Affect, Representation, and the Standards of Practical Reason

by

Paul S. Boswell

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2016

Doctoral Committee:

Professor Peter A. Railton, Co-Chair
Professor Sarah Buss, Co-Chair
Professor Allan F. Gibbard
Professor Sekhar Chandra Sripada

2016

To my parents
who deserve all that I can give.

A C K N O W L E D G M E N T S

People

This dissertation and I have taken a long and winding route, and now that the journey is over it is my distinct pleasure to thank those who have provided me with essential help and encouragement along the way.

I would first like to thank my dissertation committee: Peter Railton, Sarah Buss, Allan Gibbard, and Chandra Sripada. I came to Michigan to work with Peter largely on the basis of his philosophical work, and it goes to show how able of a philosopher he is that what I have come to appreciate most about working with him is his ability to impart the significance of, and communicate the motivations behind, philosophical views that neither of us may ultimately share. Sarah, in turn, was a perfect complement to Peter as a co-advisor, as she provided critical guidance and wisdom in clarifying and extending my thought. Her dedication to her students is extraordinary. Allan has had a deeper impact on my philosophical development than he may realize, as I spent much of my time in graduate school grappling with his expressivist program in metaethics. Finally, Chandra (along with Peter) has always been a model for how to combine rigorous and penetrating philosophical work with empirical research.

Before I left Princeton for Michigan, my undergraduate thesis advisor Michael Smith told me that I was likely to learn as much from my fellow graduate students as I was from the faculty. He was right. Throughout my six years at Michigan I have been blessed to be surrounded by such good philosophers and colleagues. Rohan Sud has a wonderful way of providing encouragement along with truly helpful feedback. Daniel Drucker, polymath and philosophical lodestar, sharpened my philosophical skills. Others who have helped me become a better philosopher include Chloe Armstrong, Dmitri Gallow, Nils-Hennes Stear, Steven Schaus, Chip Sebens, Mara Bollard, Patrick Sherriff, Damian Wassel, Ira Lindsay, Will Thomas, Umer Shaikh, Ian Fishback, Billy Dunaway, Sara Aronowitz, and Warren Herold.

It is a privilege to have benefited from the completely supererogatory help of the wider philosophical community. Uriah Kriegel provided a welcoming presence during my time in Paris, and helped me better see the dialectic of some of my own arguments. Sergio Tenenbaum was very generous to me in providing kind and helpful feedback. Special thanks also goes to Chris Howard, Daniel Wodak, Adam Lerner, Victor Kumar, Sam Asarnow, and Sabine Döring. (I will also take

this opportunity to thank Dale Dorsey, from whose book the structure of this acknowledgments section was lifted.)

Last but not least, I could not have completed this dissertation without the material and emotional support of friends and family. First among these are my parents, Mark and Deborah Boswell, and sister, Lesley Boswell, who have always been there for me. I am also especially glad to have had the support of Ann Gong, Siddharth Bhaskar, Nina Frey, Laura Lanzoni, Nadan Sehic, Jack Morgan, Laura Grigereit, and Harry Gaples and Christy Lange.

Places

Large parts of this dissertation were composed in Sweetwaters Coffee, Black Diesel Coffee, Roos-Roast, and Mighty Good Coffee (Ann Arbor); Kaldi's Coffee (Chesterfield, Missouri); Astro Coffee (Detroit); and Coutume Instituutti (Paris). Thanks to these establishments for putting up with me and my computer. Thanks also to Ann Arbor for being a pretty great place.

Things

Work on this dissertation was completed with the help of a Chateaubriand Fellowship from the Embassy of France in the United States. Thanks also to Springer for permission to reprint "Making Sense of Unpleasantness: Evaluationism and Shooting the Messenger", which is forthcoming in *Philosophical Studies*. It appears here as Chapter 2 and is edited only for style and minor typographical corrections.

TABLE OF CONTENTS

LIST OF APPENDICES

ABSTRACT

Affect, Representation, and the Standards of Practical Reason

by

Paul S. Boswell

Co-Chairs: Peter A. Railton and Sarah Buss

How does human agency relate to the good? According to a thesis with ancient pedigree, the connection is very tight. Known as "the Guise of the Good" (GG), it states that human action or motivation to act, of some special kind or another, is only possible insofar as the agent performs or is motivated to perform the act because of the good she sees in so acting. But how might agents see their actions as good? Recent research in moral psychology, the philosophy of mind, and the cognitive sciences suggests that affective states may play a deep role in cognition and action as *representations of value:* for instance, pain may represent an injury as *bad for one*. This dissertation begins by defending just such an evaluationist account of unpleasant pain from an objection, and then develops and defends an affect-based version of GG.

The first part of the dissertation (Chapter 2) considers a foundational problem for an evaluationist theory of affect. The theory is motivated by its ability to make sense of our aversive intentional responses to pain as responses to value, but the shooting the messenger objection charges that it is unable to make sense of our aversive behavior to the sensations themselves. I propose a solution to this problem on behalf of the evaluationist: when we introspect our pains we also turn our emotional distress inwards, enabling it to represent our pains as bad.

One crucial question GG theorists must face is just what the good of GG is. Chapter 3 argues that, lest the thesis be too weak, it must hold that actions must appear to their agents to meet a standard of practical reason. The chapter then shows how the intelligibility motivation for GG can lead naturally to the view that the standards so presented are shared publicly. Chapter 4 argues against the standard understanding of GG in terms of essentially evaluative desires and contends

that it should be replaced by *hard-line affectivism*, the view that GG is true because actions are based on affective states that represent there as being reason for those actions.

CHAPTER 1

Introduction

Our actions make sense to us, at least when we do what we mean to. But *what is it* for an action to make sense to its agent, and how do we make sense of our actions? These questions are not exactly the same. The first seeks a constitutive account of an action's making sense to its agent, an account of its very nature. The second question admits of the possibility that there may be a particular way in which we human animals make sense of our own actions. Human animals are a special kind of agent, and it may be that we have a special way of acting and of making sense of our actions.

Philosophers of action, however, have tended to pursue the first question at the neglect of the second. The standard tool chest of the philosopher of action includes beliefs, desires, and intentions as mental states. Suppositions, hunches, plans, emotions, and traits of character can be found in deluxe packs, but philosophers take pride in doing more with less — and they have found plenty to work on even with beliefs and desires alone. And the interesting thing about beliefs, desires, and intentions is that, at least as most philosophers understand them, they are quite general to agency. More cautiously, they must be very common to all agents who play certain very abstract functional roles, such that of an agent who is effective in the world or who is limited in reasoning powers and must plan over time.

For example, it may well be that full accounts of belief and of desire as they are realized in humans would be rather complicated and appeal to peculiarities of human psychology, but most philosophers' arguments depend only on an understanding of them in terms of simple dispositions. Michael Smith (1994, 115) offers the following platitudes about beliefs and desires: "a belief that *P* tends to go out of existence in the presence of a perception with the content that not *P*, whereas a desire that *P* tends to endure, disposing the subject in that state to bring it about that *P*." Given that any agent will need to perceive the world in order to effectively act in it, it is quite difficult to imagine an agent without such dispositions.

Or take an influential job description for intention due to Michael Bratman (1987, 15-17): intentions are conduct-controlling pro-attitudes which tend to cut off deliberation and resist reopening the question of what to do, and they also involve dispositions to reason to further intentions.

These are important capacities, and arguably crucial to any agent who must coordinate with herself and others over time. Finally, many philosophers are happy with explanations that appeal to such broad capacities. John Broome (2013), for instance, sets himself the task of explaining how it is that often when we believe that we ought to do something, that belief causes us to intend to do that thing. The explanation he defends is, in a nutshell, that often we bring ourselves to do this through reasoning. This is not a toothless explanation — one could certainly imagine creatures for whom the same causal link is wholly unmediated by reasoning, and Broome has much of value to say about the nature of reasoning — but it is a rather bloodless one.

Because of their emphasis on characterizing agents with respect to their abstract, functional capacities, such approaches reflect what might be called a "function-first" program for understanding the psychology and explaining the behavior of agents. One methodology in the function-first program is conveniently encapsulated by Paul Grice's (1974, 36-41) hypothetical "creature construction" process. Grice proposes that in order to devise and defend a psychological theory for actual beings, we should imagine constructing a creature roughly of that type. In order for the constructed creature to serve as a model for actual beings, we should begin with three things: (a) a description of the creature's environment, especially the opportunities, restrictions, and perils it affords the creature; (b) a description of the limitations of the creature, including its physical and physiological constraints; and (c) a description of the function that the creature is to carry out in that environment, which Grice supposes must at the very least include survival and reproduction.[1] The idea is then that as designers we would devise for our creatures only the psychological tools necessary for it to carry out its function in its environment, which ensures that the psychological states that we appeal to have a role in explaining the behavior of the actual creatures we wish to model.

Creature construction is a powerful method for psychological investigation, but it has a limitation: it depends on *first* specifying the function the creature is to carry out. With respect to human agents this is not a problem for the kinds of functions Grice has in mind, for we know that human agents survive and reproduce. And there are other functions too which we know humans perform, such as planning over time. But many of the interesting philosophical questions about agency concern deciding among different ways of specifying the functions that agents perform, many of which do not clearly demand different kinds of observable behaviors from the creatures that carry them out. Do we deliberate about ends, or only about the things towards the ends? Do we generally deliberate *about* our desires, or do desires normally operate in the background of deliberation? Should we think of human agents as directed towards producing good in the world, or as satisfying their own desires, or as specifying and carrying out the actions they intend to do,

[1] I have extracted condition (b) from Grice's discussion of condition (a) and from his description of the "engineer" of the creature, who differs from the designer. The difference between these two roles does not matter here, however.

or what?

To answer such questions we may need to complement our investigation with a *psychology-first* approach. We start with psychological states we know that human agents have and, applying what we know from our best psychological and philosophical theories, see what role these states are best thought of as playing. The question here is not "What kinds of psychological states and laws would best explain how this being can behave so as to carry out its function?" but, roughly, "What function for these states would best fit our best philosophical and psychological theories?"[2]

This dissertation is, in large part, an application of the psychology-first approach to *affective states* — that is, roughly speaking, valenced psychological states such as emotions, moods, and pleasures and unpleasant pains. In it I argue that affective states play a deep role in human agency, especially in motivation and in making sense of actions, through their capacity as representations of value.

In developing this view I rely upon two general philosophical programs. The first, which only plays a significant role in Chapter 2, is intentionalism. This is roughly the thesis that phenomenal character — the what-it's-likeness of experience — is exhausted by representational content. On this view, phenomenal experience is the head-up display of the mind. The particular version of the view defended in Chapter 2 is *evaluationism*, which holds that unpleasant pains as affective states are evaluative representations, and in particular that they represent an alteration in the body as bad. Recent critics of evaluationism have held that although the view may explain why it makes sense to us to respond to our bodies when we are in pain, evaluationism cannot explain why it makes sense to get rid of the pain itself. Since the pain is merely a representation on this view, getting rid of the pain would seem to amount to shooting the messenger. Chapter 2 clarifies the nature of the objection and offers a solution. It shows how a proper understanding of introspection and the role of emotion in pain experience allows the evaluationist to defuse the objection.

The second philosophical program, the Guise of the Good theory, takes center stage in the two chapters that follow. An old family of views that can claim among others Plato, Aristotle, and Aquinas as illustrious proponents, it holds that we only act or are moved to act because we see some good in doing so. However, it is often unclear exactly what unites the views in this family and what theoretical benefits it is intended to bring. Chapter 3 argues that these views should be considered as united in proposing that actions appear to their agents to meet a standard of practical reason. In it I develop a popular motivation for the view, the intelligibility motivation, and demonstrate how it could be used to derive a surprisingly strong restriction on the kinds of goods which may motivate agents. I argue that a plausible result of the intelligibility motivation is that the good of the Guise

[2]To be clear, I am not expressing opposition to functional or dispositional accounts of mental states. My point is the epistemic one that we are unlikely to address all important questions in philosophical psychology by starting with the functions which we know an agent to carry out.

of the Good is a fully public good, one whose normative force extends to all agents.

Finally, Chapter 4 argues against a standard understanding of the Guise of the Good and offers an affect-based version in its place. Typically the view is held to be a *necessary* truth about rational motivation, and the the way in which actions are standardly held to "appear good" to their agents is by holding that *desires* are essentially evaluations. I demonstrate that there is very good reason to deny both theses. I instead propose a view*, hard-line affectivism*, which holds that human actions are based in affective states that represent there as being reason for those actions. Hard-line affectivism thus secures a non-necessary version of the Guise of the Good. The upshot of such a view is that actual humans, at least, cannot be entirely counter-normative agents. When humans act for a reason, they take their action to be supported by a normative reason for doing what they do. Finally, Chapter 5 concludes and takes note of some outstanding issues.

Reading notes

Although Chapters 3 and 4 complement each other, all three of the main chapters were written so that they could be profitably read independently of the others. For that reason a few minor points made in Chapter 3 are repeated in Chapter 4, and definitions are sometimes also repeated.

Reason

Philosophers typically distinguish among at least three kinds of reasons: normative, motivating, and explanatory. An explanatory reason explains a fact: "Why are leaves green?", you ask. I reply, "Because plants generate energy by photosynthesis, photosynthesis relies on chloroplasts, and chloroplasts contain chlorophyll, which reflects green light." Here I have explained why a fact obtains by citing more facts, which together form the reason why it obtains. The general category of explanatory reasons will not play a significant role in this dissertation.

A motivating reason is a kind of explanatory reason which usually shows an action to be intentional. As Anscombe (1963) would put it, these reasons are the considerations that an agent might cite in response to someone asking her why she did what she did. One must be careful, however, since not every true and enlightening explanation of action involves giving a motivating reason for it. "The workers revolted to throw off the yoke of capitalist oppression" may offer a motivating reason if it means that most of the workers intended by revolting to throw off the yoke of capitalist oppression, but it may not if it means to analyze the social forces at play in Marxist terms the workers themselves would reject. And not every explanation of an action in terms of the agent's mental states gives a motivating reason: "He went berserk because he was overcome by the stress caused by working at the bank" does not give a motivating reason. It is not the agent's reason for doing what he did, for he did not go berserk for any reason at all. Finally, a normative reason is

a reason that counts in favor of something. "Why eat kale?" "Because it's healthy." Normative reasons *pro tanto* justify actions, beliefs, intentions, and the like.

In this dissertation, 'reason' will by default mean a normative reason for acting. However, the word 'reason' in 'acting for a reason' refers to an agent's motivating reason for acting. As discussed in §3.2.2, the phrase 'the standards of practical reason' refers to to the standards for normative assessment of action and intention. I will often use 'practical reason' as a synonym for the standards of practical reason, though sometimes I will use it to refer to the study of those standards.

Finally, I will occasionally write of an agent *having* or *possessing* a reason, and of something's *giving* a reason to an agent. Sometimes these terms are used in slightly different ways in philosophy, so I will here indicate how I use them. There can *be* reasons for an agent to act which she does not possess. One reason for an officer to discipline a soldier is that he has disobeyed her command, but this is not a reason she *has* if she is unaware of his disobedience. Something *gives* her that reason if it affords her access to it, such as evidence she possesses or beliefs she has that conjointly entail that he has been disobedient.

