments are self-ascriptions of attitudes that are completely transparent to the judger. To be sure, such a view would deserve the label of "subjectivism." But in that case, his discussion of definitional

naturalism would have left out the middle ground of a relationalist inner reality theory.

Furthermore, even if Smith would admit that this middle ground does allow there to be "much to argue about," he might still insist that the second part of his criticism against subjective definitional naturalism, i.e. that it does not allow participants in a discussion to argue *against each other*, will refute volitional inner reality theories as well. The worry, then,—and I think this is shared by many moral philosophers who hold nonrelationalist views—is that relationalism undercuts the *purpose* of moral discourse. In order to address this worry, we must be able to construct this purpose in relationalist terms. Or, since it will turn out not to be a single purpose, we must explain the *reasons* that we can have, according to relationalism, for participating in moral discourse with each other.

10.2 Shared Psychology and the Intersubjectivity Principle

The first and most important reason should be familiar by now from my discussions in the previous chapters: when A judges in approval of P while B judges against it, then because of the opacity of their normative volitional attitudes, those attitudes might favor P in the case of both A and B. Furthermore, such a similarity of our volitional inner realities need not be thought of as accidental or incidental, but can instead be explained with reference to our shared genetic background, and when applicable, our common environment or upbringing. Therefore, if A and B cannot find specific reasons to suppose that their underlying dispositions with respect to P are likely to be different, their moral discussion concerning P may be founded upon their agreement that their normative volitional attitudes towards P are most likely the same, even though they disagree about whether those attitudes support P or not. I should now like to make a number of additional remarks about this idea, with my eyes on both the task of responding to the No-Purpose Objection, as well as the task of accounting for the intuition behind the Intersubjectivity Principle.

10.2.1 Volitional Similarity Judgments

To begin with, it might be objected that on the proposed analysis, the practical judgments that A and B express still do not strictly speaking contradict each other. After all, logically speaking their dissimilar self-

ascriptions of their attitudes towards *P* are still compatible, because it is at least conceptually possible that the normative will of *A* favors *P* while the normative will of *B* does not. It is only *in conjunction* with the additional contingent assumption that *A* and *B* have the same normative volitional attitude towards *P*, whatever that attitude turns out to be, that a contradiction can be derived from their different practical judgments. But even if *A* and *B* have agreed that this assumption is true, that does not make the assumption part of what they are saying when they utter their approval and disapproval of *P*. And that is an implausible consequence, the objection might run, because our intuition is that the practical judgments *themselves* contradict, regardless of further assumptions. Put differently, our intuition is that they are talking about the same thing, while my analysis would seem to imply that they are not.

My response to this objection is twofold. First, I think that in order to understand disagreement and discussion in everyday conversations, we must not focus on what is being said in such a strict sense in order to construct a contradiction that would explain the disagreement. Instead, we must look for the *conversational implicature* of each utterance in the course of the discussion. Consider the following example:

> John: "I don't think I should take another pain killer tonight, as I have already taken three earlier today."
> Sarah: "I have been taking those for years and going beyond the regular dose never killed me, John."

Aside from the fact that *dying* is probably not the side effect that John is worrying about, Sarah has on the strictest mode of interpretation said nothing about John at all. Logically, her claim entails a relation between the painkillers and herself, which is compatible with the absence of such a relation between the painkillers and John. However, saying that the painkillers never killed *her* is Sarah's way of telling John that he shouldn't be so sensitive about going beyond the official dosage once in a while. The assumption that John's body will be similar to hers in its response to the painkillers can be inferred from the context of the conversation. Thus it becomes part of the conversational implicature of her utterance.

By analogy, if in a moral conversation, *B* expresses her disapproval of *P* after *A* has just made a statement in support of it, the context of that conversation may include a similarity assumption that will turn the conversational implicature of *B*'s disapproval into a contradiction of *A*'s

approval. Regardless of the specific phrase that *B* may have used,[1] this implies that what *B means* to contribute to the conversation exceeds her practical judgment itself, on the relationalist analysis. And this brings me to the second part of my response to the objection: the relationalist is not committed to the view that practical judgments are the only judgments that we bring to the table in a moral discussion. Instead, we can construe what I shall call a "volitional similarity judgment" in approval of *P* as a judgment about a group or class of agents that they all share a normative volitional attitude in favor of *P*. Thus, often when a person uses moral language in a discussion, the implicature of his claim will be a volitional similarity judgment about a group of people, or possibly the class of all people, to which he himself belongs, such that the volitional similarity judgment will still also entail a practical judgment on his own part in the strict relationalist sense, while at the same time carrying implications for the practical judgments that the other members of this group ought to make.

Note that when *A* and *B* are expressing volitional similarity judgments concerning a group of people to which they both belong, where *A*'s judgment is in approval of *P* while *B*'s is in disapproval of it, we can even literally agree with the intuition that *A* and *B* are talking about the same thing: they are talking about how they might be similar in their normative volitional attitudes towards *P*. Of course, there is also a further intuition that nonrelationalists sometimes allude to, which is that *A* and *B* must not merely be talking about the same thing, but that this same thing they are talking about must be *P* rather than their attitudes towards *P*. However, this is a much stronger claim which no longer seems to be about establishing reasons for having moral conversations in the first place. To that end, at least, the claim that the participants in the conversation have the same normative volitional attitude towards *P* seems sufficient. The stronger intuition may sometimes apply as a claim about the surface grammar of moral discourse, which I will discuss in section 10.5.4 below. However, if instead it is meant really as a claim about the truth conditions of moral judgments, then it amounts to a statement of nonrelationalism itself, which means once again that it no longer qualifies as a premise of an argument against relationalism.

[1] A matter to which I shall return in section 10.5.4.

10.2.2 Species-Wide Similarity: Human Values

With this response to the objection that the participants in a moral conversation would be talking past each other (so to speak) in mind, let us now take another look at the Intersubjectivity Principle. According to the principle, the same considerations which justify *A*'s approval of *P* may, under the appropriate conditions, disconfirm *B*'s disapproval of *P* and require *B* to judge in approval of *P* instead. Note that this formulation is also consistent with the idea that both *A*'s approval and *B*'s disapproval are self-ascriptions. The "appropriate conditions" for the possibility that *A*'s justification disconfirms *B*'s self-ascription are simply those under which *B* has reason to believe that she is volitionally similar to *A*, either because she already had such a reason on independent grounds, or because she may recognize how the considerations cited by *A* apply to her own emotional life, or make sense of her own affective dispositions, as well, in a way that would challenge her current self-ascription.

How often would it be plausible to say that these conditions obtain? In sections 7.4.2 and 8.1 I have speculated that with respect to certain issues, volitional similarity may be a matter of empirical fact about human psychology as such, and thus apply species-wide, with possibly only a few individual exceptions. In section 9.1.4 I have argued in greater detail how such a species-wide volitional reality may be understood with respect to the Holocaust. We have seen that there are psychological reasons, supported by empirical findings, to believe that even most of the perpetrators of the Holocaust have been similar to us in the relevant respects, and must therefore have been acting against their normative will. I shall not repeat the argument here, but I want instead to follow up on it with some additional remarks, as my focus in chapter 9 was on establishing a strong case of opacity, whereas now it is on investigating the extent and scope of intersubjectivity.

If we assume that the Holocaust was indeed a violation of a species-wide volitional inner reality, then how much does that tell us about the extent to which a core set of moral values might be normative for human beings generally? The Holocaust is about as extreme in its immorality as our history has shown us to be capable of, and as most of us are willing or even able to imagine. From our point of view, then, the normative premises needed in order to disapprove of the Holocaust are fairly minimal, and would not settle much else in the way of moral theory or ethical practice. In the moral discussions we are used to having, its evil is not only universally

agreed-upon, but even the suggestion by one participant that the view defended by his opponent is in a relevant respect akin to Nazism will be taken as a grave insult.[2]

Be that as it may, the awful historic fact, however, is that despite its extremity the Holocaust has not been exceptional. The victims of atrocities in Rwanda, for example, or more recently in Darfur, and on a longer timescale in countless other displays of genocide, massacre, and massive and collective torture and rape, have as much experienced the hatred and malice that human nature can give rise to as did the prisoners in Nazi concentration and extermination camps. The philosophical discipline of ethics, including meta-ethics, should be as much about getting to grips with this phenomenon, as about settling or conceptualizing subtle disputes between sophisticated rival normative theories.

In the light of this, being able to solve the Facts Problem with respect to such atrocities in a way that establishes the normative reasons that even most of the perpetrators must have had to not have committed them, is in my view a significant result. I do not claim to have achieved this result, as my account has been programmatic in many ways, and as I am sure that many counterarguments have yet to be dealt with. But my point is that *if* a theory delivers this result that I have attempted to demonstrate, then that result would be considerable.

Furthermore, I do think that we can be optimistic about a more inclusive hypothesis concerning what values human beings generally would support under ideal conditions of rational agency—what I have called "human values." I think these may include such things as school and play for children, freedom from oppression and torture, standards of health and nutrition, and so on. These values may constitute a psychological reality that I think is presupposed in the justification of political ideals and principles that purport to be "universal" in some sense, such as the Geneva Convention or Universal Declaration of Human Rights. I am not saying that these specific documents get it right exactly, or that the extent of species-wide volitional similarity would cover the level of detail that such documents tend to go into. But I am optimistic that various dispositions which we may share as human beings pertain to their subject matter.

[2]In many Internet communities, arguing that one's opponent is wrong by analogy to Nazism is known as a "Godwin," from Godwin's Law which states that the probability of a comparison involving Nazis or Hitler approaches 1 as an online discussion grows longer. Committing a Godwin is considered to be a fallacy that immediately causes one to have lost the discussion as a matter of general etiquette.

10.2.3 Psychopaths and Other Anomalies

My next remark concerns the status of the relatively small amount of human individuals that would still not be volitionally similar to us even in the aforementioned respects. As I noted during my discussion of the Holocaust in section 9.1.4, even if it is empirically plausible that most Nazi soldiers would share our normative volitional attitudes, it is implausible to think that there would be no exceptions. Psychopaths may be an example of this, but we should not equate deviation with pathology in this context. For the sake of the argument, let us assume that there are some "successful" individuals out there who are similar to dysfunctional psychopaths in their lack of a disposition for mercy, but who suffer from no symptoms and who manage to get along and cover their tracks. The consequence is that no proceduralistic deliberative route would lead them to disconfirm whatever crimes they commit. Whether their abnormal psychology has been caused by some genetic anomaly, a highly unusual environment or upbringing, or what ever may have been the case, is irrelevant in this example. The question is, would such individuals be counterexamples to the claim that there are human values?

And the answer, I think, is no, at least not in any reasonable sense of the term. By analogy, consider other exceptional deviations within our species from our regular biological and psychological properties. It is a basic fact about human biology that we are born with ten fingers and ten toes, even though cases of *polydactyly* occur in which babies are born with more than ten, and even though having twelve toes need not be pathological in the sense of posing a clinical problem to the person who has them.[3] Furthermore, in contexts where it is obvious that the subject matter involves humans and not other animals, we can even leave out the reference to 'humans in general' altogether and say something like "there are five toes on each foot" even though this would be false for many animal feet. The same goes for statements such as "500 mg of paracetamol will not cause stomach problems," "the liver is located in the right side of the body" or "the prefrontal cortex is involved in mechanisms that cause merciful behavior": there may occasionally be a human being for which one of these statements would be false, and they do not apply to non-human

[3] Polydactyly has a prevalence of one in every 500 births, but in most of those cases the additional digits are not fully functional and may therefore be understood as pathologies with reference to the 'design' of having 10 digits. However, rare cases where the sixth digit on each hand or foot is fully developed and fully functional *do* occur.

animals generally.

Of course, when a human being has been born with his liver on the left side, he cannot be treated on the basis of the assumption that his liver must be on the right side.[4] And when it is known beforehand, the surgeons better take it into account before they start cutting him open. In this case, the adjustment is straightforward, but I suppose there can be biological anomalies that would make common medical tools or practices ineffective or nonsensical to use on the patient in question, even though the anomaly may not be dysfunctional in itself. A similar breakdown might be expected, I suppose, if one were to have a moral discussion with a psychopath (or his more successful counterpart). However, that is hardly an implausible implication of the theory: most of us have never been in such a conversation, and conversations with incarcerated psychopaths that have been recorded are certainly out of the ordinary in moral respects.

In a nutshell, then, my view is that where human values are involved, the statements we utter in moral discourse can be just as objective and intersubjective as those we utter in medical, psychological, or biological discourse. This analogy even extends to the surface grammar of such statements: the expressions that "the liver is located in the right side of the body" or that "500 mg of paracetamol will not cause stomach problems" hide the reference to the class of organisms to which they exclusively apply in the same way that the expression that "genocide is wrong" may hide an implicit relationalist reference to the class of deliberators for whom this claim would reflect their inner normative reality. I will return to this point in section 10.5.4 below.

10.2.4 Intracultural Similarity

Let us now go back to the question of when the conditions under which the Intersubjectivity Principle applies would obtain. So far, I have argued that they obtain when human values are involved, because then volitional similarity is guaranteed among all participants in a moral discussion, the rare psychopath being the negligible exception to the rule. But would it be plausible to think that all moral disputes could be settled in such a species-wide manner? I myself am not inclined to believe this: it is not empirically

[4]There is a condition known as *situs inversus* in which all internal organs are positioned in the opposite lateral locations compared to normal human beings. If the inversion is 'total,' so to speak, then the condition need not pose any problems, and may remain undetected until the person in question is examined for an unrelated medical matter.

unlikely that the development of certain normative volitional attitudes will depend on aspects of the individual's environment and upbringing that vary across different cultures, for example. Thus, while I think that certain moral questions will have answers that are universally valid for human beings generally, I accept that the correct answers to other moral questions may be different for people with different cultural backgrounds.

However, note that this does not make me a cultural relativist in the familiar sense that the correct answers would be those that accord with the views or values commonly *held* within or *definitive of* a culture. My view is still that they must instead accord with opaque attitudes of the individual, and even if it is true that these would have been different should his cultural upbringing have been different, then that doesn't imply that they must be straightforwardly identical to the views or values the individual was raised to hold as part of that upbringing. Rather, they might depend on cultural teachings in more subtle ways. Furthermore, who knows what other environmental factors might have a developmental influence that would be relevant in this respect: perhaps basic living conditions, the amount of violence or disease people are used to seeing around them, or even biological and behavioral facts about the mother during pregnancy. The bottom line is that even if people in different cultures have their own different normative realities, they might still be getting those realities wrong. And that is not how I think cultural relativism is usually defended.

In fact, this picture allows us to account for a whole further range of situations under which the conditions for the Intersubjectivity Principle obtain. Because if many correct moral answers depend on cultural aspects while remaining opaque in the sense of not being identical to the views *held* within a culture, that means there can be a lot of *intracultural* volitional similarity between members of some culture who *disagree* with each other about the moral questions they have similar normative volitional attitudes towards. More simply put: they could defend different volitional similarity judgments against each other about people within their culture.

Of course, cultures are hard to individuate, the boundaries between them are vague, and so is the very idea of belonging to the same culture. Perhaps it would be better to speak about the 'amount of shared cultural background' that any two individuals might have with each other. Then my claim would be as follows: even if human values may not be forthcoming on certain moral issues, then the conditions which make the Intersubjectivity Principle applicable will still obtain in many of the moral

conversations we are actually used to having, because of the amount of cultural background we tend to share with the people that we have such conversations with.

Once again, the hard-core nonrelationalist may find this unsatisfactory. But what he cannot deny is that this argument supplies the relationalist with a reason for us to have moral conversations in such cases, which means that in such cases, the objection that moral discourse would be without purpose has been defeated. What the nonrelationalist *can* do, by contrast, is question whether these cases of volitional similarity are really as common as I have suggested. In particular, there are two possible counterarguments which I think my account needs to address.

The first is that shared psychology with respect to the development of normative volitional attitudes, either across the human species or within a culture, need not always give rise to volitional similarity because it might also merely lead to *volitional isomorphism* instead. This line of criticism basically takes my own argument against the nonrelationalist constitution strategy and uses it against me.

The second argument concerns cases where the shared psychology itself is absent, either because it is a conversation between participants with very different cultural backgrounds, or because it concerns a moral question with respect to which we might expect affective patterns to differ significantly even when cultural backgrounds are more or less the same. Since it seems plausible to say that we have a reason to participate in moral conversations in such cases as well, the objection against relationalism still stands. I will now discuss these two arguments in turn.

10.2.5 Volitional Isomorphism Instead of Similarity

Recall the distinction between practical beliefs that are *similar* and those that are *isomorphic* from section 1.3.2: if Barack Obama and Mitt Romney both believe they should be president, then their beliefs are isomorphic but not similar, whereas if Mr. Obama and his senior campaign advisor both believe that Obama should win the presidency, then their beliefs are similar, but not isomorphic. Let us also say that normative volitional attitudes of different agents are similar when they make similar beliefs true, and that they are isomorphic when they make isomorphic beliefs true. Thus, if a wife and husband both prefer the wife to have a full-time job and the husband to stay at home to care for their kids, then

their attitudes are similar, whereas if both would prefer themselves to stay at home and the other to have a full-time job, their attitudes would be isomorphic. As this example illustrates, volitional isomorphism may involve psychological similarities while volitional similarity may actually be a result of psychological *dissimilarities*: it is because the spouses in the former case are psychologically inclined towards different roles that they can complement each other in a common arrangement, while the fact that the spouses in the latter case are psychologically of the same type, so to speak, requires them to negotiate an arrangement that will be less than ideal for one or both of them.

Given this insight, it may no longer seem plausible to think that as a general rule, psychological similarities between all human beings will lead to volitional similarities between them. Instead, they constitute isomorphic relations which may just as easily lead to conflict: when every human being favors the interests of her own children, her own family, her own tribe, country, or ethnic group. History shows us that at least at the level of our *cognitive* volitional attitudes this phenomenon has been undeniable, so what reason do we have to suppose that at the level of our *normative* volitional attitudes, things would be so different?

A parallel may be drawn between this problem and the problem I raised for the constitution strategy in sections 3.2.2 and 5.4.2. We have seen that if a disposition which makes some practical belief true is constitutive of being an agent, then all agents will exhibit isomorphism with respect to that belief. But the only plausible candidates for such dispositions that we have been made to consider seem to involve a crucial indexicality: in order for some organism to succeed at being an agent it must want its own attitudes to be means-end coherent, perhaps, or its false beliefs to be corrected, but from that we could not derive an interest in the rationality or knowledge of other agents. And it kind of made sense to expect that any interest constitutive of agency would be likewise self-directed and therefore ill-suited to breach the gap from isomorphism to similarity.

At this point, the nonrelationalist might be tempted to say that if the constitution strategy fails for this reason, then my case for human volitional similarity on the basis of shared psychology must fail for the same reason. However, this objection would misrepresent my argument from shared psychology. I am not saying that, merely because general features of human psychology are likely to be constitutive of our normative volitional attitudes, those attitudes must therefore exhibit similarity. I am not even

trying to derive or construct particular human values from known human psychological properties. Instead, I am arguing in the opposite direction: I start from certain values that many different people seem to uphold, and then argue that given the fact that much of our psychology applies species-wide, the view that Nazis who do not uphold these values are getting themselves wrong may offer a more plausible explanation of their behavior, psychologically speaking, than the view that their inner normative realities are radically different.

Furthermore, we should note that some of these values do not even seem to involve the sort of indexical structure that makes the dissociation between similarity and isomorphism possible. For example, the idea that some people go through intense suffering is something that abhors me, but so does the idea that nonhuman animals that are capable of suffering do so as well. And if I were to believe that other types of agents are suffering intensely, that suffering would abhor me just as well. Just *there being a subject of intense suffering* is something that I value negatively, without this involving any indexical reference back to me. Hence, whenever I have reason to believe that the psychological feature which implements my normative volitional attitude against such suffering is shared by someone else, I can make a volitional similarity judgment against suffering about us both.

Now I do not know how I came to have this attitude, neither ontogenetically nor phylogenetically. Perhaps it will be explained one day by evolutionary psychologists, and then again, perhaps it never will, as its evolutionary history may be too complex or at crucial turns too accidental for us to uncover. Perhaps it will turn out that such attitudes are an almost inevitable by-product of the evolutionary processes that best explain our level of intelligence and social cognition. If that were true, then the Martians that we have been talking about would be an unlikely product of natural selection, but that still would not make them conceptually impossible. Even though I am against suffering wherever it occurs, I see no contradiction in the idea of an agent who has no reason to care about the suffering of others or of members of a different kind. My argument against the constitution strategy is meant to explicate why. It shows that my attitude against suffering is essentially *contingent*. But given that I have this attitude, it can be very plausible to think that it is contingently shared by all human beings in the light of what we know about our psychology.

Of course, it does not follow that *all* normative volitional attitudes

that we have as a result of shared psychological features will be similar rather than isomorphic. Some of these attitudes will involve the sort of indexicality mentioned above. For example, most people care more about their own children than about those of others, and this may well be in accordance with their volitional inner reality. By itself, this might not rule out normative similarity: for it might be that all would value the state of affairs in which each parent takes a special interest in their own child, in a manner that is analogous to the sense in which similarity is compatible with everybody keeping their own promises or pursuing their own tastes. Smith discusses the following example:

> Suppose you are standing on a beach. Two people are drowning to your left and one is drowning to your right. You can either swim left and save two, in which case the one on the right will drown, or you can swim right and save one, in which case the two on the left will drown. You decide to swim right and save the one and you justify your choice by saying "The one on the right is my child, whereas the two on the left are perfect strangers to me." (1994, p. 169)

He then goes on to argue as follows:

> [I]f I had been standing on the beach instead of you, and if the one on the right had been my child, then surely I too would have been able to justify the choice of swimming right and saving the one by saying "The one on the right is my child." Indeed, if we think that a parent who fails to save their child in such circumstances fails to act on a reason available to her—as it seems to me that we do—then we are in fact obliged to say this; obliged to assume the non-relative conception of normative reasons. (p. 169)

But this argument actually falls short of Smith's conception of normative reasons. The mere fact that I would do for my child what you would do for your child under the same circumstances only establishes isomorphism. What Smith's conception of normative reasons requires, recall, is that in order for someone to have a normative reason to ϕ under circumstances C, the ideal selves of all agents would desire her to ϕ under C, and this implies similarity, not isomorphism. But suppose that while you are the parent on the beach whose child is drowning on the right, I am actually

the father of the two children on the left (and I am not near enough to save them myself). To say that you have a normative reason to save your own child under these circumstances, on Smith's conception of normative reasons, entails that my ideal self must also desire you to save your own child, *rather than mine*, just like I may desire you to keep your promises rather than mine. But in the drowning children case that implication seems false: if I care most about my own children, then it makes sense for me not only to desire that they will be saved at the expense of others by me when I am in a position to do so, but that they will be saved at the expense of others regardless of who will do so.

Suppose that the parent who is in a position to save the children has the sort of utilitarian view that tells him to save my two children in such a case at the expense of the single child of his own. Now if I were to think that he is mistaken in his view and that he actually has an all-things-considered normative reason to save his own child, then I will be *glad* that he is so mistaken. I certainly would not pick that moment to argue with him if I could. It seems that I can have a normative reason to keep it so that he will not act upon his normative reason. But note that if we find this intuitive, that means we actually find the lack of normative volitional similarity in such cases the more intuitive outcome, which means it would no longer be a problem for the relationalist at all that his view would make such an outcome likely.[5]

Furthermore, note that the absence of similarity would not make moral discussion about such a case impossible. For suppose that this other parent and I are merely discussing the example as a theoretical possibility. And assume once again that he and I are psychologically isomorphic in the sense that our ideal selves would each desire our own children to survive at the expense of the other's children. If I am making a volitional judgment reflecting this fact about me, while he is making a mistaken utilitarian volitional judgment that gets his normative volitional reality wrong, then it seems that he and I can also make contradicting *volitional isomorphism judgments* about the both of us, which explain our disagreement with

[5]On the contrary, this seems like a promising argument *against* nonrelationalist dispositionalism. Rather than merely showing that relationalism can account for moral practice just as well as nonrelationalism but at a lower metaphysical or epistemological price, we now have a case that is *better* accounted for by relationalism, challenging the nonrelationalist dispositional analysis *directly* in addition to its implausible metaphysical or epistemological implications. A further development of this argument is beyond the scope of this chapter, but it deserves a full statement that I intend to give elsewhere.

reference to our assumed normative volitional isomorphism.

10.2.6 When the Relevant Psychology is Not Shared

However, and this brings us to the second objection, what about cases in which there is neither similarity nor isomorphism due to the absence of any shared psychological features pertaining to the moral issue in question? We have already seen that a substantial amount of shared psychology underlying the common ground in many of our every day moral conversations may be due to shared cultural backgrounds rather than common biological ancestry. Therefore, intercultural conversations may be more difficult to accommodate. And even in intracultural cases, psychological differences may result in volitional dissimilarities.

Let us actually start with the latter type of case. Consider once again the discussion about choosing between saving your own children or saving more other children. Now suppose that both participants agree that it is permissible, say, to save a school bus containing 34 children including your own instead of another bus containing 35 children, assuming that you cannot save both. Suppose also that both agree that it is no longer permissible when the choice is between saving 1 child of your own or 10,000 other people.[6] But they might not agree about where to locate the cutoff point between these two extremes. And given the Affective Pattern View it seems they wouldn't have to. For the two imperatives in this context, the imperative to save your own children first and the imperative to save as many lives as possible, may be understood as rivaling patterns in the web of our affective dispositions. When the difference in amount of children saved is 35 vs. 34 the first pattern may dominate the second for most of us. When it is 10,000 vs. 1, it is probably the other way round. So far, the two participants may be isomorphic. But the precise point at which the one pattern begins to overtake the other is likely to differ a bit from agent to agent given the sort of view I have defended about the nature of the normative will.

In fact, on the basis of that view it is first of all most likely that there will not be a precise *point* for *any* human being at which the situation would suddenly 'flip.' Instead, we should expect a gray area in which the two patterns are of similar magnitude to such an extent that the truth

[6] Although somewhat theoretical, this sort of thought experiment does seem to appeal to our imagination, as popular television series like *24* present us with people having to make that choice every hour.

conditions remain indeterminate. Nevertheless, the region in which we encounter that area may still vary from individual to individual. Perhaps some of us would already be in that area when the choice is between one child of their own and three other children, while others would definitely be getting themselves wrong if they gave up their own child under those conditions. Let me stress that I do not mean to make any claims concerning the specific numbers in this context. My point is entirely about the sort of normative *structure* that we may expect, and can account for, on the view I have proposed.

Although this analysis confirms the implication that intracultural normative volitional dissimilarity is likely to occur, I think it nevertheless also undermines the objection somewhat. Because even though we are used to having moral discussions, I am not sure that we are also used to establishing, through reasoning and argument, where an intersubjectively valid cutoff point should be drawn when it comes to weighing the amount of sacrificed loved ones against the amount of saved strangers. Instead, what we do know from experience, I think, is that even though we may agree on the general values at stake, and even though we realize that this means we sometimes have to weigh these values against each other in particular cases, the task of determining a kind of exchange rate between loved ones and strangers strikes us as patently absurd. Rather than clear and precise cutoff points or exchange rates, I think we experience exactly the sort of vague and gray area that the Affective Pattern View predicts. Furthermore, when a moral conversation leads us into such an area, I think most of us experience the limits of reasoning and argument. It is at that point that we turn to intuitive judgment and say things like "you have to follow your conscience" or "decide what feels right to you." We may agree on this method, but I do not think our everyday experience with moral conversations carries an implicit commitment to the idea that normative similarity must be guaranteed in such instances.

Nevertheless, I do not think this is sufficient to counter the objection completely. For we do not always agree on the values at stake in every particular discussion, not even when all participants share most of their cultural background. Consider debates surrounding freedom of speech. Some people think there is a trade-off between freedom of speech and such things as nondiscrimination, respect and decency, or freedom of religion in some cases. Others believe the freedom of speech to be absolute, and take the issue of judgment rather to be about when a speech act involves

more than just speech, such as when it is a call for action on the part of others. And in the Netherlands, "freedom of speech" is not even the phrase that we tend to use or that our constitution protects. Instead, we have a "freedom of opinion expression" which some consider to have a more narrow application. Now these sorts of differences we do experience, I think, to be matters of reasoning and argument. And given the subtleties involved, my relationalist proposal might not make it very plausible to think that there will be no room for the affective patterns in our emotional lives to vary from individual to individual with respect to these matters as well. And so here we really have a case where the objection does seem to kick in: without normative similarity, what would be the reason for us to discuss these matters with each other?

Finally, in the intercultural case these differences become even more pronounced. Sometimes we do enter into moral conversations with people of very different cultural backgrounds. Suppose we talk about the importance of "honor." If relationalism is true, then it might be that not only the practical judgments, but also the truth conditions for those judgments, are going to vary significantly from culture to culture when it comes to things like the honor of the family, or honor as a reason to make personal sacrifices, and so on. The relationalist must be committed to the view that it is an open empirical question to what extent volitional similarity will happen to be the case on such matters. But when it is absent, would there then still be a reason to engage in intercultural moral conversation?

My answer to this question is going to be yes, both in cases where much and where little cultural background is shared. In fact, there is not just one such reason, but there are several. In section 10.3 below I will articulate each of these reasons and explain their soundness even when normative volitional similarity is absent. Now perhaps this might seem as if I am changing my strategy: first I tried to account for the purpose of moral discourse by making such similarity seem plausible, and now I am going to say that such similarity is not needed in order to explain the purpose of moral discourse. However, these two lines of argument are meant to be complementary. With respect to moral discourse I think there are two intuitions that we need to account for. The first intuition is that it makes sense to us to have moral conversations, which means that an intuitive account should explain the purpose of such conversations. The second intuition is the more specific idea, captured by the Intersubjectivity Principle, that considerations which give you a normative reason to support

P may disconfirm my practical judgment against *P*. When normative volitional similarity is present, then both intuitions are accounted for, because the sort of intersubjectivity that the second intuition requires is sufficient as a purpose of the sort that the first intuition requires. However, what I think is intuitive about the Intersubjectivity Principle is that this sort of intersubjectivity *does exist*, which means that it must play a role in *some* moral conversations. But that does not imply that it will play a role in *all* such conversations for which there is a purpose. Which means that we may try to come up with additional reasons to have moral conversations that are less demanding in the sense that they do not require normative similarity, while having a wider application, in the sense that together with the reasons in virtue of normative similarity, they can account for the purpose of all moral conversations that seem intuitively purposeful.

10.3 Additional Reasons We Have to Discuss the Reasons We Have

I will discuss four types of reasons that we may have to conduct moral conversations that do not depend on us contingently sharing psychological dispositions (pertaining to the moral issue under discussion) as a matter of empirical fact. The first is that participants may improve their self-understanding by *contrasting* themselves with each other. The second is that other people may sometimes know us better than we know ourselves. Third, moral discourse increases our awareness of the *alternatives* that we might choose from. And forth, moral discourse could still focus on certain principles of reason that, even though they cannot establish ethical *similarity*, might make certain volitional *isomorphism* judgments conceptually necessary.

I turn to each type of reasons below. Together with the type of reasons that I discussed in the previous section on the basis of shared psychology, the *five* types of reasons that we thus end up with provide the relationalist with a purpose for moral discourse in a wide range of cases. Note, however, that these reasons are meant to refute the objection that there would be no purpose for moral discourse if relationalism were true. It does not follow, and I am not claiming, that all participants of actual conversations always have these reasons *in mind* when they talk about ethics. I return to this latter issue in sections 10.4 and 10.5.

10.3.1 Contrast

As we have seen in sections 1.3.2 and 5.1.2, there is no real disagreement between relationalism and nonrelationalism about differences between people as a matter of personal taste. If some people like baseball and others like tennis, then the nonrelationalist can say that everybody should approve of the state of affairs in which those who like baseball play baseball and those who like tennis play tennis, and the nonrelationalist can agree with the relationalist about whatever contingent psychological fact it is that establishes whether someone likes baseball or not.