Good

Evaluative terms such as 'good', 'bad', and 'value' must be used with less consistency in this dissertation, unfortunately. Chapter 2 adds to a preexisting literature on the nature of pain in which evaluationism is a major player. The view I ultimately hold, and which I defend in §4.6.3.2, is that unpleasant pains represent *reasons*. This view is not much represented in the current literature on pain and the problem that I solve in Chapter 2 is a problem for both views, and so I present that Chapter as a defense of evaluationism — though I do signal my interest in what I call Broad Evaluationism, the view that affective states represent justifications for acting. Another difficulty is that in spite of its title, many Guise of the Good theories do not work with specifically evaluative properties. One can be a Guise of the Good theorist by holding that in acting one sees one's action as permissible, for instance.[3]

But what should we call the Guise of the Good if it does not entail that one act under the guise of the *good?* We might reasonably retain the name while we discuss the view. Since there are reasons to use evaluative terms to pick out the evaluative and reasons to extend their use much more broadly, I have decided to do both:

[3] Schroeder (2008, 127) notes that Sergio Tenenbaum's (2007) defense of the Guise of the Good relies on a notion of *good* that behaves more like *permissible*.

NOTE TO THE READER

In Chapter 2, evaluative terms mean what they usually do. In the Chapters that follow Chapter 2 evaluative terms take an extended sense, much as 'normative' can denote a much broader category than the narrowly deontic. When I mean to denote the narrowly evaluative concepts — that is, the usual evaluative concepts — I will then use an asterisk, e.g. 'good*'.

On stylistic conventions

I use italics both for emphasis and, depending on the context, in order to name the concept, principle, or proposition expressed by the italicized phrase. For instance, '*good*' is equivalent to 'the concept of the good'.

Single quotes mention the quoted expression. Double quotes may use or mention the quoted phrase and sometimes may do both at the same time, as in scare quotes. In both single and double convention I do my best to include punctuation within the quotation marks only if it would appear within the quoted expression if used without quotes.

CHAPTER 2

Making Sense of Unpleasantness

2.1 Introduction

Pain and other unpleasant sensory experiences intuitively seem to bear a tight normative relation to certain aversive actions. I do not respond strangely to scalding bathwater by leaping out of it and then hopping around to distract myself from the pain, at least under the circumstances. Nor does it seem to me that I do. But someone — call him Yellowman — who flails about in order to avoid yellow things and experiences of yellow *is* acting strangely, and surely would *see himself* that way too. (It is crucial for my purposes that we not think of Yellowman as finding yellow unpleasant in any way. His experience of yellow is just like ours — except that he has dispositions we do not.) In contrast to the experience of yellow, my pain makes sense of both my body-directed aversive behavior (leaping from the bath to save my skin) and my experience-directed behavior (distracting myself from the pain). That is to say, these behaviors both seem to me to be *pro tanto* justified and really are *pro tanto* justified.[1]

So I take it that one task of a philosophical theory of unpleasant sensations is to explain how our unpleasant sensations give us access to a justification for *both* body-directed and experience-directed behavior.[2] How can we account for the way in which, in experiencing unpleasant sensations, we quite generally have access to a genuine justification for these kinds of actions? I'll call this question **Q1**.

A new and hotly-contested view hopes to give a satisfying answer to just this question. Call a view *evaluationist* if it holds:

[1] Hereafter I'll often drop explicitly mentioning that the kind of justification I have in mind is *pro tanto*. It also bears noting that I am throughout only concerned with our *intentional* responses to pain.

[2] By 'access to a justification' I mean more than *de re* awareness of something that *is* a justification; the subject must be aware of it *as* a justification. But otherwise I use 'access' as a placeholder for whatever sort of epistemic state the above description picks out. I am not here concerned with determining the quality of our access, e.g. whether it constitutes knowledge.

Evaluationism Necessarily, a sensory experience's being unpleasant amounts to its being a certain kind of representation as of an event in or near the subject's body being bad for the subject.[3]

The hedge "a certain kind of representation" alludes to conditions on the manner of representation. So, for instance, Tye (1995b) holds that, among other things, phenomenally conscious representations must be *abstract*, *non-conceptual*, and *poised* to affect a subject's conceptual systems in a certain way. The hedge is necessary to avoid over-inclusion, since obviously a pictorial representation as of a wound's being bad for the wounded is no unpleasant sensation itself.

Recent evaluationists include Helm (2002), Tye (2005), Cutter & Tye (2011), O'Sullivan & Schroer (2012), and Bain (2013).[4] Because so much more is known about pain than other unpleasant sensations, pain has monopolized the literature's attention, and it will be the focus of this chapter as well. Evaluationists typically hold that pain in particular represents an alteration or injury in the body as bad for the subject.

We can now see the outlines of a response to our initial question: if we can show that unpleasant sensations are representations of events as bad, and if we can show those events to generally *be* bad, we not only will have given a justification for our aversive behaviors, we will have explained our access to that justification too.

But note that our concern so far has been with *sensory* experiences, even though *all* unpleasant experiences, worries and fears as well as pains and itches, seem to bear a similar normative relation to action.[5] Anxiety typically prompts world-directed behavior such as avoiding other people, and it also motivates attempts to relieve oneself of that very anxiety. So one reason to favor evaluationism is that it coheres well with an attractive general account of unpleasant experience:

Broad evaluationism Necessarily, an experience's being unpleasant amounts to its being a certain kind of representation of a *pro tanto* justification for aversive behavior.[6]

There is another reason to hope the broad evaluationist project succeeds, too. Unpleasant experiences clearly motivate action, and so can be used to explain it. But if these experiences contain evaluative representational content, then they also show us the way in which the action *seemed good* to the agent, and so they suffice for a *rationalizing* explanation of that action. The result is that many cases of intentional action that have thus far resisted subsumption under the

[3] As I use it, the schema "*X* amounts to *Y*" is neutral as to whether *Y* explains *X* or *vice versa*. See the discussion on p. 15 below. I will also suppose throughout that the badness represented by pain experiences, as well as the sense in which pains really are bad, is badness *for* the subject of the pain, though for verbal economy I will generally leave out this qualification.

[4] Critics include Aydede (2005), Jacobson (2013), and Cohen & Fulkerson (2014).

[5] Among evaluationists only Helm (2002) has thus far shared this broader concern to extend the account to unpleasant experience generally.

[6] Obviously, something similar can be said for pleasant experiences. I will follow the literature in focusing exclusively on unpleasant experiences and in hoping that the account developed can be extended in a fairly straightforward way to pleasant ones.

story of rational action will turn out to be explicable by it after all. For instance, Hursthouse (1991) famously accuses some cases of emotion-driven intentional actions, such as that of a woman angrily defacing a photograph of her enemy, of being "arational" because they seem to require us to attribute absurd beliefs to the agent if we are understand them as aiming at some good. (Must the woman believe that the photograph *is* her enemy? Certainly not.) But attributing such a belief is unnecessary if her anger *already* presents her action as good.[7] And if the representational content of unpleasant sensation is non-conceptual, then we have a rationalizing explanation of small children's intentional responses to pain.[8]

Evaluationism thus offers a scheme for tackling at least one difficult question, and it promises fruitful explanations down the road. Yet its progress has recently been stymied by the objection that it endorses *shooting the messenger*.[9] Consider Siddharth and Ann, both of whom suffer from occasional shooting pains in the left knee. Siddharth's pains are caused by small pieces of shrapnel around the joint that were acquired in a war, and when he feels them Siddharth knows to keep pressure off his left leg to avoid aggravating any damage. In the stream of Siddharth's experiences, his response is immediate: there is the pain, and then there's his keeping off his leg. Here evaluationism gives just the sort of response to Q1 we are hoping for: the pain represents the existence of a bad alteration, and this motivates Siddharth to address his injury. Moreover there *is* an injury in his knee, and it is bad for him. So the evaluationist can explain how Siddharth is accessing a justification for the action he is motivated to perform.

But poor Ann suffers from a disease whose only symptom is the shooting pain. Not that there's anything wrong with her knee. The root of her problem is in her somatosensory cortex, where some crossed wires confuse pressure for pain. But Ann is well aware of her condition, and when she gets the pain in her knee she is inclined, not to stop walking and inspect it, but to reach for an oral painkiller she keeps in her pocket.

Now, the problem for the evaluationist is that:

1. There must be similar explanations of Siddharth's and Ann's actions. Intuitively, in both cases the response is *practically immediate*: as soon as they experience the pain, or are aware of their pain, they engage in aversive behavior.

2. The explanation of Ann's action must satisfactorily answer Q1. For Ann's pain itself really is bad, and in taking the painkiller as she does Ann is accessing that justification for acting.

3. It seems the evaluationist cannot offer similar explanations of Siddharth's and Ann's actions that give satisfactory answers to Q1.

[7] Döring (2007) sees related opportunities for affect to play an explanatory role in action.
[8] See also Tappolet (2000, 178-183) and Hawkins (2008).
[9] See Jacobson (2013) and Aydede & Fulkerson (2015).

The worry behind 3 is that extending the evaluationist's explanation of Siddharth's action to Ann's yields but two events: the tokening of an experience that represents a harmful event in Ann's knee, followed by an immediate attempt to get rid of that experience. And that, it seems, makes about as much sense as shooting the messenger who bears bad news. Even if we know the message to be false, what justifies shooting the messenger? To the extent that evaluationism offers an answer to this question it seems to wrongly accuse Ann of confusion, for it only explains Ann's access to one (apparent) reason for acting, *that there's a bad alteration in her leg* — which is surely no reason to get rid of her *pain*.

Responses to this problem thus far in the literature have not been encouraging. O'Sullivan & Schroer (2012, 755) seem to endorse shooting the messenger. Bain (2013, S87) asserts that there is reason to shoot the messenger because it is intrinsically bad to be in a state in which something else seems bad for you, even if you do not believe it to be bad. I find the premise here questionable. There are, after all, less affect-laden ways of entertaining normative propositions. "The polls might look bad for us", the adviser tells the politician, "but given the demographics, I'm confident we will win". Surely it would be odd to say that the adviser is in a bad state.

Cutter & Tye (2014, 428) are prepared to deny that there is anything bad about unpleasant pain itself, and regard non-instrumental aversive behavior to it as *a*rational. But one might reasonably complain that this fails to fully address Q1. True, as Cutter & Tye point out, pain can have negative consequences: it is distracting, and chronic pain can cause a host of problems. At best this establishes that unpleasant pain is *extrinsically* bad, and that it would be good to learn to avoid it, much as it is good to learn to be careful with sharp knives. But unpleasant pain seems to be importantly different from a sharp knife in that it is bad intrinsically, *of itself*, and that we can appreciate its badness by being aware of it.

I have two goals in this chapter. First, I aim to get some clarity on just what the shooting the messenger (STM) objection amounts to. There are in fact two parts to STM: (a) what can the evaluationist say about the *badness* of pain? and (b) how can she explain our *access* to that badness? The literature has concentrated on the badness problem, but I give reason to think that an adequate response is within reach. I argue that it is the access problem that is especially difficult.

Second, and more importantly, I offer a solution on behalf of the evaluationist to the access problem. The clinical literature on pain tells us that pain's unpleasantness really has two components, sensory and emotional. I argue that when we introspect our pains the intentional object of our emotional unpleasantness is the pain itself, and thus represents it as bad.

§2.2 addresses some background issues concerning the phenomenology of unpleasantness and the commitments of evaluationism. §2.3.1 addresses the badness prong of STM, and §2.3.2 argues that the seriousness of the access prong has been underappreciated. §2.4 develops and defends the solution to the access problem, and §2.5 concludes.

2.2 Preliminaries

2.2.1 The phenomenology of unpleasantness

Following the clinical and philosophical literature on pain, in this section I distinguish pains from *unpleasant* pains and contrast the latter to unpleasant emotions. The main conclusion I draw is that our access to unpleasant emotion, unlike to unpleasant pain, must be inferential — where "inferential" is used in a broad sense to include abductive inferences and guesses based on evidence, and more generally any sort of reasoned transition from thought to thought. This result is used in §2.4.4 to explain why the solution I propose to STM does not simply move the bump in the rug and push the objection upwards to unpleasant emotional states.

It has long been known in the clinical research on pain that it is possible to experience pain-like sensations that are not unpleasant. The experiences of patients of frontal lobotomies, those with congenital indifference to pain, and especially pain asymbolics all challenge the notion that pain *by itself* possesses any special normative property.[10] Asymbolics and lobotomy patients use strikingly similar language: "I feel the pain, but it doesn't hurt".[11] This, together with the fact that what it's like to have one of these conditions is clearly very different from what it's like for the rest of us, suggests that asymbolics and the lobotomized share with us some phenomenological component of our overall pain experience, *the pain sensation*, and wholly or partially lack some other component, *the unpleasantness*.[12] This comports well with clinical research on pain, which divides normal pain experience into sensory-discriminative and affective-motivational components,[13] and also reflection on our phenomenology: some pains are more sharp than unpleasant, and sometimes the unpleasantness of a throbbing pain fades before the throbbing does.

Here I take this standard interpretation of these dissociation cases and their connection to normal pain experience for granted. One's overall *pain experience*, I will say, is generally a composite of a *pain sensation* and *unpleasantness*, though some pain experiences involve pain sensations with no unpleasantness. So token pain experiences, or *pains*, may be unpleasant or not, and when they are not unpleasant they feel quite a bit different from normal pain.[14] Finally, I also take it for granted that unpleasantness is affective: unpleasant experiences feel *bad*, where 'bad' here

[10] For a case of congenital indifference to pain, see Frances & Gale (1975). For lobotomy as a treatment for chronic pain, see Freeman & Watts (1950); Hardy et al. (1952). For asymbolia, Schilder & Stengel (1928); Weinstein et al. (1955); Berthier et al. (1988).

[11] Compare the case reports in Freeman & Watts (1950) with those in Berthier et al. (1988).

[12] See Grahek (2007) for an influential view of this kind, though at times he does seem uncertain whether asymbolic pain is indeed pain; see *op. cit.,* 111-12.

[13] See Fields (1999); Price (2000).

[14] Klein (2015) has recently questioned whether asymbolics' pains feel different from those of non-asymbolics, but I fail to see how the truth of his positive account — that asymbolics lack concern for their bodily integrity — undermines the standard interpretation. One significant symptom of asymbolics' lack of concern is their total lack of pain affect. See Bain (2014) for further criticism.

describes its negative phenomenological valence, not its normative status.[15]

Now, what should we say about how pain and its unpleasantness are related? First consider how attention relates the two. Recall the last time you stubbed your toe: a sharp, relatively well-localized and only moderately unpleasant pain was followed about a second later by a more unpleasant and diffuse throbbing pain that radiated outward from the injury.[16] The latter kind of pain is more memorable, so concentrate on the way that felt. Do you think you could switch your attention from the sensory aspect of the pain to its unpleasantness, and back again? No, at least not more than you can switch your attention from the timbre of a trumpet blast to its pitch. In both cases you are attending to what seems to be a single spatially-located event with two aspects.[17]

Next, consider the nature of our access to our unpleasant pain. Famously, it seems that to have an unpleasant pain *just is* to have an experience with the phenomenal character of an unpleasant pain, and that introspection can make us aware of an experience with the phenomenal character of a pain *as* an experience with that character.[18] Since our introspective awareness of our phenomenal experiences is non-inferential, we have non-inferential access to our unpleasant pains *as* what they are — that is, as experiences with their particular unpleasant character.[19]

But unpleasant *emotions* contrast with pains in both respects just considered. Take your fear of a snarling dog, and note that whereas the unpleasant pain seemed to be about a body event in part in virtue of seeming to be located there, your fear does not seem to be about the dog in virtue of seeming to be co-located with it. The dog is before you and the fear inside you, if anywhere. In contrast, the unpleasantness of your fear is not experienced as having a specific location. After all, *what* is unpleasant about your fear? What should you attend to if you wanted to attend to the unpleasantness of your fear?[20]

Also, how might you come to be aware of your fear *as* what it is — that is, as an episode of

[15]It should be noted that some motivation or attitude-based theories of unpleasantness seem to deny that unpleasantness itself contributes to phenomenal experience; see for instance Tye (1995b, 135), Clark (2005), and Heathwood (2007). But lest they deny what seems to me to be an obvious truth about phenomenology, these theories are often better construed as offering a reductive account of unpleasant experience in terms of motivation or an attitude.