It is sometimes said that one cannot argue about taste. And in some cases, this might be true: I like the taste of coffee, for example, but I do not have arguments that establish why coffee tastes good, or why others should like the taste of coffee as well. As far as I'm concerned they don't have to. Nevertheless, the range of cases for which the nonrelationalist must invoke the above construction also includes less trivial matters of preference, such as whether joining the Army is right for you or whether you should revise your decision to study philosophy and enroll in the chemistry curriculum instead. As I have argued in chapter 9, the attitudes that determine what we should choose in such cases are opaque and deliberating on them may require lots of reasoning. Furthermore, when we face such personal choices, we do tend to discuss our options with the people around us, even if we know that they do not share our preferences in this matter. It is not useless for a person who wonders whether he should join the Army to discuss this with someone that already knows he wants to, or with someone who would never want to do such a thing himself. Suppose that the doubting person realizes that he shouldn't join the Army by coming to see how he differs from the guy that already made up his mind. In such a case, it is by *contrasting* himself with the other guy that he better understands his own reasons.

So far, the relationalist and the nonrelationalist can agree. But given that it makes sense in the nonmoral case, the relationalist can now argue in similar fashion about moral cases where normative volitional similarity is absent. Thus, if relationalism is true, then a moral debate may help a person to explicate and understand his own moral values better by contrasting them with the views put forward by others. We can understand the reasons based on similarity and contrast as complementary ones: insofar as people are alike, they can learn from each other by exploring what they have in common, and insofar as people are different, they can identify their own

normative volitional attitudes by investigating how they are different from each other.

A moral debate often begins with different initial views about a concrete case and then takes the form of a search for the source of the difference. In this search, participants try to explain to each other why they have the practical beliefs that they have. It is possible that that will reveal fundamentally different emotional response profiles, so that each participant can understand why the other judges differently. However, explaining how you arrive at your moral judgments to someone else, especially if that person does not share your moral intuitions, will require you to explicate your reasons very precisely. In the light of such a demand, you may discover weak elements in your reasons, and be forced to revise your volitional beliefs. And even if your initial judgement about the case at hand remains unchanged, your self-understanding may have improved, and the degree of conviction in your volitional beliefs may have increased.

10.3.2 Knowing Someone Better Than He Knows Himself

Sometimes we say that someone knows us better than we know ourselves. People who live close to me for a long time may observe things about me that I never noticed myself, or never realized were distinctive of me as a person, simply because I am so used to being me. Discussion with such a person may help me discover a pattern so obvious that I missed it all the time. I may be in doubt about something, or lost in false presumptions about myself, or hiding (unconsciously) behind excuses, when another person may tell me that what I am about to do is, in her opinion, not what I really want. Another example of such a case would be when a therapist understands the predicament of her client better than the client himself, for example because it may be part of his problem that he misunderstands his own psychological state. And although the therapist may be relying to some extent on her knowledge of general human psychology, she will also observe the ways in which this client differs from other people with a sharpness that even his closest friends or relatives might lack.

Now as I see it, this reason for conversation cuts across the similarity–dissimilarity distinction. The things that my friend observes in me may also apply to himself, or they may not. The point is that he observes them *in me* and that I can learn about his observations by talking to him. And this goes for moral cases as well. Perhaps in some moral cases the person

who knows you better than you know yourself is not going to tell you about it if the reasons she thinks you have conflict with her own. But not all cases are like that, even if volitional similarity is not a given. For example, my friend might face a moral dilemma so unique to his situation that I will never have to face it myself, and which does not concern my own interests as far as I know them. I may not have a clue what *I* would want him to decide, and I may be happy that I do not have to make that call. Nevertheless, he may want to talk to me about this problem, and I may share my ideas about what I think matters to him in a way that can contribute to his deliberations. In order for such a conversation to be meaningful, it does not seem to be required that we assume normative volitional similarity.

This especially applies to dilemmas involving the sort of trade-offs and gray areas which I discussed in section 10.2.6 above. As we have seen, volitional variations from person to person on moral issues are especially likely in such cases. I already noted that we can experience the limits of reasoning and argument when moral discourse touches on such gray areas. Nevertheless, even if my friend and I are not strictly volitionally similar with respect to the dilemma he is facing, and even if neither of the values that he must weigh against each other clearly outweigh the other, then he might still want to talk to me about his decision and I might still try to sort of reflect back to him how I see this problem affecting him emotionally in order to help him get through it and make his decision.

10.3.3 Increase of Alternatives

There are certain obstacles to volitional interpretation that we have already discussed, which may be overcome by engaging in moral conversation. In section 7.4.2 I referred to the psychological fact that human beings exhibit a strong self-confirmation bias, not only in the sense that we are masters at ignoring inconvenient evidence, but also in the sense that our experiences themselves are influenced by our beliefs regardless of whether they are true. Thus, even though the agent *would*, under the appropriate conditions, experience an unexpected affective response to the intended consequences of his action that would disconfirm the practical belief he acted upon, and even if his disposition to have that experience makes it the case that the belief is false, then in actual practice his current self may not experience any such response whatsoever. That is because all his experiences are

'following the lead,' so to speak, of his false practical belief. But one factor that may alter this situation, I argued, is whether the agent is made to consider alternatives to that belief by his peers.

In chapter 9 I developed this line of thinking a bit further with respect to the Nazi example, which is a case involving species-wide shared psychology, or so I have been arguing. But given that we know about these mechanisms of dogmatism and self-confirmation, we may now assume that these same mechanisms can play a role in cases where normative volitional attitudes vary from individual to individual. If my father really wants me to do *X*, then my own stubborn belief that I want to do *Y* may prevent me from having the experiences that would lead me to discover that I really want to do *Z* instead. And talking to my father may help me to overcome this problem, even if *Z* differs as much from *X* as it does from *Y*. Perhaps talking to him will temporarily convince me that I want to do *X* but without the stubbornness with which I used to believe that I wanted to do *Y*, so that I am now open to the experiences that will ultimately teach me that I want to do *Z*. Or perhaps my father simply discusses all the options with me, explaining the reasons that would justify, from his perspective, *X* over both *Y* and *Z*, thereby *introducing* me to the *Z* alternative as well.

Of course, the point is not just about being introduced to an idea in the sense of simply never having heard of it before or never having thought about it yourself. That seemed to be George Orwell's notion in *1984*: the language of "Newspeak" was meant to prevent the future citizens of Oceania from even formulating the views and ideas they were not supposed to be having. By contrast, in *Brave New World*, which I discussed in section 9.3.3, children *are* being thought about the awkward customs of their ancestors, such as couples having long-term relationships and raising their own children as parents, but of course they are being taught to respond to the idea of such a custom with horror and extreme embarrassment. And to go back to the real world, it is not as if the Nazis didn't know about liberal democracy. But they probably made fun of it amongst each other.

So the point is really about taking alternatives *seriously*. And this also applies in cases that lack normative volitional similarity. For example, let us assume that as an English speaker, Mark of course knows about vegetarianism. He knows what the word means, and he knows there are people who do not eat meat. But if those people are always ridiculed in

his peer group, he might not take the alternative seriously. Now suppose his new girlfriend Janet turns out to be a vegetarian, and he finds himself defending her against his friends. It might be that, as a result of this, Mark eventually figures out that even though he does want to keep eating meat, he no longer wants to buy meat from the factory farming industry. Now it might be that Janet would really want no animals ever to be killed by humans for food, in which case there is a dissimilarity between her and Mark. Nevertheless, through his engagement with her ideas, Mark got to correct his practical judgments.

10.3.4 The Discovery of Isomorphic Principles of Reason

I have argued in section 5.4.2 that the constitution strategy for defending nonrelationalist dispositionalism has failed to demonstrate that it can lead to similar rather than isomorphic desires under ideal conditions of rational agency. That also means I have allowed, for the sake of the argument, that it *can* lead to isomorphic desires. In fact, I had already argued in section 3.3.2 that my Distinctness Principle is consistent with Williams's idea that certain desires or principles of deliberation may 'come for free' with the idea of self-governing agency. Now as we have seen in section 10.2.5 above, even in cases where shared psychology does not lead to similarity but merely to isomorphism instead, what is being shared may still be the focus of rival *volitional isomorphism judgments* that contradict each other in a moral discussion. But since I have already allowed that certain principles of reason may be isomorphic for all agents as a matter of conceptual necessity, we may now conclude that such principles can also be a target for volitional isomorphism judgments in moral discourse.

The most obvious example of such a principle is of course that of means-end coherence. This principle is a requirement of rationality for any agent, but only in the isomorphic sense: I should desire my attitudes to be means-end coherent, you should desire your attitudes to be means-end coherent, but I am not rationally required to desire that your attitudes be means-end coherent. The principle itself is never questioned in actual discussions, but a person's moral views, or a well-known system of moral views (such as those associated with a religious tradition or summarized in the program of a political party) can be criticized in a discussion for being incoherent in this respect.

Other principles pertaining to the coherent organization of one's atti-

tudes may be constitutive of self-governing rational agency as well, perhaps involving considerations of prioritizing and scheduling various intentions and plans through time. And if so, then again our views and plans may be criticized for failing to comply with such requirements. In that case, the validity of such criticism need not depend on the criticizer having similar normative volitional attitudes, or even *contingently* isomorphic attitudes as a matter of empirical psychological fact.

Finally, a certain willingness to reflect upon our imperfections and the limitations of our faculties may be a rational requirement for any conceptually possible agent, since a perfect agent with unlimited capacities might well be conceptually impossible.[7] Thus, it may be that the volitional isomorphism judgment that every agent has a normative reason to be critical with regard to his own practical beliefs, which I briefly discussed in sections 5.4.3 and 9.3.3, is true for agents across all possible worlds.

In fact, the reasons for having moral conversations that I have been discussing so far may be considered, in their most general form, as isomorphic requirements of this sort. Since every agent must be critical of his own volitional beliefs, every agent may benefit from discussing those beliefs with his peers. However, as I already explained earlier on, the problem with the idea that self-criticism is a requirement of rationality is that *insofar* it is rationally required, the resources that we are supposed to commit to this task remain undetermined. For it depends on the contingent extent of my limitations, as well as the contingent opportunities that my peers provide for me to improve myself through conversation with them, how much of my energy it will be worth to invest in such reflective self-evaluation.

Thus, in practice, when we criticize a person, tradition, or institution for being too dogmatic, our judgment that the agent or agents in question should exhibit a *greater amount* of self-criticism must always tie in contingent facts about them. Nevertheless, the ultimate reason behind it, we might say, is conceptual: it derives from the finitude of all agency. Therefore, I think it is important to note that the relationalist can accommodate this principled aspect of moral discourse as well.

[7]Recall that on my view, the ideal self of a conceptually possible agent need not itself be construed as another conceptually possible agent, but can instead be understood as the limit of a succession of less erroneous alternative selves.

10.4 The Semantic Objection

I have argued that there are several reasons for us to engage in moral discourse if we understand our practical judgments as having relationalist meaning. This refutes the no-purpose objection, but it also invites a new objection. Because from the fact that people *would* have such reasons if they *did* mean their moral statements in the relationalist sense, it does not follow that people actually *do* mean their statements in that sense, and therefore neither that they *have* moral conversations for those reasons. On the contrary, the nonrelationalist might argue, there is a good reason to think that they don't, because that is simply not what moral statements mean. The words in our moral vocabulary, such as "right," "wrong," "duty," "forbidden," "good," "evil," and so on, have nonrelationalist meanings. So when you use these words correctly, you are not making claims about whether something is right-for-you-but-maybe-not-for-some-other-person. Instead, you are making claims about what is right and wrong, period! And so that is what our moral conversations, in which we use this vocabulary, are about. Let us call this the "semantic objection."

In response, we might ask the nonrelationalist how he *knows* that our moral terms do not have relationalist meanings. Perhaps he would be inclined to answer that "everybody knows this" because it is "common sense." I suspect that several moral philosophers have this intuition that nonrelationalism is the view that takes moral discourse 'at face value,' so to speak, and that the attribution of a property of rightness to an action, for example, is a more straightforward interpretation of what people mean when they call something right than the more complex attribution of a right-for relation between the action and the speaker. Nevertheless, the semantic objection would not be very strong if it relied on the assumption, without further argument, that it is *trivial* that moral terms do not have relationalist meanings. Instead, the objection becomes more interesting if we understand it as another articulation of the idea that nonrelationalism is a *platitude* about moral judgment. Smith, for example, has argued as follows:

> Let's, then, confront the conceptual question head on. Is our concept of a normative reason relative or non-relative? The relativity of a claim should manifest itself in the way we talk. [. . .] The question to ask is therefore whether the way in which we talk about reasons for action and rational justification reflects

> a relative or a non-relative conception of truth conditions.
>
> One reason for thinking that it reflects the non-relative conception comes from the broader context in which the question is being asked. For it is important to remember that we have a whole range of normative concepts: truth, meaning, support, entailment, desirability, and so on. Between them these concepts allow us to ask all sorts of normative questions, questions about what we should and should not believe, say and do. But how many of these other normative concepts are plausibly thought to give rise to claims having relativized truth conditions? As I understand it, none of them do. (1994, pp. 166–167)

As we can see, Smith thinks that the nonrelationalist semantics of practical judgment is something that can be argued for with reference to the way in which we use language. Thus, in the argument in this passage he tries to make such a semantics plausible by showing how it would be similar to the semantics of other normative concepts. I think this particular argument is easily refuted, by the way. For there are many modes of normativity, and each mode must have at least some unique features that distinguish it from the others. One such distinguishing feature of practical normativity is its conceptual relation to motivation and self government, as articulated in my Authority Principle. And we have seen that Smith's Practicality Principle captures roughly the same intuition. Clearly, then, he does not think that our moral concepts cannot have this motivational dimension merely because it would set them apart from other normative concepts. In the same fashion, relationalism might be a feature of moral concepts that sets them apart from our concepts of epistemic normativity or logical necessity, say. In fact, if my arguments for relationalism are sound, then it is in virtue of the motivational aspect that sets them apart and that Smith already acknowledges, that our concepts of practical normativity also have a relationalist aspect.

However, what concerns me now is the more general idea that a nonrelationalist semantics of moral discourse could be argued for with reference to the way in which we talk when we use moral language. Isn't it rather common for people to talk about morality in a way that follows nonrelationalist inference patterns? For example, consider moral disagreements about homosexuality. Isn't it obvious that people who claim that "homosexual acts are immoral" will infer from that claim that *anyone*, who judges that it is okay for consenting adults of the same sex to have sex, is getting

it wrong? And conversely, isn't it equally obvious that their opponents in the gay rights movement, who claim that "discrimination on the basis of sexual orientation is immoral" will infer from their claim that *nobody* can have a good reason to deny same-sex couples the same rights that heterosexual couples enjoy?

Can it not be demonstrated, not just on the basis of the surface grammar of such claims, but moreover from the way in which people make use of these claims in moral and political debates, that they disagree about them in the same manner in which people can disagree about whether, say, homosexuality is partially determined by certain factors during pregnancy? And if so, doesn't that mean that nonrelationalism is a platitude about moral judgment after all, and hence, that the semantic objection against relationalism succeeds?

10.5 Conceptual Revisionism and Folk Meta-Ethics

As I have already explained briefly in section 4.3, my response to this issue is not to deny that people sometimes mean their practical judgments in a nonrelationalist sense, but rather to defend a *semantic pluralism* according to which different people can mean their practical judgments in different ways. Whenever they mean their judgments in a way that commits them to nonrelationalism ("NR-practical judgments"), those judgments must strictly speaking be false, if relationalism is true. But on other occasions, our judgments may be assigned relationalist truth conditions ("R-practical judgments"), in which case they can be true. In other words, I am a realist about R-practical judgments and an error theorist about NR-practical judgments.

I will now defend this approach in further detail. As we will see, semantic pluralism upsets an assumption that has been implicit in much of moral philosophy, which I will call "conceptual conservatism." Rejecting this assumption, I shall argue that we should be "conceptual revisionists" instead. But before we go into that, I want to make a few preliminary remarks about what we might call "folk meta-ethics" (in analogy to *folk psychology*): the ideas about the nature of morality, insofar these can be inferred from their speech acts in moral discourse, held by ordinary people, bless their souls, whose mastery of the concepts they use may vary.

10.5.1 How Nonrelationalist Is Folk Meta-Ethics?

Is the average moral language user a nonrelationalist realist? And if not, would that undermine the semantic objection? One view would be that most users of moral language do not have meta-ethical opinions in the first place, just like most of us do not have opinions about disagreements in specialized fields of medicine, aeronautics, or quantum theory. There are many things that we can use without having opinions about their inner workings or the principles that guide their application: I can use my own eyes without having studied the physiology of human vision, I can use my glasses without knowing the principles of optics, and in similar fashion, we might be able to use moral language without having thought about meta-ethical questions. On this view, proper meta-ethical scrutiny of the way in which moral language operates might reveal that we mean something nonrelationalist when we make practical judgments even if most of us have never given this a thought.

However, it seems to me that many people actually do hold meta-ethical views and that they do make meta-ethical claims in moral conversations. For starters, many people have religious beliefs and many religious belief systems include meta-ethical ideas. Thus, Catholic doctrine includes the nonrelationalist idea that certain acts are morally wrong because they are "intrinsically unordered." More generally, all monotheistic traditions seem to involve beliefs about an essential or conceptual relation between morality and the will of God, and different views about the nature of this relation have fueled heated theological debate throughout the centuries.

Now perhaps a proponent of the view that moral language use need not involve meta-ethical reflection might want to object that neither need religious practice involve theological reflection. Many religious people do not worry about these questions very much. However, I do not think this observation applies across the board. It is simply false that only theology students would worry about theological questions: many people struggle with a crisis of faith, or simply wonder whether God might exist even if they lack specific religious commitments. When people change their religious beliefs, or lose them altogether, they often have to evaluate the manner in which they arrive at moral judgments or the sources that they used to rely on for moral justification.

Furthermore, even if theology often involves nonrelationalist beliefs, there are many cases in which people (including religious people) seem to advocate relationalist ideas as well. In my experience a lot of people tend

to be even far more *relativistic* in their claims than I have been in this thesis. Of course, their formulations are usually less precise than those used by academic philosophers who have specialized in this field. For example, some people will say that moral judgments are "just opinions, not facts." This phrase can be used to claim that although these opinions are opinions about facts, they have not been established *as facts* in a scientific manner. But from the context it is sometimes clear that people actually mean that there are *no* facts in ethics beyond the facts about the opinions themselves. I have also heard several people who did not study philosophy say about morality that it is "subjective"—often with an air as if this insight should be beyond any doubt for someone who has given the matter some thought.

In fact, when friends and relatives want to know what my thesis is about, I have experienced more trouble defending the idea that at least some moral judgments might be true for all human beings including those who currently make opposite judgments, than I have defending the idea that there are no moral facts independent of the attitudes we happen to have. Finally, I see people making all sorts of relativistic sounding remarks in online communities. Whenever someone proclaims a certain moral assumption or recommendation as true, people respond with claims like "but that is your truth, I have my own truth." When some cultural practice is renounced, you can expect that somebody will not be able to resist responding that "it is true in their culture." And so on.

Of course, these observations are anecdotal. But my impression, at least, is that people do have meta-ethical ideas, and that these are very diverse, from extremely relativistic forms of relationalism to extremely fundamentalistic versions of nonrelationalism. It does not follow that people are also very clear in their views; in fact I think they usually aren't.

So folk meta-ethics at least does not seem to be univocally nonrelationalist in the sense that not all the folk subscribe to nonrelationalism. But could it perhaps still be nonrelationalist in the sense that most, if not all, average moral language users follow nonrelationalist inference patterns, even if those among them who have relationalist ideas about morality apparently are not aware of this? After all, like so many dispositions, our inferential dispositions might not be fully transparent to us. If they where, then conceptual analysis would be easy, which it is not.

According to the Affective Response View that I have defended, however, one of the core inferential dispositions that we have with respect to our practical judgments is that we expect our affective responses to

the intended consequences of our actions to confirm the judgments upon which we acted. And as I have argued in section 7.3, the presupposition that our practical judgments are predictive of our affective responses in this manner implies relationalism. So the Affective Response View essentially postulates a *relationalist inferential disposition* as one of the most common dispositions that we have with respect to practical normativity. And I have been arguing for this view on independent grounds. If anything, I suspect it will more often be the case that someone who proclaims nonrelationalist ideas about morality turns out to have relationalist inferential dispositions, than the other way around.

Nevertheless, I do not think that the meta-ethical ideas that people hold cannot influence their inferential dispositions whatsoever. On the contrary, the whole point of my semantic pluralism is to be able to account for the fact that even though I think the Affective Response View is the proper story about how we should, and often do, come to revise our practical views, there will nevertheless be some people who are so explicitly committed to their own epistemic and metaphysical assumptions about morality that their inferential patterns will not be in accordance to what the Affective Response View requires. For example, someone who very strongly believes in a certain interpretation of the Bible according to which moral obligations are to be established on the basis of Scripture alone and have nothing to do with what we may happen to like or dislike, may be able to stick with his moral judgments even in the face of overwhelming negative affective responses that he has to endure as a result of his way of living. In such a case, it no longer seems plausible to identify this person's practical judgments with volitional judgments, and I propose an error theory about his judgments instead.

Summarizing, my conclusion about 'folk meta-ethics' is that we have no good reason to believe that average moral language users are predominantly nonrelationalist, neither in their own ideas about morality nor in their inferential dispositions. By itself, this does not yet refute the semantic objection, however. Recall that on Michael's Smith's view, the meanings of our moral concepts are to be determined on the basis of platitudes about the inferential dispositions of those moral language users who have "fully mastered" those concepts. Therefore, one of the things that Smith might say at this point is that what is common or average in our moral practice is completely irrelevant because average language users need not have full mastery of their concepts. Even if the Affective Response View

captures a common inferential disposition, the meaning of everybody's moral judgments may still be determined by a small minority of people whose inferential dispositions strictly reflect the Principles of Reason View, for example, if they are the ones who get the concepts right.

But is it plausible to say that certain people 'mean' their judgments in ways that go completely against their own understanding and application of the concepts they use to form those judgments? I will turn to this question below.

10.5.2 Conceptual Conservatism

Recall that Mackie made a division in meta-ethics between *conceptual* and *substantial* claims: the former provide the truth conditions for practical judgments while the latter identify types of facts, if any, that satisfy those conditions. The semantic objection is based upon a conceptual claim: regardless of whether there really are any nonrelationalist facts about what we should do, the objection is that the truth conditions of our practical judgments require such facts because that is what it *means* to use the moral concepts that we do. Should it turn out that there are no such facts, then our judgments must be false, but it wouldn't change the meaning of our concepts.

In so far as concept mastery is not determined by contingent sociocultural facts about which inferential dispositions or which interpretations of the concept are most common among ordinary language users, it presumably must be determined by considerations of conceptual necessity: any *widespread* lack in mastery of some concept must involve some *incoherence* in the application or understanding of the concept by the user. Otherwise, we might just as well consider this widespread usage to be the 'mastery' and rival usages to be 'lackings in mastery' instead:

> To say that we make inferences and judgments along these lines is, of course, consistent with the possibility of our coming to think that it is wrong to do so. Our prereflective inferential habits are, after all, corrigible. We would, for example, change our inferential habits if we were shown that the judgments and inferences that we make as masters of the term "red"—the platitudes themselves—were incompatible or inconsistent with each other. However it is to say that it would take something

> like such inconsistency to make us change our inferential and judgmental habits. (Smith, 1994, p. 30)

However, suppose that the *substantial* claim of nonrelationalism would be false. And suppose that it would be false for the reasons which I have discussed in chapters 5 and 6: in order for nonrelationalist values to exist, they must follow from conceptual considerations of coherence, but such considerations seem to leave no room for such values to exist. In that case, it seems to follow from the falsity of the substantial claim, that the corresponding conceptual claim must be false as well, for it would have turned out that the very idea of nonrelationalist value is incoherent. Hence, the conceptual claim might not be so independent from the substantial claim after all.

This insight leaves us with a choice. For there are two things that can happen when we detect some inconsistency in our inferential habits pertaining to some concept. If we judge one of the conflicting inferential dispositions to be not such an important part of the concept, then we may *improve* our mastery of the concept by removing that disposition or by adjusting it so as to remove the incoherence.[8] But if we judge instead that both of the conflicting dispositions reflect essential or core platitudes about the concept, then we may instead feel compelled to simply *abandon* the concept altogether. Let us now apply this choice to the idea that there will be no convergence of all possible agents onto similar attitudes, and that therefore, nonrelationalism cannot be reconciled with the Facts, Authority, and Distinctness Principles. If we take the first option, we end up with the revisionist error theory that I shall be defending in combination with my semantic pluralism below. Instead, if we take the second option, we get the nihilistic error theory that Michael Smith has been considering as the most plausible alternative should his realism fail, which we have seen in section 6.3.5.

So the conceptual claim that Smith is defending, and that the semantic objection requires, is not just that nonrelationalism is part of our concept of practical normativity, but rather that it is an *essential* part of the concept in such a way that it cannot be removed from the concept should it turn out to be part of an inconsistency. It is here to stay, so to speak, and if it really must go, then it will take the entire concept down with it. Regarding the

[8] If we judge *both* of the conflicting platitudes to be unessential, we may also disambiguate the incoherent concept into two coherent concepts: one which loses the one platitude and one which loses the other.

issue of nonrelationalism, on this view, our concept of practical normativity is fixed and immutable. It cannot change. Let us call this view "conceptual conservatism."[9]

Furthermore, if nonrelationalism cannot be separated from the concept of practical normativity, then any average moral language users who did not have nonrelationalist inferential dispositions in the first place presumably cannot become masters of their moral concepts until they *adopt* such dispositions—or at least, that is how I mean to understand "conceptual conservatism" here. It follows that conceptual conservatism rules out my semantic pluralism. According to the conceptual conservative, then, the semantic principle that assigns truth conditions to a judgment in approval of *P* is always the same, regardless of who is approving of *P* or when.

Note that a conceptual conservative can also be a relationalist: if he favors a relationalist analysis of the meaning of practical judgments, then his claim, which I shall call "conservative relationalism," is that the truth conditions of a judgment in approval of *P* will always be relationalist, regardless of who is approving of *P*. Of course, that means the facts that might satisfy those conditions *do* depend 'strongly' (as explicated in section 5.1.2) upon who is making the judgment, because that is what relationalism tells us, but the thing that remains the same regardless of who that agent is, is that those conditions always contain that reference to him or her. By contrast, the "conservative nonrelationalist" claims that the meaning of no practical judgment can ever involve such strong dependence, but only weak dependence (again as explicated in section 5.1.2) upon attitudes of the judging agent, regardless of who she is. What the conservative relationalist and nonrelationalist have in common is that they both believe their favored meta-ethical semantics should be applied to moral judgments across the board.

In my discussion of folk meta-ethics, I have been talking about the conceptual understanding and inferential habits of average moral language users. But the most striking illustration of conceptual conservatism comes from its application to the moral judgments made by leading academic philosophers in the field of meta-ethics itself. Thus, according to the

[9]I realize that this label is not entirely satisfactory, as the willingness to eliminate rather than modify something may seem like an odd way of "conserving" it, but I haven't managed to come up with a better term so far. In any case, "conservatism" does imply a resistance to change and contrasts nicely with the label of "revisionism" that I am using for the alternative to it.

conservative nonrelationalist, whenever Gilbert Harman utters a moral judgement in approval of *P*, he must be judging that there is an objective nonrelationalist fact that *P* should be the case, even if Gilbert Harman does not believe in such facts, does not believe his judgement to have that meaning, and published many famous articles about why he does not believe these things. Furthermore, suppose that as a trained philosopher, Harman is at least consistent in the inferences he makes from his own moral judgments with the relativism that he defends. Then even that would not, on this view, make the meaning of those judgments relationalist. Instead, what it would mean is that Harman has failed to master our moral vocabulary. Conversely, the conservative relationalist thinks that even the moral judgment of a staunch believer in objective nonrelationalist moral facts such as Derek Parfit should be evaluated in terms of a relationalist strong dependence upon Parfit's attitudes. And if Parfit's inferential dispositions are strictly nonrelationalist, then the implication would be that Parfit has not mastered his moral vocabulary.

I find these implications deeply implausible. Surely people like Harman and Parfit must have mastery of our moral vocabulary? If *they* don't even know how to apply moral language, then who does? And surely, if Harman does not believe in nonrelationalist values, then it would be odd to think that when he approves of something, he is making a claim about something having nonrelationalist value, even if he himself would never make any nonrelationalist inference from his approval?

10.5.3 Conceptual Revisionism and Descriptive Conceptual Pluralism

Despite these strange implications, I take it that conceptual conservatism is pretty much the standard view in analytic meta-ethics, and that it is often assumed implicitly or without argument. However, some notable philosophers have taken steps towards a somewhat different view. Richard Brandt, for example, has made the following remark:

> [T]here is no reason to think there is any language-wide single meaning for these terms. In general, our meanings are entwined with our total conceptual system, and as our total beliefs change, our concepts change. Why should this not also be true of practical concepts? Take the moral concepts of religious people. Suppose a religious man says that what he means by "wrong" is "prohibited by God." Now, philosophers have used a dialectical

> device, as early as Plato in the *Euthyphro*, to show that this theological concept cannot be what is meant by "wrong." But the dialectical device at most shows that religious people are confused, not that their moral concepts are identical with those of Bertrand Russell. Nor is there reason to think that the moral concepts of unruly boys from the east side of London or New York are identical with those of Moore or Sidgwick. (1979, pp. 6–7)

From a conceptual conservative point of view, one might say that if the religious man in Brandt's example is confused in his understanding of what is meant by "wrong," then the man has simply not mastered the concept, but the concept itself is still determined by the inferential habits of those who do have such mastery. But what Brandt is saying is that the religious man's understanding of wrongness as that which is prohibited by God *counts* as a concept of wrongness in its own right, even if this concept turns out to be confused. Like Brandt, I want to claim there are no single meanings for the terms in our moral vocabulary, and that whatever analysis of these terms we propose as part of our philosophical theory need not apply as the semantics for the judgments made by someone like the religious man in his example.

Furthermore, we have seen in section 4.3 that Bernard Williams not only allowed different people to have different concepts of a reason, but that he also claimed that the reason statements uttered by these people can therefore have different sorts of truth conditions. In particular, the utterances made by those who mean them in the external sense have truth conditions that cannot be satisfied by facts about reasons: these utterances are all "false, or incoherent, or really something else misleadingly expressed" (1980/1981a, p. 111). Finally, we have also seen that Williams's view resembles that of Mackie in this respect, and the remarks from Mackie which I have discussed suggest that we should not understand his error theory along the nihilistic lines considered by Smith, which presuppose conceptual conservatism.

My own view, which I shall call "conceptual revisionism," is that we should extend Mackie's classification of conceptual vs. substantive claims in meta-ethics by making a further distinction between "descriptive conceptual" and "prescriptive conceptual" claims. *Descriptive conceptual* claims are about what people actually mean when they make practical judgments, depending upon their concepts as they, possibly incoherently,

understand them, their possibly false beliefs about the nature of morality, and/or their possibly conflicting inferential habits with respect to those judgments. By contrast, *prescriptive conceptual* claims are about how to properly analyze practical normativity, which must be fully coherent and cannot depend upon false beliefs or conflicting habits.

So the prescriptive conceptual thesis I want to defend is relationalist (we could call it "prescriptive relationalism"), whereas my descriptive conceptual thesis is the semantic pluralism according to which people make R-practical judgments that follow a relationalist semantics as well as NR-practical judgments that are nonrelationalist (let us call this "descriptive conceptual pluralism"). The combination of these claims is revisionist in the sense that the people who make NR-practical judgments should revise their concepts and inferential habits in order to start making R-practical judgments instead.[10]

Now the essential difference between revisionism and conservatism is that the revisionist allows the confused concepts or conflicting inferential habits of average language users to influence the truth conditions of the judgments they express. However, this idea comes in two flavors, which I shall call "hard" and "soft" revisionism.

Hard revisionism is the view that the truth conditions are fully determined by the descriptive conceptual thesis. Thus, if nonrelationalism is incoherent, and if Derek Parfit is completely nonrelationalist in his beliefs and inferential dispositions, then the hard revisionist must conclude that all Parfit's practical judgments are *false*. Only once Parfit allows himself to be convinced by his opponents that relationalism is true, which will never happen, could he start to make true practical judgments. Now at first glance, this view may seem implausible, especially if we consider that I rejected conceptual conservatism in part because it required me to say that Parfit would not have mastery of his moral concepts. Trading the conclusion that Parfit has no such mastery for the conclusion that Parfit has no true moral beliefs may hardly seem to be an improvement.