[16]See Price & Aydede (2005, §4.1) for an overview of psychophysical studies concerning these two pains.

[17]It's possible that the phenomenal character of pain and unpleasantness are not experienced as exactly co-located: perhaps the pain is limited to the area of perceived damage while the unpleasantness radiates further outward from it. In that case it would be hard to say that the unpleasantness is an aspect *of the pain*. I'll ignore this complication in what follows, since what matters for my purposes are the claims about the location of unpleasantness and how pain and its unpleasantness are "bound together" as about the same thing. There is some psychophysical evidence that the unpleasantness itself is experienced as having a body location. Ploner et al. (1999) describe a case study of a stroke patient who, for some nociceptive stimuli to the hand that would normally be painful, experienced an unpleasant sensation "somewhere between fingertips and shoulder" (*ibid.*, 213) — but no pain.

[18]*Pace* Rosenthal (1991, 17), who suggests that pain can be unconscious if it is unnoticed.

[19]Note that I am assuming only that our introspective access to our *phenomenal experiences* is non-inferential, not that introspection of *any* mental state must be non-inferential. Determining the nature and scope of introspection is beyond the scope of this chapter.

[20]It is instructive to compare the introspectionist Titchener (1896, 96) on this point, who writes of the unpleasantness of affection that it "pervades the whole consciousness of the moment".

fear? If you attend to where in your body you feel the fear, you might find a tightening in your chest.[21] But surely that alone does not license you to infer that you are afraid. That feeling is common to other unpleasant emotions, such as anxiety and surprise, and to non-emotional states such as angina. Similarly for other aspects of the overall experience of fear, which include an attentional bias towards the dog, a feeling of arousal, and the adoption of a new goal to get away from the dog together with an inclination to pursue that goal immediately.[22] Even if certain processes sufficient for an experience of fear each contributed to the phenomenal character of one's experience, and even if the contributions of each of these processes are each accessible via introspection, it seems we would only become aware of the emotion as the emotion that it is by being aware of its phenomenal markers and *inferring* that we must be experiencing the emotion of which they are markers. Moreover, many of the processes which arguably are necessary components of the experience of an emotion like fear (e.g. a change in one's attentional bias or in the goals that one is inclined to pursue) may not have a phenomenology.[23]

On a somatic feeling theory of emotions, according to which emotions are a kind of perception of bodily states, introspection may at first seem to give us better access to our emotions. If the emotion of fear *just is* the feeling of one's heart pounding, one's palms sweating, one's veins constricting, etc., then isn't it the case that we have introspective access to that set of feelings as we experience them? Here it is important to note that, even according to the contemporary defender of the strictest version of this view, Jesse Prinz (2004a,b), emotions are not *just* occurrent, conscious perceptions of bodily states. According to Prinz, some emotions are dispositions to feel (Prinz 2004a, 50) and even occurrent emotions need not be consciously felt (Prinz 2004b, Ch. 9). We will not have introspective access to these emotions, it seems. Also, on his account a bodily perception is an emotion only if it is characteristically caused by certain organism/environment relations, for that is part of what makes such a perception an emotion (Prinz 2004a, 53). But it seems it would be a category mistake to suggest that we have *introspective* access to the characteristic environmental causes of our perceptual states. So, even on Prinz's view we do not have introspective access to our emotional states *as* emotional states, which is the key claim defended above. Thus, the truth of a somatic feeling theory of emotion would not threaten the claim that we do not have that kind of introspective access to our emotions.

On the whole, it seems that our access to an emotion is at a distance from our access to the phenomenological aspects of that emotion. Our emotions are not simply and obviously reflected in the character of our experience, and therefore we must infer their existence *from* that experience.[24]

[21] For a study that maps where emotions are experienced in the body, see Nummenmaa et al. (2014).

[22] See Tappolet (2010, 327) for a similar list of components of fear.

[23] See Schwitzgebel (2008, 249-50) for further worries about the introspectability of the emotional phenomenology.

[24] Note that among intentionalists (see below) there has been some recognition of our relatively poor epistemic access to our emotions in light of their complexity; see Seager (2002).

To simplify the prose below, I'll use 'introspectable' in place of 'non-inferentially introspectable'. Thus, emotional experiences are non-instrospectable experiences — though of course they have introspectable components.

2.2.2 Intentionalism

Nearly all the evaluationists cited above see their views as a development of a major position within the philosophy of mind known as *intentionalism*,[25] which can be *very* roughly glossed as the view that *representation exhausts phenomenology*.[26] One of the major intuitions supporting the view, and which also gives some sense to the gloss, is the supposed transparency of experience: gazing at a blue square painted on the wall before you, the square very much seems to be *out there*, not some private mental object.[27] Try as hard as you might to concentrate on your *experience* of the square, you only end up concentrating harder on the square and its properties. The phenomenal character of experience is "inseparable from" what that experience represents.[28]

I take it that a defense of evaluationism should not threaten intentionalism. Partly this is because a defense of evaluationism that almost no extant evaluationist could accept would lack dialectical punch, but it is also because evaluationism itself borrows much support from the possibility that *all* phenomenal experience is representational, and with it the phenomenology of pain. For, putting this theoretical consideration aside, the unpleasantness of pains is often thought to be a prime candidate of a "raw" non-representational feel if ever there was one.[29] And a lack of transparency really would be a mortal threat to intentionalism, for then we could switch our attention from what our experience represents to some *other* feature of the experience, say to its "raw feel". In that case it would be hard to maintain that representation exhausts phenomenology. For that reason I will assume a moderate transparency thesis TP:

TP One cannot attend to, nor become directly aware of, one's own phenomenal experience.[30]

[25]Cutter & Tye (2011) and O'Sullivan & Schroer (2012) are explicit in their commitment to intentionalism. In outlining an earlier, injury-perceptualist view of pain Bain (2003) takes himself to be defending intentionalism. Helm (2002, 2009) is more carefully characterized as *intentionalist-friendly.* He aims to account for the *distinctive* phenomenology of emotions in terms of their intentional content, but it is unclear whether he thinks this accounts for its phenomenology without remainder and whether he is willing to extend a representational account to all phenomenal experience.

[26]A traditional formulation is that *phenomenal character supervenes on representational content* (see Byrne 2001), though at least one of the most prominent intentionalists, Tye (2014), now rejects this characterization in favor of what he calls "property representationalism".

[27]Tye (1995b, 30).

[28]From Horgan & Tienson (2002, 521). See also Harman (1990) and Dretske (1995).

[29]See Block (1996).

[30]Aydede & Fulkerson (2014) have recently argued that representationalism cannot explain in a manner consistent with transparency what, exactly, affective qualities like *awful* qualify. Are experiences awful, or are the objects they present us with awful? I think their challenge can be met, but it is also important to recognize that it is only aimed at representationalists committed to a stronger version of transparency than TP.

Importantly, TP does *not* imply that one cannot attend to the phenomenal *character* or what-it's-likeness of one's experience. According to Tye's intentionalism, for instance, the character of one's experience just *is* the cluster of properties it represents non-conceptually; the experience is the internal representational vehicle of that character.[31] Nor does TP deny us introspective access to our experiences, for it allows that we can be indirectly aware of them, via the access they give us to extra-mental events and properties.[32]

However, it is also important to recognize that many evaluationists take on additional commitments in elaborating their broader philosophical projects. Some aim to defend reductive naturalism about the normative property represented in unpleasant pain (O'Sullivan & Schroer 2012; Cutter & Tye 2011), others a tracking psychosematics (Cutter & Tye 2011). And all evaluationists so far have been *representationalists*, intentionalists who aim to explain phenomenal experience in terms of, or even reduce it to, a certain kind of representation. Not all intentionalists share these views.[33] Indeed I am more sympathetic to *phenomenal intentionalism*, which differs from representationalism by inverting the order of explanation: it explains the intentionality of experience in terms of its phenomenology.[34] For that reason the phrase "amounts to" in the statements of evaluationism and broad evaluationism should not be understood as offering a reduction or even as assigning a priority of explanation. But the solution I offer to the access problem is independent of any of these further commitments; I point them out only because, as I explain in the next section, different commitments will differentially affect the evaluationist's options for responding to the badness prong of STM.

2.3 The STM objection

2.3.1 Badness

We are ready to return to the shooting-the-messenger objection. In my brief response to Cutter & Tye (2014) above (p. 10) I noted the intuition that

4. Unpleasant pain is intrinsically bad for its subject.

[31] Tye (2014).

[32] I give a mechanism for introspection consistent with TP later in this chapter. Note that TP does entail that some of the phenomenological investigation in §2.2.1 above is misdescribed: one *attends to* the phenomenal character of one's experience but *introspects* the experience itself. Addressing this wrinkle above would have unnecessarily complicated the presentation, however.

[33] See Chalmers (2004) for an overview of intentionalist positions. Note that imperativism about pain (Klein 2007; Hall 2008; Martínez 2011) is often considered a form of intentionalism, and that 'representationalism' is often used not for a subtype of intentionalism but as a synonym for it.

[34] See especially Horgan & Tienson (2002) and Kriegel (2013).

Pace Cutter & Tye, 4 has recently received support from both sides of the dispute over evaluationism: the evaluationist Bain (2013) lists 4 as a constraint on any account of pain's unpleasantness, and critics Aydede & Fulkerson (2015, 16) agree, understanding pain's intrinsic badness in terms of its constituting a non-instrumental reason for experience-directed aversive behavior. The badness prong of the STM objection then amounts to the following argument:

5. If evaluationism is true, then an unpleasant pain is merely a representation of a distinct event as bad for its subject.

6. No state that is merely a representation of a distinct event as bad for its subject is *itself* intrinsically bad for its subject.

7. (From 4, 5, and 6) So, evaluationism is false.

The principal difficulty with this argument is introduced by the term "mere". Premise 6 is most plausible when interpreted as denying that the representation of badness is itself a bad-making feature, or that representations inherit the badness of what they represent. For the former, recall that our objection to Bain's response to STM (p. 10) was that something's merely *seeming* to be bad need not itself be a bad state. Korsgaard (1996, 155) seems to hold a view of the latter sort for pain, but it is hard to see how it can be made to work. Fire alarms represent a very destructive kind of badness, but a ringing fire alarm is bad only because it's *annoying*, not because it is a representation of something destructive — let alone because it *is* destructive. Something similar seems to be the case for pain: the principal badness of the injury and the badness of the unpleasant pain are of different sorts, for the former is a matter of physical harm and the latter is not. Unpleasant pain does not inherit the harmfulness of the injury.

On this interpretation of 6, a "mere *P*" is something that has no relevant properties other than *P*: if unpleasant pain is nothing but a representation, *of some sort or other*, of some other event as bad then it seems the pain is not intrinsically bad at all. The problem with this interpretation is that it renders premise 5 false. As we saw above, evaluationists hold that unpleasant pains, as conscious states, are representations *of a certain kind*, and evaluationists are free to appeal to any special features they take conscious representations to have in order to explain the badness of unpleasant pain.

What features might evaluationists appeal to? First we will need a rough sketch of a theory of the badness of unpleasant pain, and then consider how it can be understood in intentionalist-friendly terms. There is no space to adequately carry out that project here; my goal in this section is simply to point to features evaluationists *could* plausibly appeal to.

One highly intuitive proposal for the badness of unpleasant experience is that, at bottom, all unpleasant experiences impose on one's freedom. This is clearest in cases where the pain is severe enough to undermine one's agency and so one's freedom to act, but even in more mundane cases

it still discourages one from basic goods like the free use of one's body, or it disrupts the free flow of thought. And even in cases where it is not severe enough to interrupt or prevent one's pursuits, one's engagement in the pursuit will be less than free: if Ann dances, it will be *in spite of* the pain. Her dancing is compromised by being painful. So it is not merely that pain may *deprive* a person like Ann of something good, a carefree dance. It constitutes a way of engaging in otherwise good activities that is less than fully voluntary. That is itself a bad thing, so unpleasant experience is thus intrinsically bad.[35]

So much for a sketch of the intrinsic badness of pain. But how is it consistent with intentionalist scruples? Here I merely note that there are many options intentionalists might pursue. The way is perhaps easiest for the phenomenal intentionalist who reduces phenomenal consciousness to psychofunctional role, for everyone agrees that there is a very tight connection between unpleasantness and motivation. Affect primes motor systems for action, alters the weighting of goals, changes what information counts as relevant or significant for present action, provides steady input into conceptual evaluations of one's situation, and enables one to learn appropriate avoidance behavior.[36] A phenomenal intentionalist may plausibly hold that unpleasant pain reduces to or is realized by a state that in part plays that functional role. The resulting theory would be well-poised to explain why unpleasant pain constitutes an imposition on one's freedom.[37]

The way forward is less clear for representationalists, who typically explain phenomenology not in terms of the realization of an internal functional role but in terms of the tokening of a state that bears some external (teleofunctional or tracking) relation to extra-mental content or properties. But here too there are options. A tracking theorist might argue that at the level of cognitive architecture the state that causally covaries with, and thus represents, harmful injuries is a hybrid of sensory and motivational states. Or one might propose that a conscious *affective* representation of bodily harm must, in addition to being poised to affect cognitive, concept-applying systems, be poised to affect lower-level motivational and goal-setting systems as well.[38]

[35]The idea that unpleasant pain is bad because, or at least when, it interferes with agency is common one in the recent literature; see Swenson (2009); Klein (2015); Martínez (2015).

[36]See Panksepp (1998); Rolls (2014); Navratilova & Porreca (2014); Aydede & Fulkerson (2015) for recent contributions to theorizing on this issue.

[37]It is also important to recognize, *contra* Cohen & Fulkerson (2014), that evaluationism is not itself committed to denying causal accounts of unpleasant experience. Evaluationists do think that no *mere* causal account will suffice except one that *rationalizes* aversive action. But this only commits them to thinking that if a causal account of unpleasant experience is true, it will be of a very special sort — one that makes it a representation of a certain kind.

[38]This latter option does bear the cost of making representationalism less pure since it takes away some explanatory work from the content of a representation and gives it to the *way* that content is represented. (See Chalmers 2004 on the distinction between pure and impure representationalism.) But as nearly all representationalists are to some extent impure, the cost is a matter of degree.

2.3.2 Access

Now we turn to the access prong of STM. Recall from 1 (p. 9 above) that we want to account for the way in which pain-directed aversive behavior like Ann's is *practically immediate*. Yet because we are also looking to account for the normative relation between her unpleasant pain and her attempt to get rid of it, we need to explain how Ann is motivated in part by her access to the badness of her unpleasant pain. This requires that before acting (or at the latest, *in acting*) Ann *takes* her unpleasant pain to be bad, and that in turn requires that her pain be the *object* of one of her mental states. And surely, the state that gives Ann access to her unpleasant pain is a state of introspection.

This makes for a small difference with respect to Siddharth's action, where introspection of the unpleasant pain is not required since evaluationism holds that the unpleasant pain by itself provides the access he needs to rationalize his action. But intuitively, upon introspecting her unpleasant pain Ann is in a position to act out of recognition of its badness. Given that introspective access to one's pain is non-inferential, this means that Ann must have non-inferential access to its badness. And this seems right. Ann does not need the *further thought* that pain is bad in order for her painkiller-taking to be rationalized. All she needs is to recognize that she is having an unpleasant experience.

But the problem for the intentionalist at this point is that introspection *only* gives Ann access to the fact that she is having an experience with a *bad-injury-in-knee* character. There she is, attending to the (apparent) bad injury in her knee and rightly recognizing that really she has only an *experience as of* a bad injury in her knee. If, as we are supposing, her experience is transparent, then the badness of the pain itself has disappeared from view. Indeed, it seems the evaluationist must say that in order to access the badness of her pain Ann must do some further thinking — she must, say, remember that pains compromise her freedom and that this is a bad thing. This is inconsistent with 1, which requires her to have better access than that.

Now, it may seem as if there is an easy way out for the evaluationist at this point, one that Cutter & Tye (2014, 428-429) seem inclined to take. Why not suppose that Ann, upon introspecting her unpleasant pain, immediately forms a non-instrumental desire to get rid of it? Doesn't this give her all the access she needs to the badness of her pain?[39]

Here I'll argue that such a response falls prey to a dilemma, one that shows the true depth of the STM problem. The argument is inspired by one David Bain (2013, §4) levels against what he calls "mere inclination" views of pain's unpleasantness, and it turns on the nature of the desire that is thought to give the subject access to the badness of unpleasant pain.

Suppose first that the desire to get rid of the pain *cannot* be characterized as itself an evaluation and that instead it can only be characterized in terms of its motivational role, as a mere disposition or inclination to end one's pain. Citing Warren Quinn's (1993) infamous Radio Man, Bain argues

[39]Thanks to an anonymous referee for pressing this objection.

that such a desire does not itself give the subject access to a justification for ending her pain. Radio Man has a bare disposition to turn on radios in his vicinity — not because he likes to hear what's on, nor because he expects it'll be pleasant. From his perspective, he doesn't turn on radios for any reason at all — it's just something he does. But Radio Man's reaching to turn on a radio is rather too much like Yellowman's flailings, for to understand Radio Man correctly is to understand that, as it seems to him, his reaching for the radio very much stands in need of justification. And if that is so then the non-evaluative desire that causes his reaching cannot give him access to a justification for it — and neither can a similar desire explain the difference between Ann's taking a painkiller and Yellowman's avoiding yellow experiences.

So it must be that the desire to end the pain admits of an evaluative characterization: to desire that p is in part to see p's obtaining as good, and to desire that $\neg p$ is to see p's obtaining as bad. Evaluative views of desire are common enough in philosophy,[40] and are certainly consilient with the general project of evaluationism.

But here we run into another dilemma, this time turning on whether or not the evaluative nature of the desire is to be explained in terms of phenomenal representation. The first horn may look attractive in this case since it is plausible that the practically immediate desire Ann forms to get rid of her pain is an *experienced* desire — it is such that there is something that it is like to have it. Indeed, one might think it is the paradigm case of a motivation with distinctive phenomenology. Furthermore it is natural to think that the phenomenal feel of the desire is in part *unpleasant:* it feels bad to want to get rid of one's pain, one might think. A solution thus appears at hand: if, on broad evaluationist lights, bad feelings are representations of something as bad, and a desire to get rid of one's pain has a bad feel to it, doesn't that entail that the feel of the desire represents the pain as bad?

The problem for the evaluationist is that, if anything, she seems to be committed to answering *no* to this question at this point. For the hypothesized unpleasantness of Ann's desire to get rid of her pain is clearly part of her overall pain experience, and it appears that evaluationism is committed to holding that the unpleasantness of a pain experience transparently represents *the apparent injury*, not the pain, as bad. It seems that the evaluationist must say that although the desire itself is about the pain, the *unpleasantness* of that desire, as a component of Ann's overall pain experience, is not about the pain at all. Indeed it is unclear how it could in a way consistent with TP (p. 14): if Ann has a feeling that represents her pain as bad, then can't she attend to the way her pain *feels* to be bad as much as she can attend to the way her knee feels to be bad?

Now, there are other ways to understand Ann's desire to be rid of her pain as an evaluation of it, ways that do not appeal to its phenomenology; this is the second horn. One could construe desire as an evaluative attitude by holding that the good is the formal end of desire (Tenenbaum 2008) or

[40] See for instance Stampe (1987); Oddie (2005); Tenenbaum (2007).

is part of its Fregean force (Schafer 2013). But given the evaluationist's explanation of Siddharth's injury-directed behavior, appealing to a non-phenomenal evaluative seeming at this point leaves her with an inelegant and unmotivated theory. Contrary to desideratum 1 (p. 9), it gives fundamentally different explanations of Siddharth's and Ann's actions: each accesses a justification for acting, but while Siddharth's access is explained by the content of his phenomenal experience, Ann's is explained non-phenomenally by the nature of her desire. Moreover, presumably it's true that Siddharth *desires* to avoid his injury. On the view under consideration, that fact alone suffices for Siddharth's access to the badness of the injury. Appealing to a general thesis about the evaluative nature of desire that is independent of its phenomenology thus undercuts a principal motivation for evaluationism, that it is otherwise necessary to postulate that Siddharth's unpleasant pain has evaluative content in order to explain his access to a justification for acting. If desires are by nature evaluative, that postulation is superfluous.

Here is another way of looking at the problem. There is a philosophically popular sense of 'desire' according to which nearly any motivating state is a desire.[41] If desires come so cheaply it is no problem to say that Siddharth desires to be rid of his injury and Ann to be rid of her pain, and that these desires explain their actions. The distinctive contribution evaluationism makes is to explain why acting out of such desires counts as acting out of one's access to a justification for acting: it claims that *the phenomenology of pain* provides the evaluative access. Pre-theoretically, we might expect this to work for Ann just as well as it works for Siddharth. When Ann introspects her unpleasant pain, doesn't it feel to her just as awful as her knee does? And shouldn't that count as access to her pain *as bad?*

But the transparency thesis and evaluationism together seem to exclude this, for the only sense that they can make of unpleasant pain's "feeling bad" is its being an unpleasant feeling. And on this view, all we find when we attend to the phenomenal character of our unpleasant pains is an (apparent) event in our body and *its* normative properties. But when we introspect an experience as of some external object or event's having property *P*, we do not thereby come to have a seeming as of that very experience's being *P*. When I introspect my experience as of a red apple, I do not come to visualize my *experience* as red. Why, upon introspecting the unpleasant pain in her knee, should Ann come to feel her unpleasant pain as bad?

2.4 How to shoot the messenger

In this section I will show how, contrary to appearances, it *is* possible to feel your unpleasant pain to be bad, and in a way consistent with TP. The main pieces of the solution involve recognizing a distinction in *kinds* of unpleasantness and applying a theory of introspection to that distinction.

[41] For more on this, see Finlay (2007).

2.4.1 Secondary unpleasantness

The empirical literature on pain indicates that there is more to unpleasantness than the kind sketched in §2.2.1, which I will now call 'primary unpleasantness' or 'unpleasantness$_1$'. There is also what is called 'suffering' (Wade et al. 1996; Price & Barrell 2012), 'secondary unpleasantness' (Fields 1999), or 'secondary affect' (Price 2000). This aspect of pain experience, which I will follow Fields in calling 'secondary unpleasantness', is considered to be processed in series with pain and primary unpleasantness, and is also mediated by higher-level, cognitive processing related to the implications of the pain (Price 2000; Wade et al. 2011; Roy 2015). In contrast to primary unpleasantness, which may be processed in parallel with pain as well as in series (Price 2000) and is sometimes thought to be a kind of sensory discrimination (Fields 1999) that is independent of cognition (Gracely 1992), secondary unpleasantness is considered by all these authors to be distinctly emotional. Pain, as we might have suspected pre-theoretically, reliably causes some fear or anxiety, though the unpleasantness that occupies our attention in a given episode is primary unpleasantness.

The main disagreement over the two kinds of unpleasantness is how dependent on cognitive processing primary unpleasantness is. Price (2000) believes that both kinds of unpleasantness are mediated by cognitive processing, and that the distinction between them is largely in their intentional object: primary unpleasantness is directed at the stimulus and the immediate threat it poses, and secondary unpleasantness is directed at the pain's meaning and long-term implications. Fields (1999) believes that primary unpleasantness, unlike secondary, is *not* mediated by cognitive processing.[42] I think the phenomenology of primary unpleasantness described in §2.2.1 provides some evidence that Fields is right here, for it is characteristic of sensory experience to seem to have a spatial location, and sensory experience is not dependent on the kind of high-level cognitive processing Price has in mind. At any rate, I will assume that primary unpleasantness can be felt prior to *conceptualization*: like the pain itself, it does not require that you have a concept-laden thought about, say, injury or harm before you experience it.

Now, given that it is primary unpleasantness, not secondary, that occupies one's attention, one might wonder why we should consider secondary unpleasantness part of one's overall pain experience. Why is it not like being sad about a breakup and having a toothache at the same time? We normally wouldn't count the sadness as part of one's overall toothache experience.

But there is good reason to consider secondary unpleasantness as part of the overall unpleasantness of one's pain experience, for there is evidence that when experimental subjects are asked to rate the unpleasantness of their pain, they combine primary and secondary unpleasantness in their rating. Cancer patients tend to rate their most intense pains as less intense than do women

[42] For an excellent overview of the dispute, see Aydede & Güzeldere (2002).

during the most intense stages of childbirth, but they also rate it as more unpleasant (Price et al. 1987). Women in labor who focus mainly on pain and avoiding it find their pain just as intense as those who focus on the birth of the child, but rate it as considerably more unpleasant (*ibid.*). Induction of a sad mood while experiencing clinical pain is associated with an increase in catastrophizing thoughts about the pain and an increase in the unpleasantness attributed to it (Berna et al. 2010). In all of these cases, what most explains differences in unpleasantness ratings seems to be concept-laden emotional factors.

This data also suggests a straightforward explanation of why the secondary unpleasantness of pain is genuinely part *of* the pain experience while one's sadness at a breakup is not part of one's simultaneous toothache experience: the former and not the latter bear the right kind of intentional relationship. The sadness is experienced as about the *breakup*, not about the tooth or the toothache. One's dread in experiencing cancer pain, however, is very much about either the pain or the damage it represents (or both). Something similar goes for the study by Berna et al. (2010), who induced a sad mood in subjects by playing them sad music at half speed. The subjects' feelings *about their pain* were manipulated just as a film score manipulates our emotional experience of the characters.

So, say that an emotion is *intentionally attached* to a sensory experience if it is experienced as about either the experience itself or what that experience represents. This evidence should lead us to conclude that secondary unpleasantness is intentionally attached to unpleasant$_1$ pain. Although it does not allow us to say that secondary unpleasantness is about the pain *as opposed to* the injury it represents (say), it does entitle us to a disjunction: the secondary unpleasantness of pain is about the pain or what it represents.[43]

2.4.2 Transparent introspection

But recall that the intentionalist has additional commitments here. On intentionalist lights, both primary and secondary unpleasantness are in Siddharth's case about the *injury*; they *both* represent it as bad, though there may be subtle differences in the kind of badness they represent. (Perhaps secondary unpleasantness takes into account broader implications for the subject's well-being.) And as we saw in §2.3.2, this leads to the STM problem. The solution to the problem, then, is that when the subject introspects her unpleasant$_1$ pain the intentional object of her secondary unpleasantness shifts to *the pain itself*.

To explain this the evaluationist needs two things. First she needs to explain how to introspect her experiences in a manner consistent with TP, which is what I will take up in this section. Second,

[43] Of course, this is not to say anything about what exactly attached the two experiences intentionally, nor how tight the bind is. As in the case of misattribution of arousal (Dutton & Aron 1974), what mental states one's emotion attaches to will often depend on context and what information is salient. The same appears to go for moods; see Schwarz & Clore (1983) for a classic study.

she also needs to explain when the intentional object of an emotion can shift from something in the external world to one of the subject's own mental states. I address this in the following section.

As noted in §2.2.2, introspection to our phenomenal experience is consistent with TP so long as it is *indirect*. In practice this amounts to a kind of *displaced perception* along the lines of Dretske (1995, Ch. 2): one comes to have knowledge of one's experience via an awareness of the properties of physical objects those experiences present one with. An ordinary example of displaced perception involves seeing that the gas tank is empty by reading the gas gauge. The gauge wouldn't read 'E' unless the tank were empty, the thought goes, so it must be empty. Dretske's proposal is that introspecting my experience as of a red apple works in a similar way. I first form the belief that there is a red apple in front of me, on the basis of my visual phenomenology. Then, via an appropriate "connecting belief" (*ibid.*, 58), I derive the belief that I am having an experience as of a red apple.

There are a number of problems with this proposal. For one thing, the tank analogy suggests that the connecting belief in the introspective case is "There wouldn't be a red apple in front of me unless I were seeing it", which I in general have *no reason* to believe. This fails to account for the epistemic quality of introspection.[44] For another, on this view introspection is, at bottom, an inferential process: I *derive* the introspective belief from a visually-based belief and a connecting belief. And intuitively, this cannot be how introspection works. In introspecting I simply look at the apple and form the belief that I am having an experience as of a red apple.

Fortunately, the intentionalist has a way around both of these problems. All she needs is to hold that there are two kinds of beliefs that an agent can form from an experience that represents (say) propositional content p: the external-world belief that p^* corresponding to the content p of her experience,[45] and the introspective belief "I am having an experience as of p", formed by attending to the content of her experience and using it to refer to her own experience. More precisely, the view is that an agent first attends to the phenomenal character p of her experience and applies her concept of an experience to it: "EXPERIENCE AS OF (p)"; this is the introspective step. She then comes to believe that she is having just such an experience.

[44] See Aydede (2003); Lycan (2003) for similar criticism. Dretske does anticipate this worry and points out that the inference from a visually-based belief that an external world object has property p to a belief that I am having an experience as of an object's having p is infallible (*op. cit.*, 61). As he himself notes, this is a "very unusual form of inference" that secures a true belief whether or not the premises are true: it goes through even in the hallucinatory case where there is no object seen (*ibid.*). But this raises further issues. If we think of an introspector as not paying attention to whether her beliefs are visually-based when she makes these inferences — for that would presuppose introspection already — then it seems she should be strongly tempted to conclude from this unusual feature that she is having an experience of everything. For either there is an apple there or not, and she has just been informed that her usual inference from the apple's existence to her having an experience of an apple goes through even when there is no apple. This is absurd.

[45] The contents may not be the same if, as many intentionalists think, the phenomenal experience has non-conceptual content.

This account gives us the right kind of access to our experiences. As a form of displaced perception it is consistent with TP: it does not require attending to or directly introspecting an experience, as if one could just "see" one's own experience. Since it holds that introspection is a matter of forming a belief from a phenomenal experience, it is non-inferential, and for that reason it also does not require reliance on an unjustified belief. And it is formed via a reliable process: basing one's belief that one is having an experience as of p on one's experience as of p will *always* generate a true belief. Furthermore, it gives us access to our experiences *as* the phenomenal experiences they are. As noted in the case of pain, there is nothing more to being a phenomenal experience, *as* a phenomenal experience, than to be an experience with a certain phenomenal character. According to this proposal, introspection uses that very character, the content p, in specifying precisely what experience one is having, thus giving access to its nature.

2.4.3 Turning unpleasantness inward

We are at last in a position to solve the STM problem and thereby answer Q1. The explanation of Siddharth's action is largely unchanged: as before, Siddharth attends to his injured knee via his unpleasant pain, which in turn represents the injury as bad for him. He thus acts upon accessing a justification for acting. The only difference is that now his unpleasantness is understood to be a composite of primary and secondary unpleasantness, both of which represent his injury as bad.

Now as we noted in §2.2.1, unpleasant$_1$ pain is an introspectable phenomenal experience. The *secondary* unpleasantness of one's experience, as an emotional state, is not. So what happens when Ann introspects her unpleasant$_1$ pain?

Well, consider first a rather different case. When the Marquess of Glastonbury returns to her estate to find her armoire rifled by thieves she is sad, but she becomes positively distraught when she sees what the burglars did to the portrait of her mother: they ripped it, thereby giving it a gruesome appearance. Her emotional experience of the painting is very much intentionally attached to her visual experience of it, as it is the way her dear mother seems so savagely disrespected in the painting (the object of her visual experience) that is so distressing to her. But then the Marquess might briefly turn her gaze inward and introspect her visual experience as of the painting. Surely she now is just as distressed that she is *seeing* such a horrible thing. And so she turns away.