[10]Note that this 'should' is essentially the 'should' of epistemic normativity: it is about correcting our *theoretical* beliefs about what it means for something to be a normative reason. Nevertheless, I suppose it might be argued that agents generally have an interest in not being confused about what it means for them to have reasons, in which case my prescriptive claim might imply a conceptually necessary *volitional isomorphism judgment* (as discussed in section 10.3.4 above) that all agents have a normative reason to make R-practical judgments instead of NR-practical judgments.

However, it all depends on how we think about truth and knowledge. Consider the transition from Newtonian mechanics to relativity theory. Some people would say that because relativity theory has changed the concept of mass, all statements made from within the framework of Newtonian mechanics must strictly speaking be false, since they refer to a property that is never instantiated. But relativity theory can explain under which conditions those statements are nevertheless *approximately* true, in the sense that the predictions we infer from them are very reliable. Hence, even if Newtonian mechanics produces false statements, it succeeds in capturing the structure of reality to a certain extent, which may be understood as a kind of knowledge.

A relationalist who takes such a view towards revolutions in physics may have no trouble accepting hard revisionism about practical judgments. On this view, saying that large amounts of beliefs are false is just really not such a big deal. Furthermore, if the R-practical judgments that would have been true about Parfit would still be in approval of the same things that his actual judgments are in approval of, such that his actual judgments do capture his volitional inner reality in the sense that they match his affective responses, then the hard revisionist can say that Parfit's actual judgments are *approximately* true, which is perhaps all that anyone could hope for anyway.

Nevertheless, if we do not like these implications we can also opt for *soft revisionism*, the view that assigning truth conditions is a complex interplay between the descriptive and prescriptive conceptual claims, depending on the context of the utterance. Consider the physics analogy again. Suppose we say that judgments derived from Newtonian theory are true when they manage to interact with nature successfully, so to speak, in the sense that they successfully predict or explain its behavior, even if the theory itself is corrected by successive physical theories. From the point of view of relativity theory, those are the circumstances in which Newtonian statements about mass manage to say something about relativistic mass. And from the point of view of whatever ultimate physical theory the future will bring, as long as it will keep some *revised* concept of mass rather than having completely eliminated the very idea, we might be able to say that the true judgments derived from Newtonian theory manage to say something about whatever mass really turns out to be.

Now the conditions under which we are allowed to do this are limited, I think, by two contextual constraints. The first is obviously that Newtonian

judgments are false when they deviate from their relativistic successors. The acceptable margin of error is contextual: it depends, for example, on the precision of the instruments we are using. Thus, the prediction that the measurement will be "5.3 seconds" is successful if the deviation is less than a millisecond and the instrument does not even indicate hundredths of seconds anyway. The second constraint is that we might want to say that the statement is false, or at least that it is not justified and hence not knowledge, when it is made in a context where it heavily depends on features of Newtonian theory that are eliminated by its successor.

If we apply the first constraint to NR-practical judgments, then we will use the descriptive conceptual claim to render these judgments false when the agent would have made a clearly different R-practical judgment, for example when his belief in nonrelationalist value let him to ignore certain of this affective experiences as irrelevant to morality, say. The most interesting examples of this are the 'perverse cases' in which people make volitional judgments that are different from their own NR-practical judgments, which I will discuss below.

And if we apply the second constraint, then we can say that an NR-practical judgment like "stealing is a sin" might be false, or at least unjustified, when the speaker directly derives this from the Bible, say, without having used insight in his own dispositions, even if his normative volitional attitudes do happen to reject stealing. In such a case, the person is simply not talking about or interacting with the facts that would have made a volitional judgement against stealing true.

But in all other cases, we could use the prescriptive claim to interpret the practical judgments of people with nonrelationalist beliefs as nevertheless successful cognitive interactions with their inner volitional realities. In other words, we apply a *principle of charity* from the perspective of our own prescriptive meta-ethical theory in order to make as much sense of the other person's judgments as we can.

10.5.4 Surface Grammar and Quasi-Nonrelationalism

When someone says "Peanut butter is good," most of us will understand his judgment as including an implicit reference to his taste, preferences, or perhaps human taste generally, but in any case, we will not understand this claim as an attribution of a property that peanut butter has independent of contingent human features (as we have seen, nonrelationalists can

also accommodate this phenomenon, through the framework of weak dependence). Therefore, all parties may agree that the surface grammar of evaluative statements need not fully explicate the logic of the propositions they express. In the same manner, philosophers who have defended relationalist views usually note that a statement like "Charity is good" or "Stealing is wrong" may hide an implicit reference to the speaker as well.

It follows that descriptive conceptual pluralism does not require us to attribute NR-practical judgments whenever the surface grammar of someone's utterance has such a 'nonrelationalist looking' form. Furthermore, suppose that John says "Stealing is wrong" and Mike responds with "Not always, for it is okay to steal if the alternative is starving to death." In that case, adding implicit references to their respective attitudes would seem to break the inference that John's judgment must be false if Mike's is true. But as I have argued in section 10.2.1 we must look for conversational implicatures to explain such disagreements, which allows us to import a shared assumption of volitional similarity into the context of the discussion. So in this case the 'surface' also includes an apparently nonrelationalist *inference pattern* which hides implicit inferential dispositions that are actually relationalist.

This way of thinking resembles Simon Blackburn's "quasi-realist" semantics for moral discourse (1984). Blackburn is a noncognitivist about practical judgments, of course, but what noncognitivism has in common with relationalist cognitivism is that practical judgments involve affective attitudes of the judger even if the surface grammar of their expression may hide such involvement. Blackburn has been trying to show that realist inference patterns which seem part of much of our everyday moral conversations can be explained pragmatically with reference to noncognitive judgments and without postulating "real" moral properties in the external world. In similar fashion, I propose that many seemingly nonrelationalist inference patterns can be accounted for by a "quasi-nonrelationalist" semantics that translates them back into intersubjective relationalist inferences.

But if this is possible, then we may wonder how much difference there remains between situations in which we attribute relationalist beliefs to agents who make use of quasi-nonrelationalist surface grammar to express them on the one hand, and on the other hand, situations in which we attribute nonrelationalist beliefs to agents but determine that their judgments may be called true anyway because of the provision of soft

revisionism which allows us to say that their judgments nevertheless succeeded to capture their inner normative realities. My answer is that I think there still is a difference: it is a difference between the nonrelationalist meaning that Michael Smith has in mind when he says "this act is wrong" and the relationalist meaning that I have in mind when I utter the same sentence.

Nevertheless, what I can admit is that there may be a gray area of cases in between. I have argued that some average moral language users mean their practical judgments specifically in relationalist or nonrelationalist ways, but this need not apply to all of them, or all of the time. I certainly want to allow that people sometimes, or that some people, make their practical judgments in ways that leave the question of whether nonrelationalism is true completely in the open.

This means that my semantic pluralism really allows for five types of situations. 1. When someone uses relationalist formulations explicitly. 2. When someone uses quasi-nonrelationalist surface grammar or inferences, but clearly has a relationalist interpretation in mind. 3. When someone makes a practical judgment without being clear or even thinking about a relationalist or nonrelationalist understanding of what he is asserting. 4. When someone clearly has a nonrelationalist meaning in mind, but the constraints of soft revisionism allow us to assign relationalist truth conditions to his judgments anyway. 5. When these constraints are not met, and the nonrelationalist judgment must be assigned nonrelationalist truth conditions.

As it turns out, the revisionist account I am proposing only requires us to adopt an error theory in situation 5. By contrast, in situations 1–4 we can maintain that the judgments in question are satisfiable.

10.5.5 Perverse Cases

As we have seen in section 8.2, the notion of a volitional judgment, or a judgment of "identification" or "endorsement" as it is often called, is a significant philosophical analysandum in its own right, regardless of whether it also serves as a proper analysans for the concept of a practical judgement. In fact, Gary Watson and Michael Bratman have argued that practical judgments cannot be volitional because we can make volitional judgments that we want something *despite* our own practical, and in particular moral, belief that it is not what we *should* do.

> When it comes right down to it, I might fully 'embrace' a course of action I do not judge best; it may not be thought best, but is fun, or thrilling; one loves doing it, and it's too bad it's not also the best thing to do, but one goes for it without compunction. Perhaps in such a case one must see this thrilling thing as good, must value it; but, again, one needn't see it as expressing or even confirming to a general standpoint one would be prepared to defend. One may think it is after all rather mindless, or vulgar, or demeaning, but when it comes down to it, one is not (as) interested in that.
>
> Call such cases, if you like, perverse cases. The point is that perverse cases are plainly neither cases of compulsion nor weakness of will. There is no estrangement here. One's will is fully behind what one does. (Watson, 1987/2004a, pp. 168–169)

Bratman concurs, noting that "value judgment is one thing, ownership another" (2003, p. 227). It may seem that these "perverse cases" are counterexamples to the volitional analysis of practical normativity. But I think not. Instead, my descriptive conceptual pluralism allows me to accommodate these cases in the following manner. Consider a case in which someone believes that a certain act is wrong because it is forbidden in the Bible, but nevertheless judges that she really wants to do it. We can then say that her moral disapproval of the act is in fact an NR-practical judgment, while her volitional approval of the act expresses her cognitive will which tracks her normative will (accurately or not). It does follow that she must be mistaken in her idea that what she should do is not what she really wants to do, but it seems to me that it is only plausible to think that perverse cases are cases where *something* has gone wrong even if there is no "estrangement," as Watson put it, in the volitional judgment itself.

This solution may seem to have two limitations: it only works when the practical judgment is clearly nonrelationalist, and consequently, it only works when it is the judgment on the practical side, and not the one on the volitional side, so to speak, that is going astray. Regarding the first issue, it is true that I have only developed and defended conceptual pluralism with respect to the difference between relationalism and nonrelationalism, so far. But there is no reason why this approach cannot be extended to other meta-ethical differences if it should turn out that people really seem to be making, meaning, or inferring their judgments in correspondingly different ways.

Consider, for example, the distinction *within* relationalist cognitivism between the cultural relativism according to which the truth conditions of practical judgments are determined by cultural standards on the one hand, and my own relationalism in terms of opaque volitional attitudes on the other. It seems possible that a cultural relativist could make the practical judgment that he should not do something because his culture disapproves of it while also making the volitional judgment that it is what he really wants. It may be true that his culture indeed disapproves of it, but his practical judgment that *that* means he should not do it is certainly false.

As for the second issue, is the error always on the side of the practical judgment in such cases? Watson, at least, seems to be hinting at cases where the practical judgment is actually rather sound, and where it is the volitional choice that involves the 'perversity' by letting frivolous considerations such as the "thrill" supersede "a general standpoint one would be prepared to defend."

In fact, I think we can account for this possibility as well by extending our descriptive conceptual pluralism a little further. To say that the volitional is a significant analysandum in its own right is to say that there is this general idea of "judging what you really want" or "being behind it as a person," while our concepts of what that means may vary. The notion of volitional opacity is one such concept, and my Affective Pattern View articulates a more specific version of it. But as I discussed in section 8.2.2, before Harry Frankfurt made his move towards opacity, he defended a rather decisionist account that made the decision itself the source of volitional authority, in a manner resembling Sartre's early philosophy. It seems to me that someone who has such a view on what it means to want something can be "behind" his choice, from his own point of view, even when his practical judgments, which he is choosing not to act upon, are inferentially sound along the lines of the Affective Response View and in touch with his affective dispositions.

In general, my approach towards any distinctions between practical normativity and volitional authority as different evaluative modes, so to speak, is as follows. If my account is correct, then once we have revised our concepts of what we should and what we really want, these concepts will turn out to have the same meaning, because they are concepts of the same normative will (though of course the distinction between the cognitive and the normative will remains). But as long as certain revisions still need to be

made, we can have different concepts for which we use words like 'should' or 'want' and there is no reason why only the concepts for which we use 'should' could be on the wrong track. Someone who applies her concept of 'should' in accordance to the Affective Response View, while her concept of 'wanting' is that of a decisionistic radical choice, will have to revise the latter in order to realize its equivalence to the former, if my view is correct, and not the other way round.

Conclusion

I have proposed and defended a combination of two views: the *Affective Response View* of practical disconfirmation (section 7.2), which answers the Disconfirmation Question (section 1.2), and the *Affective Pattern View* of the nature of practical normativity (section 9.1), which answers the Facts Question (section 1.1). The former is the view that we disconfirm our practical judgments on the basis of unexpected affective responses to the intended consequences of our actions. The latter is the view that we get our practical judgments right when they capture certain patterns in the structured inner realities of our affective dispositions. The patterns in question are those that manifest themselves in our resultant motivations in so far as we approach ideal conditions of rational and self-governing agency.

These views are attractive, first and foremost, because they allow us to accept all five of the principles from chapter 1. Each principle articulates a widely held intuition about practical judgment. Furthermore, many philosophers who have nevertheless rejected one of these principles have done so because they judged it to be incompatible with some of the other principles. I have tried to capture these alleged incompatibilities in my formulations of the Facts and Disconfirmation Problems (sections 1.5.1 and 1.5.2). I have argued that the Affective Response View solves the latter (section 7.5), while the Affective Pattern View solves the former problem (section 9.4).

To be sure, some philosophers may reject one or more of these principles for independent reasons. The Humean theory of motivation, for example, has been criticized for being psychologically unrealistic generally, rather than merely for making moral knowledge inert. Against this, I have suggested that the purpose of the Distinctness Principle is not to model our psychological architecture but rather to explicate a core aspect of our concept of *truth* (section 3.4.2). Further development of this argument must await another occasion, however. A second example is the intuition that

self-governing agents need not care about practical normativity. Moral externalists have argued that they see no incoherence in the judgments of the "amoralist" who distinguishes right from wrong but does not care whether his actions are right (Brink, 1986). In somewhat similar fashion, Watson has defended the possibility of "perverse cases" in which people act against their better judgments without the "estrangement" associated with ordinary weakness of will. Although I have provided an alternative account of perverse cases that is compatible with the Affective Pattern View (section 10.5.5), I have not attempted to refute the externalist intuition on its own terms. In general, my main line of argument has simply been premised on the assumption that each of the principles is *prima facie* plausible on independent grounds, and that other things being equal, a synthesis of these principles is therefore preferable, if possible, to any theory that sacrifices one or more of them in order to account for the others. My conclusions in support of the views I propose are conditional on this assumption.

The Analysis of Normative Reasons

Regarding the Facts Problem, such reconciliatory solutions often involve some version or modification of the Internal Reasons View. Although I am roughly in agreement with the spirit, so to speak, of Williams's defense of this view, the devil is in the details, and there I have found much to disagree with. Williams rejected the distinction between normative and motivating reasons on the grounds that all reasons have both justificatory and explanatory implications, but I have argued that he misunderstood the distinction. Both types of reasons have both types of implications, while differing in the nature of those implications (section 2.1.2). Williams seemed to suggest that rejecting the distinction was part of his defense of internal reasons, but I have argued that the defense actually presupposes the distinction and that the Internal Reasons View is about normative reasons. Williams also placed a proximity requirement on instrumental deliberation, apparently as another step in his argument, which I have criticized for being inconsistent with the whole idea of a deliberative route that Williams employed in his later formulations of the Internal Reasons View (section 2.2.1).

Most importantly, Williams did not explicate a clear premise concerning motivation, but I have argued, contrary to Thomas, that without the

premise of the Motivational Continuity Thesis, the defense becomes vulnerable to the non-route-like deliberation objection (section 2.5). I have tried to show that this premise can be defended upon both Humean and anti-Humean theories of motivation (sections 3.4 and 3.5), allowing me to remain neutral about the exegetical question of whether Williams would have accepted the Distinctness Principle or not. Finally, in order to block the nonproceduralist objection (section 2.4), the defense requires proceduralism, and it turns out that the motivationally Humean defense already presupposes a solution to the Disconfirmation Problem (section 3.3.2).

Returning to the Facts Problem and my own preferred conceptual framework from chapter 1, and incorporating the aforementioned criticisms of Williams's defense, I arrive at three possible reconciliatory solutions: type-I, -II, and -III dispositionalism. Each is an attempt to synthesize the Facts, Authority, and Distinctness Principles, but differs in its interpretation of the other two principles: whereas type-I and -II are *proceduralist* about the Disconfirmation Principle, type-III dispositionalism is *nonproceduralist*, and whereas type-I dispositionalism is *relationalist* about the Intersubjectivity Principle, type-II and -III are *nonrelationalist* (section 5.3.2).