Here is a plausible principle that explains the Marquess' phenomenology:

IT When an unpleasant emotional experience is intentionally attached to an introspectable experience e with content p , then (i) when e is *unintrospected*, the emotion represents p's obtaining as bad. And (ii) when e is *introspected*, the emotion represents e's obtaining as bad.

When the Marquess attends visually to the ravaged painting, she represents the horribly desecrated state of the painting. Yet when she introspects her experience of it, she represents her *seeing*

the painting to be bad. That not only explains her turning away but rationalizes it, for in turning away she accesses a justification for doing so.[46]

And now we can easily apply IT to Ann's action: as I argued in §2.3.2, Ann introspects her unpleasant$_1$ pain before acting. I also argued in §2.4.1 that the secondary unpleasantness of one's overall pain experience is intentionally attached to the unpleasant$_1$ pain. Given this, IT entails that the (emotional) secondary unpleasantness of her overall pain experience then represents her unpleasant$_1$ pain as a bad thing. But this was precisely what was needed to give her access to a justification for experience-directed aversive behavior, so the access prong of the STM objection is solved. Ann shoots the messenger for a reason.

2.4.4 Objections

It might seem that by appealing to a higher-order affective state in order to explain Ann's action we have only given ourselves another messenger to shoot. Suppose that Ann has a cousin Zorba who occasionally feels an odd, pressure-like sensation in his knee. Though the feeling has no *primary* unpleasantness, Zorba cannot help but feel anxious about it when it comes, and for that reason he keeps anti-anxiety pills in his pocket. (He used to have pills to dampen the odd sensation, but they've gone missing.) Surely Zorba's anxiety is bad for him in much the same way unpleasant pain is, and taking a pill would relieve him of it. Don't we also need to explain his access to this justification for experience-directed aversive behavior?

It's true that Zorba does have access to a genuine justification for taking his pill, but the crucial fact is that it is not the kind of access that generated the STM problem. Recall from desiderata 1 and 2 (p. 9) that Siddharth's and Ann's access to a justification for acting was practically immediate upon experiencing the unpleasant$_1$ pain: nothing more than introspection on their experience was required in order for them to have that access. And that, in turn, seemed to cause a conflict with TP, for introspection on an experience presenting a property *P* not give us awareness of our experience as *itself* having property *P*. Our account resolves this difficulty by relying upon the presence of an additional state, secondary unpleasantness, and on a principle that explains how the intentional object of secondary unpleasantness shifts with introspection.

But part of the overall account developed here is that secondary unpleasantness, as an emotional experience, is *not* introspectable (§§2.2.1, 2.4.1). Our awareness of our emotional experience is not practically immediate upon experiencing it, and instead must be obtained inferentially. Indeed, this gives us a plausible account of how Zorba must access his own anxiety: his focus narrows, he feels a tightening in his stomach, he can hardly keep his mind off his odd sensation, and from all this he infers that he is feeling anxious again. And crucially, it seems that *thinking* about

[46] It is worth pointing out that the Marquess is *accurately* representing her experience as bad, since it causes distress and it is generally bad to be distressed.

one's emotional state, as opposed to introspecting it, *can* make that state self-representing: when I anxiously consider the fact that I'm anxious, one of the things that I am now anxious about is my own anxiety.[47] According to broad evaluationism, such an affect-laden thought represents one's own anxiety as a bad thing. And this, in turn, provides a plausible explanation of Zorba's access to the badness of his own anxiety.

In short, our access to our emotional states is too poor for them to cause the special problem at the root of STM, so this account does not incur a similar explanatory burden in appealing to one in order to explain how our unpleasant$_1$ pain seems to us to be bad.

Now, some might find the present account an objectionably high-brow theory of experience-directed aversive behavior, for it requires not only metacognitive capacities but a concept of phenomenal experience. Don't infants attempt to avoid pains too, and for the same reasons we do?[48]

Importantly, this is not a *new* objection to intentionalism. Most intentionalists require conceptual capacities for introspection, so the present account does not worsen the intentionalist's position. But supposing that infants *do* lack a concept of experience, it becomes hard to understand how, in apparently distracting themselves from pain, toddlers are really engaging in *experience-directed* aversive behavior. Without the conceptual capacities necessary to differentiate experiences from the content of those experiences, an infant cannot distinguish *being in a painful state* from *being hurt*. So if an infant with a pain disorder learns to distract herself when the pain comes, he is not doing it for the same reason an adult does, for he really thinks that alleviates the hurt or injury.

Before closing, it is worth noting an implication of the account. If it is possible for a person to experience unpleasant$_1$ pain without *any* secondary unpleasantness, the present account predicts that although such a pain is bad, that subject also has no *non-inferential* access to that badness, and so it should not make immediate sense to her to take a painkiller even upon introspecting the pain. This sounds odd at first, but it becomes intuitive once we understand how different such a person must be from us: while having the pain, she must be *totally unconcerned* about the pain itself. Her experience is more analogous to to that of a person with congenital insensitivity to pain who feels the burning stove to be warm to touch but not painful at all. Such a person is aware of an event in his hand that is *in fact* bad for him, but introspection on his experience will not afford him access to that badness. Instead he must infer that it is bad from what he has learned about hot stoves.

[47] Note that Colin Klein (2015, 55-56) has recently used this same fact to a similar purpose in his imperativist theory of pain.

[48] Thanks to Peter Railton for pressing me to address this worry.

2.5 Conclusion

Intentionalism provides an attractive picture of conscious experience as a fully transparent window to the world. Evaluationism elaborates that picture by adding that the world as given in experience is far from evaluatively neutral: it is full of danger and the opportunity for gain. An important advantage of evaluationism, then, is that it points to a general account not only of *unpleasant* experience but of all affective phenomenology — one that assigns it a foundational role in agency.

The heart of the STM objection to evaluationism is that intentionalism appears to make experience *too* transparent, too world-directed, to account for the practically immediate access we have to the badness of our own unpleasant pain. The main goal of this chapter was to show that objection can be met, and indeed that the evaluationist can provide an attractive account of the way in which the emotional component of unpleasant pain gives us that access.

The solution deploys a number of rather fine distinctions that have emerged in either the philosophical or clinical literature on pain: between pain sensations and unpleasantness; primary and secondary unpleasantness; experience-directed and body-directed aversive behavior; and between the experience of unpleasant pain and its intentional content. But it is important to keep in mind that it is no part of the theory that we as agents always keep careful track of all the distinctions there are to be made. Many aversive responses to unpleasant pain may have unclear aims, for instance, and it may be indeterminate whether an agent is introspecting her experience. So too the account respects the intuitive sense in which in both Siddharth's and Ann's case there is just *one thing* that occupies their phenomenal experience: it is the (apparent) *bad injury* in the knee which monopolizes the attention of both. The account just tells us what there is to be clear about, if we are being clear.[49]

[49] I owe this turn of phrase to Peter Railton.

Versions of this paper were presented in Budapest and Ann Arbor, and it benefited greatly from the comments and questions of audiences in both. I'd like to especially thank Peter Railton, Sarah Buss, Allan Gibbard, James Joyce, Rohan Sud, Daniel Drucker, and the members of the 2015 Michigan Philosophy Dissertation Working Group for all their comments and suggestions. I'd also like to thank an uncommonly helpful anonymous referee by whose criticisms and suggestions the paper was vastly improved.

CHAPTER 3

Intelligibility and the Good of the Guise of the Good

> In arguing for and applying the Guise of the Good Thesis philosophers rely on a concept with broader applications than those associated with the normal use of [the words 'value' and 'good']. There is no point in trying to describe this concept here. It is familiar from the writings on the subject, and on value theory generally. And of course, one familiar aspect of it is the absence of agreement about its nature.
>
> Joseph Raz, "On the Guise of the Good"

> It would be absurd if we did not understand both angels and devils, since we invented them.
>
> John Steinbeck, *East of Eden*

3.1 Introduction

"The Guise of the Good" (GG) refers to a family of views which all hold that human action or motivation to act, of some special kind or another, is only possible insofar as the agent performs or is motivated to perform the act because of the *good* she sees in so acting.[1] On this view, (apparent) goodness is held to have a primary motivating and rationalizing role in action. There are a number of dimensions along which members of the family differ:

- Is the special kind of motivation or action to which the thesis applies desire, intentional action, action for a reason, "full-blooded" action, or just action itself ?[2]
- Is the notion of good in play a threshold notion ("good enough") or a graded one? Does it provide a merely *pro tanto* justification for action or a sufficient one?
- What is the guise under which agents must act? Must agents *believe* that their action is good (Raz 1999; Gregory 2016)? Must it merely *appear* good to them? Or must the agent believe the action to have a property that is in fact a good-making feature? Does the good figure

[1] See Tenenbaum (2013) and Orsi (2015) for recent overviews. Note that although "sees that *P*" often has a factive sense in English, here it is not to be presumed that whenever an agent sees an action as good, it is good.

[2] For the last, see Brewer (2009, 43).

in the content of an attitude required for action/motivation, or in the nature of the attitude itself? For instance, should we understand agents as having an attitude like Believe(φing is good), or more like IsGood(φing)?

For the purposes of this chapter, I will interpret the thesis so as to minimize controversy. GG is at minimum the following thesis:

Minimum GG An agent acts for a reason only insofar as that action appears to her to be to some extent good (in a generic sense of "appears" that doesn't distinguish among the many ways in which something can be said to appear to us).[3]

The conclusions I wish to draw in this chapter extend easily to mere attempts to act for a reason and to being motivated to act for a reason, but for the sake of clarity and simplicity I will limit discussion as far as possible to acting for a reason.[4] It should be noted, however, that mere adherence to Minimum GG would not suffice for an acceptable Guise of the Good theory. A full GG theory is acceptable only if it can describe the connections among motivation, the rationalization of action, and the good that explain *why* the conditional in Minimum GG is true.

I mention these issues and qualifications only to set them aside as much as I can — though I will, as it happens, have something to say about the third bullet point both in this chapter and the next. My interest here is primarily in the *nature* of the good at the center of GG and in its connection to the intelligibility of action; I will argue that a popular motivation for GG tends to support a rather objective notion of goodness. This is an excellent development in the debate over GG since, as I argue below it addresses what I call the *inertness* worry for GG theories.

Surprisingly, very little work has been done to clarify the good at the center of GG, and many cross-cutting distinctions remain in play:

- Is there an "objective list" of familiar, particular values — fidelity, benevolence, beauty, and maybe the dainty too — under the guise of one of which agents must act?
- Is the good of GG a specifically moral kind of goodness, as many of its opponents have characterized the view, or a goodness *for* the agent in question (Saemi 2015), or perhaps the "human good" (Brewer 2009)?
- How specific does the kind of good need to be? Perhaps we can act under the guise of a generic notion of good that does not implicate any particular kind or theory of goodness (Clark 2001).
- Is the good best characterized as the formal end of practical reasoning — that is, as the end that any practical reasoner must of necessity possess (Tenenbaum 2007)?

[3]For more on this notion of appearance, see Tenenbaum (2007, 38-39) and below.

[4]Later in the chapter I will, however, presume that if GG so construed is plausible, it will be plausible for some cases of intentional action that are only controversially cases of action for a reason.

In this chapter I am especially concerned with one question that is related to the foregoing but distinct from them: Is the good of GG necessarily *shared* or (equivalently) *intersubjective*, having a normative force that extends across differently-situated agents, or might it be a purely *private* good?[5] Suppose that not long after my dear grandmother's death I decide to take up crocheting the doily she left unfinished. Making the doily won't improve my life, or least that isn't why I'm doing it. I'm doing it for grandma's sake. I reason that it would have been just as good if any of her descendants had completed this project for her, but the task fell to me. Intuitively it seems that I am motivated by my recognition of the value of helping a close family member complete a project, and it also seems that my motivation is a form of genuine appreciation: making the doily really is valuable in the way I think it is.

But what normative implications, if any, does the fact that making the doily is valuable have for *other* agents? A particularly interesting implication it might have is that it might directly — that is, in a way explained just by that value and no others — give even strangers a reason of a certain kind. In order of increasing strength, it might give strangers a reason to not to disapprove of my doily-making, to approve of it, not to criticize it to my face, not to interfere with it, or to aid me in it. If strangers have any of these reasons, then this value is, in a sense, shared among all of us: we are united by the fact that each of us has reason to tolerate my doily-making. That same value commands respect not only from those in a certain community or group — in this case, the members of a family — but from those who have no such special relation. For that reason this value cannot be fully expressed by a principle such as "One ought to help *one's own* family members with their projects." If, on the other hand, nothing of the sort follows then the value is parochial, having normative force only among the members of a given group. A private good can then be seen as a limiting case of a parochial good, having no normative implications for those not in circumstances that replicate that of a given agent. Universal ethical egoism, for example, is often thought to entail that the good is quite generally private. On this view the fact that your φing would contribute to my well-being may give *me* reason to get you to φ if I can, but that same fact need have no practical normative significance for you or anyone else.[6] A public good, by contrast, is a good that is shared with everyone.

In this chapter I will argue that one prominent motivation for GG, the *intelligibility constraint*, tends to support a fully public notion of the good, which is a stronger notion than many are prepared to accept. For those theorists who hold this interpretation of GG already or are in other respects

[5] As I use the phrase, an intersubjective or public goodness is any kind of goodness that is essentially shared among agents. Talk of public goods here follows Korsgaard's (1996) notion of a public justification, not the economic notion. Also, I follow Wallace (2009) in distinguishing publicity from agent-neutrality. An agent-neutral reason is a public reason that is a reason for *anyone* to promote some end.

[6] Of course there is a sense in which an egoist can judge that someone getting the better of her may be acting rationally, so long as that judgment is understood not to commit her to a normative judgment about what she herself ought to do in response.

friendly to it — notably Joseph Raz (1999) and arguably Sergio Tenenbaum (2007) — this paper is the first in the literature to offer an explanation, in quite general terms, of why this notion of the good is a natural fit with their views. However, I stress that that argument I will outline for this view depends crucially on a number of premises I will not be able to defend here, although I regard them as plausibly true. My aim in this chapter is merely to lay out the *route* to that conclusion. In the concluding chapter of the dissertation I will briefly return to the issue consider how the GG view developed in Chapter 4 might help justify some of these premises.

Section §3.2 explains in more detail just why there is a problem for GG theorists concerning how the good is to be specified, and it ends by way of suggesting that a natural solution is for GG to rely upon a notion of shared goodness. §3.3 sets out the intelligibility motivation for GG and proposes an account of the relevant kind of intelligibility in terms of appearances of the good. §3.4 shows how this account can easily lead to a public conception of the good and draws out a few of the conception's implications for the theory of practical reason. §3.5 concludes.

A note about terminology before moving on. This chapter puts into question the sense of good that GG uses, and to that extent I will use "good" and "value" to express generic normative concepts. This does not presuppose that GG is a thesis about *evaluative* presentations. Genuinely evaluative concepts — the sense of "good" and "value" on which they are distinct from reasons and requirements — I will express with starred terms, i.e. "good*", "value*".

3.2 The delimitation problem

This chapter attempts to establish a constraint on the notion of goodness that GG theorists can appeal to, at least for those who ascribe to one popular motivation for the view. But the constraint is of interest to GG theorists no matter their persuasion, for it constitutes one answer to a problem that all GG views face, which I call the *delimitation problem:* Is there a principled way of determining which normative notions could be substituted for *good* in Minimum GG and still leave the resulting thesis a genuine version of the Guise of the Good, and if so what is it?

3.2.1 The triviality worry

It might at first seem that the delimitation problem need not be addressed, or that it admits of an easy answer. And surely the first response that comes to mind is that the good of GG is *the good**. One might hold that as a matter of definition the Guise of the Good just is the view that we act under the guise of some value* or other. But this response is unsatisfactory. Not only are there vast disagreements over the kind of value* GG theorists take our actions to be guided by, as we saw above, but some GG theorists hold that we may be guided by so-called "deontological" goods

(Tenenbaum 2007, Ch. 5)[7] and others argue we should instead be "Guise of Reasons" theorists (Gregory 2013) or "Guise of Ought" theorists (Massin 2016). From the perspective of action theory all these disputes seem best characterized as disputes within a single family. To make this clear, consider two extreme views:

8. An agent acts for a reason only insofar as that action appears to her to *make some contribution to aggregate well-being.*

9. An agent acts for a reason only insofar as that action appears to her to *satisfy the Categorical Imperative* (where, for the sake of the example, we take the Categorical Imperative to possess some modest moral content, for instance that free-riding on others' contributions is forbidden).

According to the proposed response to the delimitation problem, 8 is a GG view while 9 is not, for the former substitutes the evaluative* notion *well-being* in the place of *good*, while the Categorical Imperative expresses a paradigmatic deontological notion. These views would indeed make different contributions to the shape of our ethical theories, but as far as action theory is concerned they share their most important features: both claim that we act under the guise of some particular, moderately ethical standard. We should, at the outset at least, treat them as species of the broader genus of Guise of the Normative views.

This in turn suggests another attempt at deflating the problem: we might count a view a version of the Guise of the Good just in case it substitutes *some* normative notion, or some class of normative notions, in for *good* in Minimum GG. GG would then be vindicated if some such substitution resulted in a true statement. This permissive orientation is encouraged by the frequently-held position that the good of GG is the formal aim of practical reasoning, which would in principle liberate the notion of good it uses from the evaluative* ones often associated with GG. This would give us:

GN An agent acts for a reason only insofar as that action appears to her under some normative guise or other.

Here, an action appears under some normative guise or other just in case it appears to the agent to possess some normative property: it appears to be good*, to be what one ought to do, to be fitting to do, etc. Of course there will be marginal cases and disagreements over which notions are properly normative, as in the case of goodness of a kind (a good toaster, a good proof) or social convention (e.g. the notion of what one is supposed to do according to prevailing social conventions). But the proper classification of these cases will not affect our purposes here.

[7] According to Tenenbaum, a deontological good is one a constitutive condition of whose value is the conformity to a rule. He offers politeness as an example: necessarily, one acts politely *by* conforming to certain rules of etiquette. Tenenbaum argues that these goods enable deontological constraints to be compatible with the injunction that an agent always should promote the most value (*ibid*., 199-200).

But even if the good were the formal aim of practical reasoning, that would not imply that GN is an adequate specification of GG. It may turn out that, due to its very nature, practical reasoning cannot occur under the guise of just any normative notion, so that a more perspicuous definition could be made. But more importantly there is a worry that the GG theorist needs to say something more to characterize the notion of the good that figures in her theory, lest it float too closely to a notion that figures uncontroversially in the concept of action for a reason.[8] If it does then we risk making GN true but at the cost of near-triviality or uninformativeness. Call this the *triviality* worry.

I'll consider two related ways of pressing the triviality worry. The first derives from Tamar Schapiro (2014, 143), who briefly offers an argument in favor of a view like GN but regards the view as trivial. Ultimately, however, I find that Schapiro's argument is most useful for setting up the second way of pressing the worry, which leans on a proposal about the content of experiences of affordances.

Schapiro puts her argument as a combination of claims about the nature of action and of action explanation. She holds that human action must be "an agent's purposive response to his representation of the circumstances", and that, if that's so, then any explanation of action must "appeal to the agent's way of looking at the world in practically salient terms", which would itself be an evaluative outlook (*ibid.*). Unfortunately the second claim does not follow from the first, since a purposive response to a representation need not itself be anything like a representation. A very simple thermostat may represent the threshold temperatures at which it turns on the heat, but it need not represent the range of temperatures it "aims" to achieve. One could thus argue that the purposiveness of human action is a similar, albeit more complex, kind of regulated response: human agency on this view is a system in which a complex causal tendency to produce certain outcomes in certain situations plays itself out, even though the system itself contains only input-output dispositions and no representation of the outcomes it tends to produce. It does no good to hold that it is just *as if* the system had a practical outlook or that it is interpretable as having one; what matters for GG is not whether humans behave *as if* they act under the guise of the good, but whether they really do.

The point that I am trying to make does not presuppose that GG must be interpreted as requiring a discrete representation of the good, for all I mean to do is point out that Schapiro needs to earn the talk of purposiveness' requiring an evaluative outlook. So too she would need to defend the claim that purposiveness is an essential feature of human action; Kieran Setiya (2007, esp. 51-52), for one, thinks that teleological explanation is a special case of action explanation. It is hard to understand how actions without a teleological explanation could be purposive.

For this reason I don't think that the triviality worry should be pressed as a claim about the uncontentiousness of a thesis such as GN, but instead as one about its weakness or uninformative-

[8]The complaint is not new; see Railton (1997, 66 n. 14); Burgh (1931, 72).

ness. Note that the move from the alleged purposiveness of action to action's involving an agent's evaluative outlook does not in context look to be any more difficult than a move from the purposiveness of action to action's being motivated by a state (a desire, if you prefer) which, through its propositional content, represents the state of affairs which is the action's purpose. That is because Schapiro uses "evaluative outlook" such that looking at the world in practically salient terms *entails* having an evaluative outlook; the outlook in question may be simply the way in which desires represent or look out at their contents. Having an evaluative outlook thus turns out to be less interesting than we thought. Yet surely the Guise of the Good cannot turn out to be true simply because in acting, practical possibilities appear to us.

It might seem that the problem here is that it is simply not true that an appearance as of there being salient practical possibilities constitutes an evaluative outlook. After all, as just mentioned, we can specify that appearance in non-normative terms as a motivational state with a non-normative propositional content. One might think that when we can describe an appearance in this wholly non-normative way it cannot *really* be an evaluative outlook. But regardless of whether this is true, there is an independent case to be made that many practical possibilities appear to us under a very weak normative guise.

Susanna Siegel (2014) has recently done some excellent work exploring and defending the thesis that affordances may be represented in experience.[9] Affordances are possibilities for action for a creature, especially as given by objects. Handles can be grasped, chairs sat on, and trees climbed. In a colloquial sense of "see" one can see affordances for oneself as for others: you can see that you can hit the baseball with your bat, or that your sister can climb the tree in front of you, or that parking spaces are the sort of thing that people generally can park in. Siegel is interested in the question of whether affordances for oneself are represented as such in perceptual experience itself, and if so what their content is. For our purposes the interesting possibility to consider is not the possibility that affordances *may* be represented in perception but the stronger hypothesis that they may *always* be represented in action, i.e. that the content Siegel associates with affordance-experiences must always be represented in intentional action.

Supposing for a moment that they are always so represented, what might we say about the content of these experiences of affordances? Siegel (54-55) makes two distinctions concerning these experiences: they may be non-soliciting or soliciting, and if they are soliciting they may also be experienced mandates. A non-soliciting experience of an affordance is an experience of a possibility for acting, and nothing more. One might see the tree as climbable merely in the sense that one recognizes one could climb it. But the experience is soliciting if one feels or takes oneself to be to some extent invited to climb it. The content of this experience, according to Siegel, is that the tree is to-be-climbed (by one) (67-68) — not in the sense that it will be climbed, but in the

[9] All page references to Siegel in this section are to Siegel (2014).

normative sense that it is suitable for one to climb, or that it is fitting or appropriate to climb it. We could, it seems, easily generalize this notion to experiences of possibilities for action that are not given by aspects of the environment: in opening my laptop I see my paper as to-be-written, in light of a deadline I see as to-be-respected.

A soliciting experience of an affordance might divert one's attention to the possibility of φing, and it may involve an experience as of knowing just how one would φ, but it need not include any inclination to φ. One can see a river as to-be-swum-in without being the least inclined to jump in. By contrast, an experienced mandate that something is to-be-φed does involve some felt motivation to φ; the content of this experience might be that one has some reason to φ, that one ought to φ, or that one must φ, as in the case in which one is playing basketball and one experiences an incoming pass as to-be-accepted.[10]

Now, it seems to me that *if* I have correctly identified the content of experienced mandates, the thesis that we experience a mandate in acting for a reason would indeed count as a Guise of the Good thesis, for reasons that will become clearer in the next subsection.[11] But note that the content adduced for non-mandating soliciting affordances seems to be an excellent way of cashing out the notion of *practical salience* in normative terms. Thus, consider for the moment the hypothesis that in φing for a reason, we merely must represent φing as to-be-φed — as suitable, fitting, appropriate, etc. for one to φ — precisely because we must have an experience of φing as solicited. Call this the "weak normative light" (WNL) thesis. Would that count as a version of the Guise of the Good?

Intuitively, the answer is no. The truth of the hypothesis would merely mark a relatively unimportant fact about action: that we represent the goal of our action — which may simply be the performance of the action itself — in normative terms. It would attest that φing involves having φing as a goal, and that having φing as a goal involves seeing φ as to-be-φed. It goes without saying that this is *not* the sort of thesis that Guise of the Good theorists ought to be or would be

[10]Here I depart from Siegel in two ways. First, Siegel characterizes experienced mandates as involving a high degree of felt solicitation, such that they always involve a considerable degree of motivation (55). But given that, as she characterizes them, solicited experiences may involve *no* motivation, I find it cleaner to conceptually separate *felt solicitation* from *motivation*. Second, Siegel's proposal for the content of an object X mandating φing is actually:

It is answered that: X is to-be- φed.

She hopes thereby to include the motivation of the experience in its content (70). But I fail to see how this works; as I understand the "answering" locution, Siegel's proposal merely represents that one is motivated by X to φ, which is not the same as capturing the motivational force of the experience. My proposal is to capture the motivational element of experienced mandates by switching to a different register of normative notions: non-motivating experiences of solicited affordances represent that φing is suitable, fitting, etc., while motivating experiences represent that one has reason to φ, ought to φ, etc.

[11]In brief the reason is that, according to this proposal, what experienced mandates represent are standards of practical reason. I also regard it as an interesting possibility that the GG theory of Chapter 4 may turn out to be a version of such a theory.

satisfied with.

Moreover, regardless of its truth, this "weak normative light" thesis is not one that opponents of GG have been concerned to deny. Velleman (1992a, 117), in arguing against GG, nevertheless allows that desires may present their objects as to-be-brought-about, so long as they are not understood as necessarily having the aim of getting it *right* as to whether the object really is to-be-brought about, as a judgment might — where Velleman seems to take a judgment on whether the object really is to-be-brought-about to be equivalent to an approval of that thing.[12] The thesis may appear not to comport well with certain construals of the Humean theory of motivation, such as the view that a desire to φ and the belief that one could now φ by ψing are sufficient to make one's ψing intentional (provided the desire and belief cause the ψing in the usual way).[13] It is true that no Humean is committed to holding that an action must appear to its agent in a weak normative light, but then again few are actively antithetical to it, and indeed Smith (1987, 38-39) and Schroeder (2007a, 157 n. 11) are sympathetic.[14] It is open to Humeans to hold that the appearance of the action as to-be-brought-about supervenes on or is analyzable in terms of the belief-desire pair that causes it, much as one might hold that what it is for someone to appreciate a snarling dog *as dangerous* is to believe that it is liable to cause harm and to desire to get away from it.[15]

Again, my goal here is not to show that it is undeniable that actions appear under a weak normative light. It is rather to point out that hardly anyone thinks that GG is false *because* they deny that in acting intentionally we feel drawn or guided to do that action in a way that could also be described as representing it as to-be-brought-about. This had better not be all GG theorists are fighting to gain. (I do think, however, that properly understood WNL is not worth disagreeing with either, since it merely registers that action involves a phenomenology of future-directed or ongoing guidance, and it so happens that some locutions in the indicative mood employing normative terms can capture this phenomenology.)

Faced with the triviality worry, in order to characterize a version of GG worth fighting for we should recall a general motivation for GG. GG is often thought to account well for a certain analogy between theoretical and practical reason. Beliefs, it is often held, have intrinsic correctness conditions: a belief is correct if and only if it is true. Furthermore, in a certain sense beliefs represent their contents *as* true, for there is no believing that is not believing something to be true. The thought is then that something similar is true of motivations: perhaps motivations and actions have

[12] Velleman (*op. cit.,* 116). It must also be noted that Velleman admits that he is unclear on the meaning of the "to-be-brought-about" locution; see *op. cit*., n. 31.

[13] See Smith (1987), and Finlay & Schroeder (2012) for another, recent formulation.

[14] Indeed, Schroeder seems to outright accept GG in that passage (*ibid.*), though it amounts to a brief remark in a footnote.

[15] According to the version of the Humean theory in Sinhababu (2009), no mental states other than this desire-belief pair are necessary for action. The supervenience move would therefore be available to him only on the additional hypothesis that this appearance state is not distinct from the belief-desire pair.

intrinsic correctness conditions, so that motivations and actions as such are liable to be evaluated as correct. Further, motivations may represent their own correctness conditions as obtaining. Since the correctness conditions of belief are standards of theoretical reason, then those of motivation and action must be standards of practical reason. Both theoretical and practical attitudes would then represent their objects as meeting the standards in light of which reason assesses each attitude as correct.[16]

These reflections point us in the direction of an improved understanding of GG, for note that the notion of *to-be-doneness* we adverted to above need have no connection to the standards of practical reason. Controversy appears when, and only when, "good" is interpreted in terms of a notion implicated in the standards of practical reason, typically *reason*, *value**, or the practical *ought*. And indeed this is how GG has been tacitly interpreted.[17] GG has also been rejected by those who think that action for a reason need have no particular connection with *good* reasons.[18]

3.2.2 The standards of practical reason

It may help at this point to clarify the meaning of "standards of practical reason". This is difficult to do in a way that meets with general approbation, for there is significant disagreement over the nature and scope of practical reason and practical reasoning, and philosophers' understanding of both of these has been changing quickly in recent years.[19] Nevertheless, for the sake of clarity it would be better to hazard an initial characterization of what I mean by the phrase. It is not simply a stipulative definition, and I do mean to elucidate the concept as it is often used in the literature, but as with nearly any elucidation of a philosophically contentious concept, the one I offer here will also be a reformative definition in some respects.

One more prefatory note. It must be admitted that "the standards of practical reason" is a misnomer for the concept that philosophers who use it have in mind, and that "the norms of practical reason" is in some respects better. A standard is just a condition, a way for the world to be, but considered as one with which the world "is to fit" in some way or another. In this sense the norm *everyone ought to be healthy* shares a standard with the imperative *be healthy, everyone!*. However, "norms of practical reason" is not perfect for the job either since it does not readily embrace

[16]This analogy is investigated in more detail in §4.2.

[17]See Velleman (1992a, 100).

[18]Humeans about motivation are often interpreted this way, and Setiya (2007) argues at length for this thesis. See also Albritton (1985, 248): "But having to do a thing does not settle magically the question whether to do it or not. ... It isn't for reasons, in the end, that we act for reasons."

[19]For instance, it is my impression that the emphasis of the literature has shifted from debates over instrumentalism and the rationality of morality towards other questions: whether one necessarily has reason to be rational, the proper relations among normative concepts including which, if any, is primitive or fundamental, and whether rational thought should be construed more as, on the one hand, following rules or conforming to requirements of coherence, or on the other as responding to one's reasons, at least as one sees them.

evaluative* criteria such as *it is good* to give to charity*. The good news is that the difference between standards and norms does not matter for this dissertation. For that reason, I will allow myself to use both terms interchangeably.