I have argued that type-I attempts to solve the problem essentially disambiguate the paradoxical implications of the Facts, Authority, and Distinctness Principles into two fully compatible statements about the dependencies between the beliefs and desires of self-governing agents under ideal conditions (section 4.2). However, this strategy raises two new problems: to explain the authority of some desires over others, and to come up with a relationalist account of intersubjectivity, which is not as straightforward as the nonrelationalist interpretation (section 4.4).

Nonrelationalist dispositionalism faces a different problem: it must explain why all *conceptually possible* deliberators would have *similar* (as opposed to *isomorphic*) desires concerning matters of ethics under ideal conditions (section 5.2). I have discussed Michael Smith's attempt to do so, distinguishing between the proceduralist type-II and the nonproceduralist type-III dispositionalism as possible interpretations of his view, concerning which some of his comments seem ambiguous (section 5.3.2). So far, his critics have typically targeted the proceduralist interpretation with very strong objections (sections 5.4.1 and 5.4.2). I have argued that the most promising response would involve a combination of both convergence and constitution strategies, but while this may allow us to construct an interesting convergence of constitutive isomorphic desire sets onto coherent

isomorphic desire sets, my ultimate conclusion has been that there still seems no logical reason to suppose that this will give rise to similar desires about substantial moral issues (section 5.4.3).

The type-III solution adds a third element to the mix of constitution and convergence: that of *ineliminable epistemic luck*. We have seen that this must be a special type of veritic luck for two reasons. First because it would be *a priori*, which requires an agent-centered rather than a world-centered account of such luck. Secondly, in order to explain such luck as ineliminable it cannot be due to the finite character of our limited cognitive capacities (section 5.3.3). The idea, then, is that the prior desires from which convergence would develop must not only be constitutive of self-government, but also 'lucky' in this particular sense, if we are to arrive at the moral truth. I have argued that even if we assume that certain eternal disputes in theoretical philosophy might be susceptible to this type of luck (section 6.1)—an assumption I myself do not share—then there would still be important dis-analogies between such theoretical matters of belief and practical matters of desire that make it unwarranted to postulate ineliminable epistemic luck in normative ethics (section 6.3).

The Account of Practical Disconfirmation

Having examined the problems for each of the three types of dispositionalism from the point of view of the Facts Question, my subsequent argument has been to shift our perspective to the Disconfirmation Question as the dialectically prior issue in order to determine which answer to the Facts Question is most plausible. We have seen that the account of disconfirmation to which type-II and -III dispositionalists are committed, the Principles of Reason View, requires rather substantial interpretations of such principles in order to explain the sort of actual revisions that we want to explain. I have argued that the simplest way to explain why agents would revise their attitudes in accordance to substantial principles is simply to count those principles, contingently, amongst their intrinsically desired ends. It seems an unnecessary explanatory burden to suppose that such revisions and the corresponding motivational changes are brought about out of sheer logical or conceptual insight (section 7.1.1). In fact, it seemed to me personally—but this may be a rather idiosyncratic intuition—that it really doesn't do justice to the existential challenge of coming to grips with the racist, sexist, homophobic, and xenophobic views people have acted

and continue to act upon, simply to attribute these to their insufficient conceptual or logical skills or capacities.

Furthermore, I have argued that even on the assumption that such substantial principles could account for belief revisions on *a priori* grounds, it would still be a mystery why *intrinsic* desires would then be updated accordingly under conditions of sustained self-government. I explained that the adoption of the Authority Principle does not take care of this, but that on the contrary, this is something that needs to be resolved in order to account for the Authority Principle (given our commitment to the Disconfirmation and Distinctness Principles). Without such an explanation, we would expect intrinsic desires simply to remain undisturbed, breaking down self-government with every practical belief revision. I have argued that the basic idea of self-government explains why certain non-instrumental desire revisions of planning and scheduling come along 'for free' with the corresponding belief revisions, but that this does not go for any principles under the substantial interpretations needed to explain actual moral conversions. So it is hard to see how the Principles of Reason View could solve the Disconfirmation Problem (section 7.1.2).

We have seen that these problems go beyond what is specific to nonrelationalist dispositionalism, since type-1 dispositionalism may also, and is often assumed to, be combined with the Principles of Reason View. The resulting account is a form of relativism according to which our normative reasons are simply the *a priori* rationally transformed versions of the desires upon which we deliberate, a view which shares the aforementioned problems of explaining actual desire changes on substantial moral disconfirmations.

Finally, we have seen that all proponents of the Principles of Reason View need an additional account of disconfirmation in order to handle non-moral matters of personal life decisions that are susceptible to error to a degree that is underdetermined by principled reasoning on any account (section 7.1.3). The Affective Response View is very well suited to handle such cases, but once we adopt this view, it seems to provide better explanations of disconfirmation on moral issues as well (section 7.2). However, the implication of explaining moral disconfirmations in this manner is relationalism: if unexpected affective responses disconfirm practical beliefs, then practical beliefs apparently entail predictions about such responses, which makes them empirical and contingently attitude-related. More specifically, given our Principles, the implication is a form of

type-1 dispositionalism according to which the attitudes that constitute our normative reasons may be *opaque* to us (section 7.3). I have developed an account of deliberation as "volitional interpretation" to make sense of this idea from an epistemic point of view, arguing that no concrete experiential or efficient motivation provides privileged access to our opaque normative attitudes over any others, making every disconfirmation susceptible to further revision in principle, and every practical belief a kind of hypothesis. This account offers a surprisingly simple solution to the Disconfirmation Problem: unexpected affective responses explain motivational changes because they *are* motivational changes (section 7.5).

An Opaque Relationalism

Although this account of disconfirmation leads us back to defending relationalism with respect to the Facts Question, the type of view we are now looking at offers some new ingredients that the Internal Reasons View, in Williams's formulation, did not include or specify. Most importantly, non-instrumental revisions are mostly not rational transformations of known desires, but *empirical discoveries* of unknown affective attitudes that reflect previously unexpected aspects of opaque volitional attitudes. As we have seen, Harry Frankfurt's recent account of normative reasons is much closer to this view (section 8.1).

Nevertheless, I have argued that Frankfurt also, ultimately, fails to account for the relevant real-life cases of practical disconfirmation. His doctrine of volitional necessity cannot accommodate disconfirmations after the fact of the act (section 8.3.1); furthermore, his appeal to claims about unthinkability presupposes what it needs to establish and applies only to extreme cases (section 8.3.2). At a more fundamental level, both ideas explain disconfirmation of the opaque in terms of the occasionally transparent, by postulating privileged experiential attitudes at the top of the desiderative hierarchy or the core of the structure of caring. I have argued that this problem traces back to Watson's original criticism of Frankfurt's early views of free will and I have diagnosed it as being due to an essentially Cartesian epistemology of practical reason (section 8.3.3).

Because the Affective Pattern View that I defend employs an epistemology of *pattern recognition*, it does not fall prey to Frankfurt's problem of coming up with some type of attitude we can deliberate upon that outranks mere desires. There are, instead, only affective attitudes with

which we deliberate that are equally unprivileged. It is the pattern in the whole that constitutes the volitional authority in which its participating parts are enshrined. I have drawn upon the work of Daniel Dennett to make sense of pattern recognition in the context of attitude attribution across time on the basis of imperfect information using predictive success as the ultimate criterion for justification (section 9.1.1). But we have also seen that the affective constituents of the pattern in my account exist at a different level of organization than the behavioral constituents in Dennett's original proposal. Hence, both theories analyze different types of attitudes in terms of different types of constituents. This makes the two theories logically independent (section 9.1.2).

I have argued that the pattern that constitutes the normative will derives its normative significance, in contrast to patterns that merely constitute a motivational character of an agent, from the wider counterfactual context in which it is to be determined. This context is constrained by a special concept of personal identity that involves "alternatives of oneself," loosely borrowing a notion from Bransen (section 9.1.3). I have argued that this way of defining the normative will makes sense of my proceduralist outlook. It allows us to explain why even most remorseless Nazis were mistaken in their practical beliefs (section 9.1.4).

An interesting result of this proposal is that it makes normative reasons present to varying degrees in different cases, which fits nicely with our intuition that not all matters of practical deliberation seem equally 'hard' matters of fact (section 9.2.1). Furthermore, as Dennett already observed, such gradually existing patterns allow for pluralism and conflict within the same set of data, which in my proposal allows us to accommodate genuine moral dilemmas (section 9.2.2). Finally, I argue that we should expect patterns to be loosened or strengthened as a result of changes that may be caused by our own practical decisions, which allows us to account for Bratman's "snowball effect" and Williams's stress on the imaginative aspects of deliberation. And so we get an account that incorporates both the cognitive dimension of *finding oneself* and what we might call the existentialist dimension of *making oneself* (section 9.2.3).

In contrast to Frankfurt's suggestion that the volitional might be a single, *sui generis* branch of attitudes, reducible neither to affect nor cognition, the Affective Pattern View comes up with two types of attitudes that are metaphysically nothing over and above—in the sense that they supervene on—our affective and cognitive dispositions (section 8.4.3). The *normative*

will consists of attitudes that are constituted by patterns in affective dispositions, while the *cognitive will* consists of structures of practical beliefs and accompanying desires at the motivational surface, so to speak (section 8.4.1). To complete this repertoire of volitional classifications, I have also defined a third concept of the *executive will*, which embodies the idea that some of our resultant motivations may be under our volitional *control* even when we do not endorse them. That is, we do not endorse them in the sense that they diverge from the practical beliefs and non-resultant motivations that constitute our cognitive will.

On the basis of these three volitional concepts I have arrived at a taxonomy of free agency and responsibility ascription that extends earlier distinctions from Watson and Pereboom (section 8.4.2). Action in accordance to one's cognitive will is free in the self-disclosing sense, which warrants the ascription of *strong attributability* if it reflects one's normative will and *weak attributability* otherwise. By contrast, if one's action is controlled by one's executive will then it is free in a sense that warrants the ascription of *weak accountability*. The familiar debate between compatibilists and incompatibilists concerns the question whether this type of control also warrants ascription of *strong accountability*, or whether that would presuppose a further, indeterministic or perhaps even incoherent, notion of control. Finally, I have distinguished all of these notions from the concept of *conscious will* that psychologists and neuroscientists nowadays often associate with free agency, but which is not immediately related to discussions about responsibility at all.

The distinction between cognitive and normative volitional attitudes has also led me to distinguish between two kinds of *wholeheartedness*: we can have *inner wholeheartedness* with respect to our normative will and *epistemic resolvedness* with respect to our volitional beliefs. I have argued that *neither* are ideals to be pursued. If our normative will is not wholehearted, then it may conflict with certain affective attitudes that, although we have normative reason not to *pursue* them, we nevertheless do have reason to *identify* with as a matter of authenticity. We thereby acknowledge their right, so to speak, to remain part of our mental economy in a non-resultant sense (section 9.3.1).

As for being resolved in our volitional beliefs, given the opacity and various forms of bias inherent in our affective psychology, the risk is that we prevent ourselves from disconfirming disastrous falsehoods if we ease into a state of full resolve too quickly (section 9.3.3). However, this risk

cannot be eliminated with a simple rule or principle for two reasons. First, because this type of wholeheartedness is a form of *passive resolution*, over which we have no deliberative control. Secondly, because it may even be constitutive of practical reason that we have a resolved horizon against which we deliberate. Instead, we can only try to develop the practical wisdom of striking the virtuous balance between criticism and confidence in our deliberative policies *before* passive resolve settles in, so that we have good reason to depend on it when it does.

In contrast to passive resolution, I have distinguished two modes of judgment that we can arrive at by deciding to do so: *active resolution* involves the self-attribution of knowledge, while the adoption of *working hypotheses* allows diachronic stability when our subjective probabilities are in flux. The latter mode is the more critical or doubtful of the two, but it restricts the adverse effects of uncertainty with a form of practical commitment. The former mode is more confident, but this confidence should be checked by policies for possible future revision (section 9.3.2). It is in the proper mixture and application of these two types of decision making that practical wisdom must be found.

And so we have arrived at a detailed reconciliation of the Facts, Authority, and Disconfirmation Principles. For it is now understandable why, the more an agent knows the facts about his normative reasons, the more his capacity for self-government would involve desires to act upon those reasons. Because the facts are about a pattern of desires that manifests itself more fully as his knowledge increases and which seems the proper object of his volitional interpretation, representing what he is really all about as a person. Hence, it seems to make sense of the idea of self-government to say that if his volitional beliefs capture this pattern, he will be self-governing when his resultant desires participate in that pattern, and that he will be lacking in self-government on the occasion when his resultant desire constitutes noise relative to that pattern.

Let us now focus on the former type of occasion where he is both knowledgeable and self-governing, and let us review the two disambiguated implications of the type-I dispositional solution about how, given the Distinctness Principle, his contingent resultant desire might have been different. The first is that his resultant desire might have been different and yet he still might have been knowledgeable and self-governing, but then his beliefs would have been different. That must be so because his knowledge and self-government would imply that the different desire

reflected a different pattern and hence, different normative reasons. The second is that his desire might have been different while his beliefs were the same, but then he would no longer be self-governing. Nor is it true that, if the different desire reflected a difference in the pattern as well, he would still have been knowledgeable. Neither of these implications are contradictions, nor do they contradict each other. The paradox has thus been resolved. The Humean theory of motivation, as captured in the Distinctness Principle, is indeed compatible with the conjunction of the cognitivism specified by the Facts Principle and the internalism specified by the Authority Principle.

Moral Discourse

In the final chapter I have explained how my relationalist theory can account for the intersubjectivity of practical judgment, as manifested by our participation in moral discourse. I have tried to capture the sentiment that relationalism is at a disadvantage in this respect by formulating two objections: the no-purpose objection and the semantic objection.

The no-purpose objection states that if relationalism were true, people would be talking past each other in moral conversations, and therefore, such conversations would be pointless (section 10.1). In response, I have been discussing three types of interpersonal scenarios: those involving volitional similarity, those involving volitional isomorphism without similarity, and those involving neither similarity nor isomorphism. We have seen that similarity is needed in order to account for the Intersubjectivity Principle, but I have argued that it is sufficient to explain that *some* moral conversations involve similarity as long as these include discourse regarding core moral values, and as long as we supply additional reasons for having moral conversations in the absence of similarity (section 10.2.6).

I have distinguished five reasons that explain the purpose of moral discourse: shared psychology, contrast, knowing someone better than he knows himself, increase of alternatives, and the discovery of isomorphic principles of reason. Shared psychology obviously leads to isomorphism, but because of their contingent empirical nature, our isomorphic attitudes need not be restricted in their content to the sort of indexicality or self-reference that would undercut similarity (section 10.2.5). Although evolutionary forces have clearly selected for the trait to care first and foremost for *our own*, we may speculate that the need to recognize suffering

and distress in our peers and children was most efficiently served by a neurological architecture that has eventually provided us with a disposition against *suffering itself*, which now motivates us to care even for those with whom we do not have family ties or reciprocal relationships. Evolution, after all, did not have to worry about whether the latter could be justified *a priori* in terms of the former as a matter of principle.

My argument does not rely on such speculation, however. Regardless of how we came to be the way we are, we have reached widespread agreement in our disapproval of the Holocaust, for example, and we have situational explanations for Nazi beliefs and behavior that make it more plausible to assume they were getting their volitional judgments wrong than to assume they were fundamentally different from us psychologically (section 9.1.4). In similar fashion, I have argued that we may reasonably assume that the dispositions we share as a species are going to have implications for other moral issues as well, including some that may yet be subject to widespread disagreement (section 10.2.2). In many such cases we can give a straightforward account of the disagreement in terms of contradicting *volitional similarity judgments* about *human* values, *human* rights, and so on.

I have also argued that the occasional human being who may turn out to be the exception to such species-wide regularities does not pose an obstacle to this line of argument. Such cases are not a relevant counter-example any more than anomalies in human biology, psychology or medicine upset the idea that the regularities with respect to which they are anomalies are nevertheless features of our species. If we think that we have objective knowledge that the human hand has ten digits, then the Affective Pattern View allows us to have, with the same level of objectivity, knowledge that we should not torture children for fun even if the occasional psychopath would lack the relevant dispositions (section 10.2.3).

Nevertheless, shared psychology does not need to involve species-wide similarity in order to explain interpersonal contradictions in moral discourse. First of all, we often disagree in the context of a shared cultural background, which merely requires *intracultural similarity*. Moreover, even when the cultural scope is unclear or left implicit, then the *conversational implicature* of their utterances may involve the assumption that the interlocutors are volitionally similar (section 10.2.1). Furthermore, since their disagreement is not about what is commonly held within, or definitive of, their cultural identity, but rather about what sort of opaque attitudes each of them may have developed in a shared cultural environment, this

account does not fall prey to the lack of objectivity commonly associated with cultural relativism (section 10.2.4).

Second, we have seen that even in the absence of similarity, a shared assumption about volitional isomorphism still gives us much to disagree about (section 10.2.5), such as when we talk about the extent of our reasons to care more for our own children than for those of others. In this type of scenario, shared psychology may be one of our reasons to have such a conversation. However, another possible reason would be the discovery of isomorphic principles that could be *a priori* (section 10.3.4). After all, I have not denied that there might be such principles of reason, but only that they would secure similarity.

Finally, the Affective Pattern View does of course allow for the third type of interpersonal scenario, in which neither similarity nor isomorphism obtains. I have argued that this sometimes means we have reached the limits of reasoning and argument, for example where individual differences are involved regarding the grey areas in which the patterns supporting rivaling values or principles fail to outweigh each other (section 10.2.6). Nevertheless, in other cases the absence of shared psychology need not make a discussion useless at all because the *contrast* between their different dispositions may help participants to clarify their own points of view (section 10.3.1).

Furthermore, there are two types of reason that may obtain in all interpersonal scenarios, regardless of similarity or isomorphism: one participant may have insights in the psychology of another that the latter lacks about himself (section 10.3.2); and participating in moral discourse increases our awareness of alternative possible viewpoints, which may counteract some of our unhelpful biases and the theory-ladenness of volitional interpretation (section 10.3.3).

In response to the semantic objection—that moral vocabulary simply has nonrelationalist meaning—, I have defended a revisionism about practical concepts, which distinguishes between descriptive and prescriptive conceptual claims in meta-ethics. The descriptive claim I have argued for is that of pluralism: not all people mean the same thing when they make practical judgments (section 10.5.1). Moreover, the conceptual conservatism according to which their truth conditions must nevertheless all follow the same semantics leads to counterintuitive results (10.5.2). Instead, revisionism allows me to argue that some people who have nonrelationalist ideas about their own judgments simply make false judgments because of

Explanations of Moral Discourse			
	Similarity	*Isomorphism*	*Neither*
A Priori	Error theory (10.5.3)	Principles of reason (10.3.4)	
Empirical	Shared psychology (10.2) Knowing someone better (10.3.2) Increase of alternatives (10.3.3)		Contrast (10.3.1)

that. According to *hard revisionism*, all such judgments must be evaluated in terms of nonrelationalist truth conditions. *Soft revisionism* is a compromise between conceptual conservatism and hard revisionism: under certain conditions we may use relationalist truth conditions to evaluate the practical judgments of people with nonrelationalist ideas. It follows that we only have to be error theorists with respect to cases in which their nonrelationalist ideas have prevented their practical deliberations from interacting with their affective dispositions (section 10.5.3). Though I favor soft revisionism, the choice between these two options depends on general assumptions about truth and knowledge that my line of argument in this thesis can remain neutral about.

Finally, I have incorporated two insights into my account from different corners in the meta-ethical landscape. The first is Simon Blackburn's idea that the surface structure of moral discourse, which includes both grammar and inferential patterns, need not be mirrored in the underlying semantics. Although Blackburn is an expressivist, I have argued that relationalist cognitivism can use a similar approach to assign relationalist semantics to 'quasi-nonrelationalist' discourse surface, which further minimizes the amount of cases about which I must really adopt an error theory (section 10.5.4). The second is Gary Watson's observation that people may distinguish between what they should and what their will is fully behind in 'perverse cases,' which I have argued can be accounted for within a volitional theory in terms of a further expanded revisionism about practical concepts (section 10.5.5).

In summary, I have arrived at six explanations of moral discourse: five that provide reasons on relationalist grounds depending on the type of interpersonal scenario, and a sixth that involves an error theory about nonrelationalist judgment (see table). My general strategy to account for moral discourse has therefore been one of 'divide and conquer': there are

many different intersubjective dimensions in play, and if anything I believe the phenomenon is actually more complicated than I have sketched so far, rather than being simpler and more unified.

In fact, a similar observation applies not only to my account of moral discourse, but to the theory I have developed as a whole. I have often sought to combine different insights about different scenarios instead of making sweeping statements that would apply across the board. I have tried to show that there are sometimes facts for our practical judgments to get right, but I have also tried to accommodate cases in which it remains indeterminate what we should do. I have explained the extent to which it is not up to us what reasons we have, but I have also accommodated cases in which we bring about our own reasons by choosing whatever we decide to choose. I have acknowledged the value of principled deliberation, and the moral significance of universalized rules, while maintaining that deontology is only one of many interpretative strategies. Virtue-ethical, consequentialist, and narrative approaches may also capture our inner realities. Which strategy is best depends on the type of problem we are dealing with.

Practical normativity, in other words, is an unclear business, where absolute precision is unattainable. Or we might say that it is diverse and colorful and multi-faceted, which makes it such an exciting subject. In any case, it is a continuous challenge to understand what our passions tell us about ourselves. Reason, therefore, is hardly in a position to obey orders. It is the *interpreter* of the passions, and while the passions are free to be cryptic, reason has the responsibility to make sense of them.

References

Blackburn, S. (1984). *Spreading the word: Groundings in the philosophy of language*. Oxford: Clarendon Press.

Brandt, R. (1979). *A theory of the good and the right*. Oxford: Oxford University Press.

Bransen, J. (2002). Making and finding oneself. In A. Musschenga, W. Van Haaften, B. Spiecker, & M. Slors (Eds.), *Personal and moral identity.* Dordrecht: Kluwer Academic Publishers.

Bratman, M. E. (1987). *Intentions, plans and practical reason*. Cambridge, Mass.: Harvard University Press.

Bratman, M. E. (2003). A desire of one's own. *The Journal of Philosophy*, *100*, 221–242.

Bratman, M. E. (2006). A thoughtful and reasonable stability: Comments on Harry Frankfurt's 2004 tanner lectures. In H. G. Frankfurt & D. Satz (Eds.), *Taking ourselves seriously & Getting it right* (pp. 77–90). Stanford: Stanford University Press.

Bratman, M. E. (2009). Intention, practical rationality, and self-governance. *Ethics*, *119*, 411–443.

Brink, D. O. (1986). Externalist moral realism. *Southern Journal of Philosophy*, *24*(Suppl.), 23–42.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. New York: Oxford University Press.

Chalmers, D. J. (2003). Consciousness and its place in nature. In S. Stich & F. Warfield (Eds.), *Blackwell guide to the philosophy of mind.* Blackwell.

Chalmers, D. J. (2009). Ontological anti-realism. In D. J. Chalmers, D. Manley, & R. Wasserman (Eds.), *Metametaphysics: New essays on the foundations of ontology* (pp. 77–129). New York: Oxford University Press.

Chalmers, D. J. (2011). Verbal disputes. *Philosophical Review*, *120*(4).

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA: The MIT Press.

Churchland, P. M. (1995). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Cambridge, MA: The MIT Press.

Davidson, D. (2001a). Actions, reasons and causes. In *Essays on actions and events* (2nd ed., pp. 3–19). Oxford: Oxford University Press. (Reprinted from *The Journal of Philosophy*, 1963, *LX-23*, 685–700)

Davidson, D. (2001b). On the very idea of a conceptual scheme. In *Inquiries into truth and interpretation* (2nd ed., pp. 183–198). Oxford: Oxford University Press. (Reprinted from *Proceedings and Addresses of the American Philosophical Association*, 1974, *47*, 5–20)

Davidson, D. (2001c). True to the facts. In *Inquiries into truth and interpretation* (2nd ed., pp. 37–64). Oxford: Oxford University Press. (Reprinted from *Journal of Philosophy*, 1969, *66*, 748–764)

Davidson, D. (2001d). Truth and meaning. In *Inquiries into truth and interpretation* (2nd ed., pp. 17–36). Oxford: Oxford University Press. (Reprinted from *Synthèse*, 1967, *17*, 304–323)

De Kwaadsteniet, E., Van Dijk, E., Wit, A., De Cremer, D., & De Rooij, M. (2007). Justifying decisions in social dilemmas: Justification pressures and tacit coordination under environmental uncertainty. *Personality and Social Psychology Bulletin*, *33*, 1648-1660.

De Muijnck, W. (2003). *Dependencies, connections, and other relations*. Dordrecht: Kluwer Academic Publishers.

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Dennett, D. C. (1991a). *Consciousness explained*. Penguin.

Dennett, D. C. (1991b). Real patterns. *Journal of Philosophy*, *88*, 27-51.

Dennett, D. C. (2001). Are we explaining consciousness yet? *Cognition*, *79*(1), 221–37.

Dennett, D. C. (2003). Who's on first? Heterophenomenology explained. *Journal of Consciousness Studies*, *10*(9), 19–30.

Dewey, J. (1957). *Outlines of a critical theory of ethics*. New York: Hillary House. (Original work published 1891)

Doris, J. (2002). *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Enoch, D. (2006). Agency, schmagency: Why normativity won't come from what is constitutive of action. *Philosophical Review*, *115*(2), 169–198.

Enoch, D. (2007). Rationality, coherence, convergence: A critical comment on Michael Smith's *Ethics and the A Priori*. *Philosophical Books*, *48*(2), 99–108.

Firth, R. (1952). Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research*, *12*(3), 317–345.

Frankfurt, H. G. (1988a). Alternate possibilities and moral responsibility. In *The importance of what we care about: Philosophical essays* (pp. 1–10). New York: Cambridge University Press. (Reprinted from *Journal of Philosophy*, 1969, *66*[23])

Frankfurt, H. G. (1988b). Freedom of the will and the concept of a person. In *The importance of what we care about: Philosophical essays* (pp. 11–25). New York: Cambridge University Press. (Reprinted from *Journal of Philosophy*, 1971, *68*, 5–20)

Frankfurt, H. G. (1988c). Identification and externality. In *The importance of what we care about: Philosophical essays* (pp. 58–68). New York: Cambridge University Press. (Reprinted from *The identities of persons*, by A. O. Rorty, Ed., 1976, Berkeley: University of California Press)

Frankfurt, H. G. (1988d). Identification and wholeheartedness. In *The importance of what we care about: Philosophical essays* (pp. 159–176). New York: Cambridge University Press. (Reprinted from *Responsibility, character and the emotions: New essays in moral psychology*, by F. D. Schoeman, Ed., 1987, New York: Cambridge University Press)

Frankfurt, H. G. (1999a). The faintest passion. In *Necessity, volition, and love* (pp. 99–107). New York: Cambridge University Press. (Reprinted from *Proceedings and Addresses of the American Philosophical Association*, 1992, *66*)

Frankfurt, H. G. (1999b). *Necessity, volition, and love*. New York: Cambridge University Press.

Frankfurt, H. G. (1999c). On caring. In *Necessity, volition, and love* (pp. 155–180). New York: Cambridge University Press.

Frankfurt, H. G. (2004). *The reasons of love*. Princeton: Princeton University Press.

Frankfurt, H. G. (2006). *Taking ourselves seriously & Getting it right* (D. Satz, Ed.). Stanford: Stanford University Press.

Frankfurt, H. G. (2008). *Demons, dreamers, and madmen: The defense of reason in Descartes's* Meditations. Princeton: Princeton University Press. (Original work published 1970)

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgement. *Psychological Review*, *108*(4), 814–834.

Harcourt, E., & Thomas, A. (forthcoming). Thick concepts, analysis, and reductionism. In S. Kirchin (Ed.), *Thick concepts.* Oxford: Oxford University Press.

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society*, *99*, 315–331.

Harman, G. (2000). *Eplaining value and other essays in moral philosophy*. New York: Oxford University Press.

Hooker, B. (1987). Williams' argument against external reasons. *Analysis*, *47*(1), 42–44.

Hume, D. (1964). A treatise of human nature. In T. H. Green & T. H. Grose (Eds.), *David Hume: The philosophical works.* Aalen: Scientia Verlag. (Original work published 1886)

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology*, *21*, 384-388.

Jackson, F. C. (1994). Finding the mind in the natural world. In R. Casati, B. Smith, & G. White (Eds.), *Philosophy and the cognitive sciences: Proceedings of the 16th International Wittgenstein Symposium* (pp. 101–112). Vienna: Verlag Holder-Pichler-Tempsky.

Jackson, F. C., & Pettit, P. (1995). Moral functionalism and moral motivation. *The Philosophical Quarterly*, *45*(178), 20–40.

James, W. (1979). The will to believe. In *The will to believe and other essays in popular philosophy* (pp. 13–33). Cambridge, Mass.: Harvard University Press. (Original work published 1896)

Johnston, M. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, *63*(Suppl.), 139–174.

Korsgaard, C. M. (1996). *The sources of normativity* (O. O'Neill, Ed.). Cambridge: Cambridge University Press.

Korsgaard, C. M. (2006). Morality and the logic of caring: A comment on Harry Frankfurt. In H. G. Frankfurt & D. Satz (Eds.), *Taking ourselves seriously & Getting it right.* Stanford: Stanford University Press.

Lewis, D. (1988). Desire as belief. *Mind*, *97*(387), 323–332.

Lewis, D. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, *63*(Suppl.), 113–137.

Lewis, D. (1990). What experience teaches. In W. Lycan (Ed.), *Mind and cognition* (pp. 499–519). Oxford: Blackwell.

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies*, *6*(8–9), 47–57.

Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. New York: Penguin.

McDowell, J. (2001). *Mind, value and reality*. Cambridge, MA: Harvard University Press.

Milgram, S. (1974). *Obedience to authority*. New York: Harper and Row.

Miščević, N. (2007). Armchair luck: Apriority, intellection and epistemic luck. *Acta Analytica*, *22*(1), 48–73.

Mullen, E., & Skitka, L. J. (2006). Exploring the psychological underpinnings of the moral mandate effect: Motivated reasoning, identification, or affect? *Journal of Personality and Social Psychology*, *90*, 629-643.

Nagel, T. (1970). *The possibility of altruism*. Oxford: Oxford University Press.

Nagel, T. (2000). The psychophysical nexus. In *New essays on the a priori.* Oxford University Press.

Neale, S. (2001). *Facing facts*. New York: Oxford University Press.

Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.

Peirce, C. S. (1992a). The architecture of theories. In *The essential Peirce, vol. 1* (pp. 285–297). Bloomington: Indiana University Press. (Original work published 1891)

Peirce, C. S. (1992b). The fixation of belief. In *The essential Peirce, vol. 1* (pp. 109–123). Bloomington: Indiana University Press. (Original work published 1877)

Pereboom, D. (2001). *Living without free will*. New York: Cambridge University Press.

Pereboom, D. (2002). Meaning in life without free will. *Philosophic Exchange*, *33*, 19–34.

Pereboom, D. (2007). Hard incompatibilism. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four views on free will* (pp. 85–125). Oxford: Blackwell.

Pritchard, D. (2005). *Epistemic luck*. New York: Oxford University Press.

Rabelais, F. (1653). The inestimable life of the great Gargantua, father of Pantangruel. In T. Urquhart (Trans.), *The works of Rabelais.* (Original work published 1542)

Russell, B. (1927). *The analysis of matter*. London: Kegan Paul.

Schaubroeck, K. (2008). *Normativity and motivation*. Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Leuven.

Schroeder, M. (2007). *Slaves of the passions*. New York: Oxford University Press.

Setiya, K. (2010). Sympathy for the devil. In S. Tenenbaum (Ed.), *Desire, practical reason, and the good* (pp. 82–110). Oxford University Press.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology*, *88*, 895-917.

Skorupski, J. (2007). Internal reasons and the scope of blame. In A. Thomas (Ed.), *Bernard Williams* (pp. 73–103). New York: Cambridge University Press.

Smith, M. (1987). The Humean theory of motivation. *Mind*, *96*(381), 36–61.

Smith, M. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, *63*(Suppl.), 89–111.

Smith, M. (1994). *The moral problem*. Oxford: Blackwell.

Smith, M. (1996). Normative reasons and full rationality: Reply to Swanton. *Analysis*, *56*(3), 160–168.