In a general sense, the standards of practical reason are evaluative[20] criteria for action and practical thought that necessarily have normative force for the agents to which they apply, at least in certain circumstances.[21] I take the notion of a standard's having *normative force*, or its *being normative,* for an agent to be conceptually primitive, so that it can only be filled out by pointing to some platitudes. In this sense normative force is a relational notion: a standard may have normative force for one agent and not another. Furthermore, standards can only have normative force *for agents*, so that it makes no sense to say that the standard *that the house be tall* is has normative force for the house; at best it has normative force for the builders. Paradigm standards that do not have normative force for most of the people to which they apply are blue laws or social conventions that one has no reason to follow in one's present circumstances. It may still be true that one legally ought not to spit on the sidewalk within the town limits even if there *really* is nothing to be said against spitting. Arguably, even intrinsic standards, those that apply to a thing in virtue of its intrinsic properties, can fail to have normative force: a good burglar steals without getting caught, but it would be sensible to hold that even a burglar has no reason to burgle. A standard recommending φ which has normative force for an agent generally gives that agent a *pro tanto* justification for φing, although I hasten to add that this need not be a justification *to others*.

Otherwise, I believe Jean Hampton (1998, 88) captures the notion of normative force nearly as well as it can be captured:

> ... besides "authority," we speak of a norm's "prescriptivity" or its "obligatory force" over us, its "compelling nature" or its "pull," its status as an "order" or a "command" (and not a mere "suggestion"). ... It is this last sense — the sense in which a reason is taken to be a directive for the agent — even if the reason has little or no motivational power — that I am interested in pursuing here.

Note, as Hampton does, that normative force is conceptually independent of motivational force. The one amendment I wish to make to Hampton's characterization is that according to my use of the term, normative force need not have *directive* strength. It comes in degrees, and a standard with less normative force can be outweighed by one with more.

Reasons are typically held to necessarily have normative force for the agent for whom they are reasons. It is frequently unclear, however, whether philosophers mean for this fact to arise out of a stipulation on the use of the word 'reason'. As a stipulation I cannot argue with it, but I wish to

[20] Again, the use of 'evaluative' without an asterisk denotes the broad sense of the term. These criteria include those according to which one is reasoning well/poorly, rationally/irrationally, correctly/incorrectly, permissibily/impermissibly, etc.

[21] As we will see below, this last qualification allows for heuristics to serve as standards of practical reason.

caution against making this stipulation. For although it does seem true to me all stipulation aside, it does not seem to be an *analytic* truth. One could, for instance, coherently hold that for some agent *S*, *S* necessarily has reason to do what *S* morally ought to do, that there are things *S* morally ought to do, and yet that nothing *S* morally ought to do has even a smidgen of normative force for him. Stipulating that reasons necessarily have normative force removes the possibility of finding an enlightening explanation of the fact that the described situation is not possible. In any case, note that it does not follow from the fact that reasons necessarily have normative force that *only* reasons have normative force, or even that a standard is normative for an agent only if there is a reason for the agent to conform to it. A rule of rationality, such as that one ought not intend the end and yet not intend the necessary means, may have normative force for an agent in some other way, perhaps simply as an ideal of rationality, even if she does not necessarily have a reason to conform to it.

There is a difference between a standard's applying to an agent and its being normative for that agent. A standard applies to a thing just in case there are true application-conditions: that it *ought* to meet the standard, that it is *good** to meet the standard, or perhaps simply that the standard is *to be met*.[22] But we have already seen an example of a standard that applies to an agent without being normative for them in the case of the blue laws: it is true that you legally ought not spit on the sidewalk, and so that legal standard applies to you, though it does not have normative force for you. This distinction is one philosophers would readily acknowledge but it is often ignored in practice. A case in point is Stephen Darwall (1992, 156) (NB that although Darwall speaks here of the rational 'ought', in older literature the standards of practical reason often went under the title "the rational norms"[23]):

> ... the rational 'ought' is almost never treated by philosophers as simply one 'ought' among others, on all fours with 'ought's internal to, e.g., etiquette, baseball, or bridge. It is regarded, if only implicitly, as *unqualifiedly normative* — not just an ought- according-to-the-norms-of-rationality, as there might be oughts-according-to-the-norms-of-etiquette, or -baseball, or -bridge. What a person rationally ought to do is whatever he ought to do *simpliciter — sans phrase*, as it were.

Darwall may be right that the unadorned 'ought' is unqualifiedly normative,[24] but it is controversial that the normative force of such a standard could be simply guaranteed by the fact that it applies to one and that its application-conditions have a certain syntactic form. Note that one

[22] More generally, a standard applies to a thing only if it provides the truth-conditions for the application of some normative concept to it. (In an extended sense a standard may be said to apply to a thing to which an imperative has been issued, but imperatives will not be at issue here.)

[23] In more recent literature, especially starting with Kolodny (2005), the norms of rationality have often come to be considered a subset of the norms of practical reason that concern coherence or good inference. (Kolodny himself holds that rationality is merely apparently normative.)

[24] Though note that it is unclear what Darwall means by the term "unqualifiedly normative". One might think that it refers to a standard that is normative *under certain qualifications*, but given the examples that follow the phrase he may mean something like "normative according to some set of norms".

could sensibly imagine comparing what one ought to do according to the standards of practical reason with what one ought to do *simpliciter* and asking oneself (if the two recommendations differ) which has normative force for one.

We should distinguish the general class of *standards* of practical reason from what we might call "the Standard" of practical reason, which — if it exists — is a rather special standard of practical reason. The Standard of practical reason would be a normatively fundamental scheme for assessing actions and reasoning in any circumstance; among other things it provides the fundamental truth-conditions for claims about what any agent all-things-considered ought to do in her circumstances, and is normative for every agent in every circumstance. Expected utility theory is often held to be a candidate for such a standard: in every circumstance, it requires the agent to maximize the expected utility of his options. Of course, talk of a *Standard*, let alone *a* Standard, may be purely honorific. If particularism is true then the Standard may be nothing more than the total of facts about what agents have what reasons in what circumstances, together with (say) the injunction that an agent ought to do what she has most reason to.[25] The Standard may also be gappy, offering little to no advice to agents in certain circumstances. It may be response-dependent: what an agent ought to do all-things-considered might depend on what his Ideal Adviser might want for him to want in those circumstances, or there may be no determinate fact to the matter until he further specifies his ends (if, for instance, some version of specificationism is true of practical reason).

It is a major project of the theory of practical reason to determine the nature and contours of the Standard. However — to bring the discussion back to the Guise of the Good — it is implausible to hold that whenever an agent acts, the Standard appears to her as such. What appears to the agent will be at best *perspectives* on that Standard: among other things the standards of practical reason that will appear to agents are particular (apparent) normative truths, truths of limited generality, heuristics and guidelines with normative force only in certain circumstances, normatively non-fundamental principles, defeasible principles, appearances of how two such standards may weigh against each other, etc. It is the appearance of one of these standards that GG requires of action for a reason, and the notion of "good" can be taken to correspond to any normative notion that appears in the expression of such a standard, or simply to the standard as a whole. So, for instance, it would satisfy GG that one's action appear to one to be *recommended*, to be what one is *required* to do, or that an action would contribute to one's health and it is *good** to be healthy, etc. Standards may be vague or highly general and multidimensional, such as *that one must not neglect one's child*. There are many different ways a child might be neglected — physically, socially, intellectually,

[25](Moral) particularism, espoused by among others Dancy (1993a, 2000) and Little (2000), can be roughly characterized as the view that there are no general (moral) principles which explain more particular normative facts and which are capable of guiding action. See Ridge & McKeever (2006, Ch. 1) for an excellent overview.

emotionally, etc. — and whether a child is neglected *simpliciter* is typically a holistic judgment that combines these dimensions. Furthermore it may not be practically possible to make explicit which behaviors are definitely neglectful, not to mention the borderline cases where there is no saying whether they are neglectful.[26] Nevertheless these are good candidates for standards of practical reason, or values (as I shall also call them, for reasons to be made clearer in the next subsection).[27]

The notion of a standard of practical reason that I have aimed to elucidate in this section is rather broader than the one that many philosophers may use that phrase to express. According to the more restrictive conception, the standards of practical reason are standards with normative force that apply to all agents, no matter their circumstance, just in virtue of being agents. Paradigm standards that fall under this concept include purported norms of rationality, such as *that one is rationally required not to intend an end without intending the means one believes necessary to it,* as well as extremely general norms such as *one ought only do what there is sufficient reason to do,* or *one ought to investigate one's options and the reasons for and against them, except when it is costly to do so*.

All of the examples just given fit the concept of a standard of practical reason in my own sense of the phrase, but I think we should also be prepared to admit standards that do not apply to every agent's circumstances. *One ought to be careful with one's words* is good advice for a pope but bad advice for a politician whose career is built on bold, brash, and spontaneous statements. What makes this a good candidate for a standard of practical reason is that it is the kind of norm that the true theory of practical reason will tell some agents to take into account. In short: the standards of practical reason include not just the formal and the fundamental standards of practical reason but *all* the standards of practical reason, including the rather non-fundamental standards that are most relevant for the actual practice of being an agent.

3.2.3 Inertness and publicity

To recapitulate: The possibility that, for all that GG theorists have said, GG could be true simply if it turns out that all action appears under a weak normative light has forced us to reconsider our characterization of the thesis in order to keep it from being too weak. We have thus restricted the notion of the good, as it figures in GG, to the standards of practical reason: GG holds that in acting for a reason one's action must appear to one to meet a standard of practical reason.

Have we thus solved the problem of weakness? In a significant sense, yes. This understanding

[26]This example derives from Soames (2011, 39-40).

[27]Here I am, for the sake of simplifying the prose, being a little careless with my own terminology. *Being healthy* is a standard; *that it is good* (for someone, in certain circumstances) to be healthy* is an application-condition for that standard. But the differences between the two will hardly matter hereafter.

of the thesis is well-motivated by the analogy to theoretical reasoning (which is not to say the analogy is unproblematic, and we will return to a few of its problems in Chapter 4), and it ties the notion of good to concepts studied in a longstanding subject of active philosophical research. Moreover, GG anchors motivating reasons in the standards of practical reason; it holds that one cannot be motivated to do what one does not see as favored, to at least some extent, by practical reason. One cannot completely avoid assessing one's actions by these standards if it is right. Thus, GG entails that anyone who would claim that practical reason is nothing to her in her deliberations either does not act for a reason or is confused about the nature of her actions. This is hardly an insignificant benefit. (We will return to this benefit below in §4.4.)

But we are not entirely out of the woods, for another problem, the *inertness* problem, remains. Note that the significance of disagreement over GG *depends* on a prior conception of the standards of practical reason. Just *what is it* that GG shows agents must care about? Without a substantive conception of practical reason, the thesis that in acting for a motivating reason we have an appearance of (to pick a notion of good) there being some normative *practical reason* to perform that action lacks enough content to exclude the relatively trivial thesis that we act under a weak normative light. Why not say that when the doorknob appears to me as a merely solicited affordance, I thereby represent the existence of a reason to turn it? Any answer to such a question must devolve from a conception of practical reason, and in particular from proposals for the meaning of "normative reason" and substantive truths about normative reasons.[28]

Now of course no one is in danger of answering, "Why not indeed?" My point is that, certain obviously false theories of practical reason aside, today there remains *vast* disagreement over the nature and scope of practical reason; even the desiderata a theory of practical reason must satisfy are under dispute.[29] Furthermore, the conception of the standards of practical reason we've thus far developed was intended to be common to any reasonable theory of practical reason. So to say that in acting we represent our actions as meeting some standard of practical reason, while not vacuous, is hardly to make a claim with well-recognized implications. The worry is rather that GG is *inert* from a theoretical perspective until we hear more about the nature of the standards that, according to GG, are represented in practical thought.

There are a number of ways one might try to press the inertness worry, but the one that concerns me most is *normative* inertness. One of the reasons we may turn to a theory of practical reason is to illuminate some corner of ethics: an ought-implies-can principle may lead us to hope that we

[28]One might think the weakness worry can arise again in another form if the concept of a standard of practical reason simply is that of *whatever* standards are necessarily normative for action, as suggested above. Doesn't that mean it is conceptually possible that *any* standard that one's action appears to meet might be a standard of practical reason? That is true, but I do not think this is a way of extending the weakness worry. Whether the constraint that one's action must appear to meet a standard of practical reason is a weak one is determined not by the concept of such standards but by the substantive facts about them. (Thanks to Peter Railton for discussion on this point.)

[29]See Millgram (2012) for a sense of some of the difficulties involved.

may discover limits on what we ought to do by investigating limits on practical thought. Or we might follow Anscombe (1958, 4-5) in thinking that a proper moral psychology is instrumental to and necessary for an understanding of the virtues. An answer to the question "Just *what is it* that GG shows agents must care about?" might have significant implications for the assessment of our own and others' actions if it shows there to be a significant divergence between the considerations which actually motivate us and those attention to which is favored by the standards of practical reason. In the same vein one might, moreover, be led to think that GG would have important implications for moral philosophy, if it truly holds that everyone acts under the guise of the *good*. Such is the basis of a complaint Bernard Williams (1985, 58) makes against the thesis in relation to Alan Gewirth's (1978) argument from GG to the claim that every rational agent must view her own freedom as good, since they must view it as necessary for them to achieve the goods of their various purposes:

> In any ordinary understanding of *good*, surely, an extra step is taken if you go from saying that you want something or have decided to pursue it to saying that it is good . . . . The idea of something's being good imports an idea, however minimal or hazy, of a perspective in which it can be acknowledged by more than one agent as good.

Hence one version of the inertness worry could press that, for all GG theorists have said so far, they are *not* using an ordinary understanding of the word "good", for they leave open that an agent could act under the guise of a perspective which could not be (correctly) acknowledged by any other agent as good. Indeed Williams seems to have something stronger in mind here, for he seems to be thinking about what he goes on to call "good, period", which he contrasts with "good for me". He admits that agents may be rationally committed to viewing their own freedom as good for them, where goodness-for requires "a perspective that goes somewhere beyond the agent's immediate wants, to his longer-term interests or well-being" (*ibid.*, 59), but he emphasizes that this does not imply they must view their own freedom as a good, period, for there is no reason why anyone else should assent to their being a rational agent with the freedom to pursue their purposes (*ibid.*). Hence the notion of goodness which Williams thinks the very name of the Guise of the Good invites requires not only the *possibility* of acknowledgment by another agent, but that *some* other agent *would be required* to acknowledge one's purpose as good.[30]

This brings us back to the question of the relative *publicity* of the notion of good, or the degree to which it is shared. Williams seems to think that any good worthy of the name must be at least moderately public, requiring as he does that it be shared by at least some other possible agent. But now it is time to seek more precision from our notion of publicity. How does one share a reason, and what does one do with a shared reason?

[30]Thanks to Sarah Buss for pressing me to be more cautious in interpreting Williams here.

One helpful distinction that is related to the public/non-public distinction is that between agent-neutral and agent-relative reasons and principles.[31] (For the sake of simplicity I will focus on reasons for the next few paragraphs, and will return on p. 45 to how this bears on standards of practical reason generally.) An agent-neutral reason has a general form that makes no essential reference to *the agent for whom it is a reason* as such; an agent-relative reason does make essential reference to *the agent for whom it is a reason* as such. In considering whether agent *A* should φ, *that A's φing would lengthen GE Moore's life* is quite a different kind of reason than *that A's φing would lengthen A's life*, even if *A* happens to be GE Moore. The former reason is agent-neutral and the latter, as it can only be properly expressed with a free agent-variable, is agent-relative.[32]

The universal egoist principle is agent-relative, and as noted in §3.1, it is generally considered a private principle: the fact that your φing would be in my interest is, according to the principle, no reason for you to φ unless you happen to be me. But many think that some agent-relative reasons are not private, and may even be fully public: I have reason to help my family, but perhaps the general principle that explains why particular considerations are reasons for me to help my family also gives others some reason to respect my helping my family. Hence agent-relativity is insufficient for non-publicity.