Smith, M. (2004a). *Ethics and the a priori: Selected essays on moral psychology and meta-ethics*. New York: Cambridge University Press.

Smith, M. (2004b). Evaluation, uncertainty, and motivation. In *Ethics and the a priori: Selected essays on moral psychology and meta-ethics* (pp. 343–358). New York: Cambridge University Press. (Reprinted from *Ethical Theory and Moral Practice*, 2002, *5*, 305–320)

Smith, M. (2004c). Exploring the implications of the dispositional theory of value. In *Ethics and the a priori: Selected essays on moral psychology and meta-ethics* (pp. 297–317). New York: Cambridge University Press. (Reprinted from *Philosophical Issues: Realism and Relativism*, 2002, *12*, 329–347)

Smith, M. (2004d). Internal reasons. In *Ethics and the a priori: Selected essays on moral psychology and meta-ethics* (pp. 17–42). New York: Cambridge University Press. (Reprinted from *Philosophy and Phenomenological Research*, 1995, *55*, 109–131)

Smith, M. (2004e). Moral realism. In *Ethics and the a priori: Selected essays on moral psychology and meta-ethics* (pp. 181–207). New York: Cambridge University Press. (Reprinted from *Blackwell guide to ethical theory*, pp. 15–37, by H. LaFollette, Ed., 2000, Oxford: Blackwell)

Smith, M. (2006). Is that all there is? *The Journal of Ethics*, *10*, 75–106.

Smith, M. (2007). In defence of *Ethics and the A Priori*: A reply to Enoch, Hieronymi, and Tannenbaum. *Philosophical Books*, *48*(2), 136–149.

Sobel, D. (1999). Do the desires of rational agents converge? *Analysis*, *59*(3), 137–147.

Strawson, G. (2006). Realistic monism: Why physicalism entails panpsychism. In G. Strawson & A. Freeman (Eds.), *Consciousness and its place in nature.* Imprint Academic.

Strawson, G. (2010). *Freedom and belief* (Rev. ed.). Oxford: Oxford University Press.

Street, S. (2009). In defense of Future Tuesday Indifference: Ideally coherent eccentrics and the contingency of what matters. *Philosophical Issues*, *19*, 273–298.

Street, S. (forthcoming). Coming to terms with contingency: Humean constructivism about practical reason. In J. Lenman & Y. Shemmer (Eds.), *Constructivism in practical philosophy.* Oxford: Oxford University Press.

Taylor, C. (1982). Responsibility for self. In G. Watson (Ed.), *Free will* (pp. 111–126). Oxford: Oxford University Press. (Original work published 1976)

Thomas, A. (2006). *Value and context: The nature of moral and political knowledge*. Oxford: Oxford University Press.

Van de Laar, T., & Voerman, S. (2011). *Vrije wil: Discussies over verantwoordelijkheid, zelfverwerkelijking en bewustzijn*. Rotterdam: Lemniscaat.

Van Zomeren, M., Postmes, T., & Spears, R. (2011). On conviction's collective consequences: Integrating moral conviction with the social identity model of collective action. *British Journal of Social Psychology*. doi: 10.1111/j.2044-8309.2010.02000.x.

Voerman, S. A. (2011). On the disconfirmation of practical judgements. *Logique et Analyse*, *216*, 569–587.

Voorhoeve, A. (2003). Harry Frankfurt on the necessity of love. *Philosophical Writings*, *23*, 55-70.

Watson, G. (2004a). Free action and free will. In *Agency and answerability: Selected essays* (pp. 161–196). New York: Oxford University Press. (Reprinted from *Mind*, 1987, *96*, 145–172)

Watson, G. (2004b). Free agency. In *Agency and answerability: Selected essays* (pp. 13–32). New York: Oxford University Press. (Reprinted from *Journal of Philosophy*, 1975, *72*, 205–220)

Watson, G. (2004c). Two faces of responsibility. In *Agency and answerability: Selected essays* (pp. 260–288). New York: Oxford University Press. (Reprinted from *Philosophical Topics*, 1996, *24*[2], 227–248)

Watson, G. (2004d). Volitional necessities. In *Agency and answerability: Selected essays* (pp. 88–122). New York: Oxford University Press. (Reprinted from *Contours of agency*, pp. 129–159, by S. Buss & L. Overton, Eds., 2002, Cambridge, Mass.: MIT Press)

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: The MIT Press.

Williams, B. (1981a). Internal and external reasons. In *Moral luck: Philosophical papers 1973-1980* (pp. 101–113). Cambridge: Cambridge University Press. (Reprinted from *Rational action*, by R. Harrison, Ed., 1980, Cambridge: Cambridge University Press)

Williams, B. (1981b). Ought and moral obligation. In *Moral luck: Philosophical papers 1973-1980* (pp. 114–123). Cambridge: Cambridge University Press.

Williams, B. (1985). *Ethics and the limits of philosophy*. London: Fontana Press.

Williams, B. (1995). Internal reasons and the obscurity of blame. In *Making sense of humanity and other philosophical papers* (pp. 35–45). Cambridge: Cambridge University Press. (Reprinted from *Logos*, Vol. 10, 1989, Santa Clara University, CA)

Williams, B. (2001). Some further notes on internal and external reasons. In E. Millgram (Ed.), *Varieties of practical reasoning* (pp. 91–97). Cambridge, MA.: The MIT Press.

Summary

As human beings, we are frequently confronted with questions about what we should do or what we think others should do: questions of *practical normativity*. In this thesis I develop a philosophical theory about how we can know the answers to such questions and what the nature of practical normativity is. My approach emphasizes the phenomenon of practical *disconfirmation*: the possibility for us to discover that we got our practical judgments wrong. I reflect upon both the manner in which we revise our moral views—concerning altruism and political value, say—as well as our non-moral decisions about what we want for our personal lives and careers.

In order to account for these phenomena, I argue that we disconfirm our practical judgments on the basis of unexpected affective responses to the intended consequences of our actions. I call this the "Affective Response View." On the basis of this view, I argue furthermore that we get our practical judgments right when they capture certain patterns in the structured inner realities of our affective dispositions. I call this the "Affective Pattern View." The patterns in question are those that manifest themselves in our resultant motivations in so far as we approach ideal conditions of rational and self-governing agency.

The thesis is divided into three parts. In the first part I construct a conceptual framework for thinking about truth, motivation, and reasons in practical philosophy. Chapter 1 introduces five principles that articulate intuitions about practical judgment that many philosophers have deemed independently plausible. The combination of these principles gives rise to two paradoxes, which I call the "Facts Problem" and the "Disconfirmation Problem." The first involves the following apparent contradiction. On the one hand, knowledge of certain normative *truths* would seem to motivate self-governing agents to act. But how can this be if, on the other hand, motivation is driven by contingent intrinsic *desires* that are independent from matters of truth or falsehood? The second problem involves a similar

tension with respect to disconfirmation. On the one hand, information that would *disconfirm their beliefs* about how they should act would seem to also change the desires that determine how self-governing agents will act. But how can this make sense, if they had those desires *independently* from their beliefs in the first place?

One way to solve these problems is to reject one or more of the principles that give rise to the paradoxes. This leads to well-known philosophical positions such as noncognitivism (which rejects the idea that practical judgments can be true or false) or the anti-Humean theory of motivation (which rejects the idea that action requires desires independent from our beliefs). By contrast, the purpose of this thesis is to resolve the paradoxes by explaining how all principles can jointly be true.

Regarding the Facts Problem, such reconciliatory solutions often involve some version or modification of Bernard Williams's "Internal Reasons View," which I discuss in chapter 2. According to Williams, reason statements can only be true if they are reachable from the agent's own motivations by a "sound deliberative route." Although I am sympathetic to the gist of his argument, I explicate a number of crucial differences between Williams's view and my own. One of these is the fact that I make a distinction between normative reasons and motivating reasons, and I argue that even though Williams has rejected such a distinction, his defense of the Internal Reasons View actually presupposes it. In chapter 3 I formulate a variety of the Internal Reasons View that follows from the five principles from chapter 1.

This view provides us with a sketch for a "dispositional" solution to the Facts Problem, which I discuss in part II. The idea is that we remove the mystery about why we would be motivated, under ideal conditions, in accordance with our knowledge of our normative reasons, by *analyzing* normative reasons in terms of the motivations that we would have under those conditions. I distinguish between three versions of this approach. According to "type-I dispositionalism," which I discuss in chapter 4, the truth conditions of practical judgments depend "strongly" on our contingent desires, which means that it is conceptually possible, if one agent approves of P while another disapproves of it, that both get their judgments right. I call this implication "relationalism." The problem for this approach is twofold. First, conflicts in an agent's desire set require an explanation of how one contingent desire could enjoy the authority to discredit another. Second, the possibility of irresolvable conflicts between

different agents requires an explanation of the intersubjective scope of validity that we attribute to moral discourse.

By contrast, "type-II" and "type-III dispositionalism," which I discuss in chapters 5 and 6, are "nonrelationalist." This means that in order for one agent to get his judgment in approval of *P* right, all conceptually possible agents must desire *P* under ideal conditions—a view defended by Michael Smith, amongst others. I argue that this leads to a dilemma, however. On the one hand, it might mean that all conceptually possible agents could in principle arrive at the same desires by eliminating incoherence from their contingent desire sets. This is the "type-II" view. In chapter 5 I argue that it is not plausible to suppose that this would yield convergence upon any desires of ethical substance. Alternatively, Smith could opt for the view that certain desire sets are irrational in such a way that they preclude the agent from ever deliberating onto the rational desires from his internal perspective. This is the "type-III" account. However, as I argue in chapter 6, this account requires unwarranted metaphysical assumptions that seem redundant from the perspective of type-I dispositionalism.

In part III of the thesis I develop a way out of the stalemate between these three accounts by shifting our perspective to the Disconfirmation Problem as the dialectical entry point for our discussion. I argue that the emphasis on ideal conditions implicit in the Facts Problem leaves too much to philosophical speculation, whereas reflection on disconfirmation in actual practice may provide us with independent grounds for theorizing.

In chapter 7 I argue that type-II and -III dispositionalists are committed to the "Principles of Reason View." This view requires rather substantial interpretations of rational principles in order to explain the sort of actual revisions that we want to account for. But I argue that the simplest way to explain why agents would revise their attitudes in accordance to substantial principles is simply to count those principles, contingently, amongst their intrinsically desired ends. It seems an unnecessary explanatory burden to suppose that such revisions and the corresponding motivational changes are brought about out of sheer logical or conceptual insight. Furthermore, this account presupposes, but does not provide, a solution to the Disconfirmation Problem: it fails to explain why such substantial principles would impact a change in the agent's actual intrinsic desires.

My alternative proposal, the "Affective Response View," is based on the independently plausible idea that our practical revisions often go hand in hand with our affective responses to the practical implications of our

prior views. I develop this view into an account of deliberation as "volitional interpretation," arguing that no concrete experiential or efficient motivation provides privileged access to our normative reasons over any others, making every disconfirmation susceptible to further revision in principle, and every practical belief a kind of hypothesis. This account offers a surprisingly simple solution to the Disconfirmation Problem: unexpected affective responses explain motivational changes in line with practical revisions, under conditions of sustained self-government, because under those conditions they simply *are* the motivational changes that we need to account for.

One implication of explaining moral disconfirmations in this manner is relationalism: if unexpected affective responses disconfirm practical beliefs, then practical beliefs apparently entail predictions about such responses, which makes them empirical and contingently attitude-related. Moreover, provided that we want to reconcile the principles from chapter 1, the implication is a form of type-1 dispositionalism according to which the attitudes that constitute our normative reasons may be *opaque* to us. On this view, in contrast to the relationalism defended by Williams, non-instrumental revisions are mostly not rational transformations of known desires, but *empirical discoveries* of unknown affective attitudes that reflect previously unexpected aspects of opaque volitional attitudes.

Harry Frankfurt's recent account of normative reasons, which I discuss in chapter 8, is much closer to this view. Frankfurt claims that there is a "reality within ourselves," consisting of empirical facts about what we love, that provides us with normative reasons about which we can sometimes be mistaken. He argues that love and caring are *volitional* attitudes which involve a more complex mode of wanting than the affective attitudes of mere desire, and he attempts to answer the question of how to authorize some desires over others with reference to how they relate in such complex volitional structures.

Although I broadly agree with this general picture, I argue that Frankfurt fails to account for many cases of disconfirmation, because he relies on special attitudes that grant us privileged access to our inner selves under the appropriate conditions. Furthermore, he maintains that these attitudes are neither reducible to beliefs or desires, which makes both their empirical status and their motivational role mysterious. Based on this critique, I make recommendations for an alternative theory: it should distinguish between a "cognitive" and a "normative" will, such that the former can be

analyzed as involving beliefs, while the latter may consist in structures of affective dispositions.

In chapter 9 I propose an account of practical normativity that follows the aforementioned recommendations. According to the "Affective Pattern View" the relevant dispositional structures are *patterns* in our affective lives, which manifest themselves as we increase our self-understanding. I borrow the idea of an ontology of patterns from Daniel Dennett, but whereas he has used it to explain *desires* as *behavioral* patterns, I incorporate various adjustments so as to explain *volitional* attitudes as *affective* patterns.

The result is an account that allows me to explain intuitions about both the determinacy and the indeterminacy of moral choice. On the one hand, I argue that my view can account for the idea that most Nazi officers in the Holocaust got their practical judgments wrong, without postulating nonrelationalist moral facts (leaving room for a possible divergent psychopathic minority). On the other hand, my view accommodates the experience of moral dilemmas, which leave the right choice indeterminate even at the 'intrapersonal' level. Furthermore, my account also combines intuitions about deliberation and decision-making as 'self-finding' and as 'self-making.' I argue that while the cognitive will can get the normative will wrong in some cases, there are other cases in which the latter may be *shaped* by the former, such that our deliberations also play a constitutive role in the genesis of our normative reasons.

With this account in place, I reflect further on Frankfurt's ideas about *wholeheartedness*. I distinguish between "inner wholeheartedness" and "epistemic resolvedness," arguing that neither are to be pursued too fervently. Instead, I claim that allowing ambivalence in our hearts may actually be a form of authenticity, of being true to the divided nature of our selves. Furthermore, I discuss the potential harm of eradicated doubt in the light of our often heavily biased emotional mechanisms.

With the Affective Pattern View in place, it is now understandable why, the more an agent knows the facts about his normative reasons, the more his capacity for self-government would involve desires to act upon those reasons. This is because those facts are facts about a pattern of desires that manifests itself more fully as his knowledge increases and which seems the proper object of his volitional interpretation, representing what he is really all about as a person. Hence, it seems to make sense of the idea of self-government to say that if his volitional beliefs capture this pattern, he will be self-governing when his resultant desires participate in that pattern,

and that he will be lacking in self-government on the occasion when his resultant desire constitutes noise relative to that pattern. In this manner, the view explains why some desires have authority over others, which provides us with a type-1 dispositional solution to the Facts Problem.

In chapter 10 I discuss the one remaining difficulty, which is the intersubjective dimension that relationalist theories are allegedly poorly placed to explain. I formulate two possible objections against relationalism on the basis of various arguments from Michael Smith. The first is that moral discourse would be without purpose if relationalism were true. My first counter-argument is that we have many good reasons to assume shared volitional attitudes on many, if not most, occasions. In particular, it is plausible that we share certain basic moral values as a species. Second, I argue that there are several further reasons for engaging in moral discourse even in those cases where our values may turn out to differ.

The second objection is that the words or concepts used by people in moral discourse, such as "right" and "wrong" or "good" and "bad," simply have nonrelationalist meanings, regardless of whether purposeful discourse about relationalist concepts would be possible or not. Rather than simply denying this outright, I admit that some people may indeed mean their judgments in a nonrelationalist sense. However, other people do not: I propose a semantic pluralism about what people actually mean, and a conceptual revisionism about what people *should* mean when they make their practical judgments.

Thus, in chapters 9 and 10 I have addressed both of the difficulties for type-1 dispositionalism that were formulated at the end of chapter 4: the authority of some desires over others and the intersubjective relevance of practical judgments in moral discourse. The resulting account continues the tradition of Humean thinking in moral philosophy, both with respect to the distinctness of motivation and the relationalism about normativity. But because of its strong emphasis on volitional opacity and interpretative self-understanding, the account has provided us with a new way to tackle some of the problems that have traditionally been associated with Humean approaches.