Some agent-neutral reasons give common aims to all agents and are thus public: according to the example above, everyone has reason to do what they can to lengthen GE Moore's life. Note that if two agents act for this reason, then one of them could in principle justify her action to the other by showing how it is instrumental to their common aim. But not all agent-neutral reasons give common aims (e.g., *that A's φing would be courageous*), and at any rate, *pace* Nagel (1970), many of the usual reasons for which agents act are irreducibly agent-relative. As Korsgaard (1993) notes, I and everyone else may correctly think that it is intrinsically good* that a yet-unclimbed mountain be climbed, and so it is true that everyone has a reason to climb it if they can — but if I, a mountain climber, have made it a special ambition of mine that *I* be the first to climb it, I will not be indifferent as to who first climbs it. In this case I am motivated by an agent-relative reason. So if we wish to defend the thesis that the goods on which agents act are exclusively public, it looks unrealistic to construe them in terms of a universal set of good* states of affairs, which is the only kind of public reason agent-neutral reasons lead us to.

[31] The distinction was first defined in Nagel (1970, 90-91), though the present nomenclature derives from Parfit (1984b, 143). I follow the characterization in Nagel (1986, 152-153).

[32] Of course, what is required is essential reference (or lack of it) apart from the fact that a general principle will specify that the reason in question is a reason *for the agent to act*, and likely also only when *that agent* is in certain circumstances; note that even the agent-neutral reason contains one free occurrence of '*A*'. The definition in Nagel (1970, 90) avoids this slight awkwardness, but at the cost of assuming that all reasons are properties of acts that agents have reason to promote.

This way of articulating the agent-neutral/agent-relative distinction presumes that reasons all have a general, principle-like form, which seems to conflict with particularism. For a way of formulating the distinction that accommodates particularists, see Ridge (2011).

But we have learned something about shared reasons by investigating the structure of agent-neutral reasons. Agent-neutral reasons are indeed shared when they give a common aim, but what makes them shared is not merely that practical reason happens to recommend that all agents bring about a certain state of affairs. Rather, shared reasons are currency in the economy of justification: If I offer *S* my reason (or a reason I possess) for my action, that reason is shared with *S* just in case *S* ought to accept it as a *pro tanto* justification (to *S*) of my action. But a *pro tanto* justification of one agent's action to another leaves the latter with a *pro tanto* requirement not to interfere with that action, or perhaps not to criticize it in a certain way, or not to disapprove of it. This gives us a nice derivation of R. Jay Wallace's (2009, 477) conception of public reasons as reasons for one agent to φ that are also reasons for anyone else not to interfere with that agent's φing.

I think we can clarify and improve on Wallace's conception in two important ways. First, note that the proper, general form of agent-relative reasons will involve pronominal reference to the agent of the action: I complete the doily not for the memory of Mrs. Norma Jean, but for the memory of my grandmother Norma. But even if this also constitutes a reason for anyone not to interfere with my doily-making, no one except Norma Jean's other grandchildren can express the reason in this way. Everyone else has reason not to interfere for memory of Mrs. Norma Jean (supposing the reason is public). Shared reasons necessarily implicate two standpoints, the *first-personal* standpoint ("the primary standpoint") of a particular agent S_1 in a position to φ, and a correlate *third-personal* standpoint ("the critical standpoint") of an agent S_2 considering what to do about the possibility or actuality of S_1's φing. This is all by way of saying that the metaphor of reasons as the currency of justification is misleading in a certain respect: what matters about shared reasons is the structure of standards of practical reason for the agents in these two perspectives, not whether everyone is responding to *the same reason*, however we individuate reasons. Second, there is no reason to fix precisely here the actions, omissions, or attitudes public reasons require of an agent in the critical standpoint, nor the strength of the reason. For our purposes we can take as sufficient for a shared reason that it enjoin, to some degree, some kind of *tolerant stance* towards S_1's φing, if not outright aid.

An ancillary benefit of characterizing shared reasons in this way is that we can easily connect them to our discussion of the standards of practical reason. As it turns out, what is essential to our concept of a shared reason is the structure of standards of practical reason that apply to agent in the primary and critical standpoints: a shared standard enjoins, requires, permits, or specifies an action φ as to be performed by a particular agent in the primary standpoint and will make a recommendation to some agents in the critical standpoint to take a tolerant stance (at minimum) towards that agent's φing.[33] A public standard is then a standard of practical reason that enjoins

[33] A shared permission of *A* to φ will generate a correlative obligation not to interfere with *A*'s φing. Note that moral obligations are public obligations, but the 'ought' of expected utility theory, to the extent that we are *obliged*

one agent to φ and that is shared with everyone else. We can even accommodate shared standards' making use of subject-relative good. Consider the standards *everyone ought to do what is best* for himself* and *everyone ought to do what is good*-relative-to-himself*.[34] According to the foregoing characterization of public standards, we could show them to be public if we could show that these principles generate reasons for others to tolerate someone doing (say) what is best for himself. And intuitively, that seems right — that would make the standards public.

It would be good to consider upfront the importance and strength of the thesis I will defend — that the intelligibility motivation for GG tends to support a public conception of the good — as well as to foreshadow the kind of defensive work it requires. Towards that end, I will close out this section by drawing out a few aspects of the relation among the primary and critical standpoints and the standards of practical reason.

The primary and critical standpoints should not be understood merely in terms of a distinction between who is acting and who is watching the agent act. These two standpoints prescind largely from two perspectives we as individual agents can take on the world and ourselves. One is a point of view, irreducibly from *where we are* in the world, in which the self is central not as an object of thought but as that in relation to which everything is experienced. It is also a point of view from which our actions, and perhaps even our thoughts, do not seem to have a origin beyond this center of experience — a phenomenology which some philosophers have taken as a seeming of our actions' not having a cause in events beyond this center. The other is a point of view almost "from nowhere" in which our selves do appear as objects of thought which are understood to be constituted independently of any point of view. From this standpoint I am to myself an object I care deeply for and can control: I am the marionette, among other marionettes, to which my strings are most attached.

Most theorizing about action and practical reason is done with respect to the former, first-personal perspective. One significant class of exceptions to this claim, however, are philosophers especially inspired by Kant. Christine Korsgaard (1996) has done the most to emphasize the place that reflection on our own motivations has for us human agents, but I wish to distance myself from her conception in two respects.[35] First, 'reflection' is too suggestive of a passive process,

to do as it commands, is not. It is a more difficult question how agents in the critical standpoint ought to respond to recommendations and other such standards; I leave this for future work. My thinking here on shared standards has been influenced by Hohfeld's (1913) pioneering analysis of rights.

[34]Relative goodness can be thought of as a generalized version of good-for: like good-for, the extension of these values differ across agents (what is good for me is may not be good for you), but unlike good-for they need have no connection to self-interest or well-being. For defenses and conceptualizations of subject-relative value, see Sen (1982); Dreier (1993); Smith (2009); Portmore (2011); Korsgaard (2013); Cullity (2015); for an influential critique, see Schroeder (2007b). I merely wish to note here that my view can easily accommodate proponents of relative value; as we will see in Chapter 4, my own view is that the standards of practical reason that are in fact represented in practical thought are *reasons*.

[35]For another important Kantian view of agency, see Schapiro (2011). Of course, the distinction between personal

as if one were to idly consider the passing scene of one's actions and motivations. I doubt that Korsgaard would welcome this association either, but it is a good occasion to emphasize that the third-personal standpoint on ourselves mentioned above still is a standpoint *of agency* in which we form intentions and plans and retain causal powers over ourselves. It is the perspective I take when I exercise certain forms of self-control, as when I avoid walking down the ice cream aisle at the grocery store because I know exactly what choices I will make if I do, or when at night I set the alarm on my alarm clock and place it on the other side of the room from where I sleep.

Second, Kantian views of agency often invite the thought that a free agent can "stand back" from all her motivations and give or withhold her endorsement of any of them, *without* that endorsement *itself* being driven by some unendorsed motive. To many philosophers this conception has seemed either incoherent or to imply a vicious regress, so it is crucial to note that the third-personal point of view that I am concerned with requires no such ability. Indeed, the very incoherence of the phrase "point of view from nowhere" should lead us to think that any sort of "objective" point of view that we finite beings can take will be impersonal only in some respects and limited in many. The impersonal point of view is distinguished by the nature of its content and by the facts that it takes into account: from it we view ourselves as just one being in the impersonal causal order in which other agents figure too, and instead of merely thinking and responding *with* our thoughts and motivations, we also think about them.[36]

So, the primary and critical standpoints reflect a deeper distinction between two general points of view agents may take, not a simple distinction between an agent and a bystander. An important implication of this way of characterizing the distinction is that, when we are considering the extent to which the standards of practical reason are shared, we should be aware that the same agent, perhaps even at the same time, may occupy both perspectives. (Indeed, the cases of self-control considered earlier are of this variety.) And we can use this fact to show that the standards of practical reason cannot be *radically* private: there cannot be standards of practical reason that only concern what to do from the primary standpoint and have no implications for what any agent is to do from the critical standpoint.

My aim in the next few pages will be to make a point about certain popular theories of practical reason, but it is not the conjunctive claim that, although many of their proponents favor interpreting these views as radically private, they cannot do so without violating plausible commitments. Instead it is that, once we see the small degree of sharedness of standards of practical reason to which we can show these views are committed, it is hard to see how they can be committed to anything

and impersonal points of view figures deeply in the work of Thomas Nagel, and my use of the phrase "the point of view from nowhere" derives from Nagel (1986).

[36]Do note as well that for Korsgaard reflection opens up space for an agent's reflective endorsement of her motivations, which in turn plays a foundational role in her theory in constituting both one's practical identity and certain norms as authoritative over us. However, these metaethical claims are orthogonal to our purposes here.

approaching fully public standards. That underscores the significance of the tendency towards a fully public conception of the good to which, as I will go on to argue, certain GG theorists are beholden.

A few examples of the popular views I have in mind are expected utility theory and other decision-theoretic principles of choice that presume an agent-relative utility function,[37] many desire-based theories of reasons and the practical 'ought',[38] Humean constructivism about reasons,[39] and universal egoism (which is admittedly more popular among non-philosophers than philosophers). These theories are particularly interesting because they imply that, for some agents with a reason to φ, there are possible circumstances in which no other agent has the slightest reason to tolerate their φing. This is clearest to see in the case of egoism: perhaps everyone else stands to lose from one person's gain. For the sake of simplicity I will concentrate below, as in §3.1, on egoism.

One might be led to think, from the fact that according to egoism no other agent may have reason to tolerate one agent's actions, that egoism entails the same about any other point of view than that from which the agent acts. But in fact this is not so. First, a moment's reflection will tell us that at the time of action, the standards of practical reason cannot both overall require that the agent φ and that from the critical perspective she not tolerate her own φing. The standards of practical reason surely cannot embody this kind of contradiction, nor can they license the kind of incoherence it would require from an agent. It would embody the same kind of contradiction I would express by saying that although Paul Boswell is 5 feet 9 inches tall, *I* am not. "Paul Boswell is 5'9"" and "I am 5'9"" (as said by me) express the same proposition and thus necessarily share truth-values. In asserting them I give voice to contrary thoughts on the same proposition. Similarly, it is reasonable to think that the standards of practical reason govern an agent's acting (perhaps for a certain reason) in certain circumstances, not the perspective from which she acts. Intentionally φing (from the primary perspective) involves being in favor of φing while being intolerant towards one's own φing involves being against one's φing. These are contrary perspectives on the same action. That is why φing intentionally and being intolerant towards one's own φing could not all-things-considered both be appropriate perspectives on the same action, and why to instantiate both at the same time constitutes a rationally objectionable form of incoherence.

Some might think that there are practical dilemmas, just as some philosophers believe there are moral dilemmas. A practical dilemma would be a choice situation in which each of an agent's options are impermissible according to the standards of practical reason. The claim I made in the last paragraph is consistent with the existence of practical dilemmas, however, because a dilemma

[37] E.g. Jeffrey (1983); Buchak (2014).

[38] E.g. Schroeder (2007a). I especially have in mind desire-based theories which, unlike Smith (2011), do not hold that the desires of the ideal agents on which reasons are based converge to a significant extent.

[39] E.g. Street (2012).

involves at least two options each of which are impermissible and not permissible; here we are concerned with the possibility of one option that is both permissible and impermissible. Some might also think that certain forms of incoherence are indeed licensed by practical reason: perhaps, in order to avoid acquiring bad habits, we ought to think we ought not φ even in cases when we ought to φ, or perhaps the correct solution to the puzzle of the self-torturer is that although the torturer ought individually to take each opportunity to turn up the dial to get more money, he ought not take *all* the opportunities altogether.[40] But odd cases such as these tend to involve either clearly distinct options (φing vs believing that one ought to φ) or tricky questions concerning how to integrate synchronic and diachronic norms; neither is the case here.

The second premise is that if the standards of practical reason require an agent A_1 to tolerate (from the critical point of view) at time t agent A_2's φing-at-t, then at least in a stable set of circumstances — where the agent's circumstances includes facts about her psychology — the standards of practical reason will also require the same toleration of A_1 as we consider A_1 in a neighborhood around time t. A *mere* change in time will not change what the standards of practical reason require of an agent with respect to an event that is fixed in time. If they require that A_2 at t be against A_1's φing, then they will require the same of A_2 at $t \pm \varepsilon$. Indeed, typically vast changes of circumstance are required. I have in mind Derek Parfit's case of the Russian nobleman who late in life renounces his youthful idealism;[41] it seems that only in such cases could the standards of practical reason allow an agent to change how tolerant he must be towards the actions that, at one time in his life, it required him to perform.

But from these two premises it follows that if an agent is overall required by the standards of practical reason to φ at t, then there is a corresponding requirement on another (possible) perspective: at times near t the agent must take a tolerant stance towards his φing/having φed. For instance, if an agent correctly φs at t, then it cannot be correct for him to immediately overall regret having done so.[42] So, the standards of practical reason are at the very least shared intertemporally within agents, and thus are not radically private.

But note that this guarantees only an extremely limited form of sharing. Moreover, it is hard to see how we could show that the standards of practical reason must be shared to a greater degree in advance of evaluating particular theories of practical reason. Suppose, for example, that there is only one standard of practical reason, the egoistic principle that everyone must act so as to maximize their own self-interest, and that it is common knowledge that this is so. George, who

[40] The puzzle is described in Quinn (1990), and this particular solution I owe to Tenenbaum & Raffman (2012).

[41] Parfit (1984a, 327-328).

[42] It may be odd to think of critical or tolerant stances towards one's own past actions as *practical* stances given that one cannot change the past. But regret for one's own actions, as a form of self-criticism, is hardly a practically neutral evaluation, involving as it often does an intense desire that one had done otherwise. The same is true of toleration, which is a kind of endorsement of what is tolerated (at least in the weak sense that what one tolerates one does not find worth actively opposing).

is visiting another town, catches Ayn the shopkeeper's attempt to shortchange him. George might recognize that if he were to do as he ought and the roles were reversed, he would do the same, and egoist that he is it seems intuitive to say that he understands perfectly well why Ayn tried to shortchange him. But for all that we have said so far, nothing prevents George in his actual situation from correctly (according to the standards of practical reason) being entirely opposed to being shortchanged. Indeed, that is precisely what egoism seems to entail of his situation.

That is what makes it all the more surprising that one of the main motivations for GG, the intelligibility motivation, may rule out theories of this form by requiring that all standards of practical reason that can motivate an agent must be *public* standards: Ayn and George must have at least some *pro tanto* commitment to respect each others' self-interest by tolerating each others' attempts to further it. The following section will begin to lay out the argument for this claim by setting out the contours of the intelligibility motivation.

Hereafter in this dissertation I will take it as established that the notion of the good relevant to the Guise of the Good is the notion of the standards of practical reason, and that talk of action appearing good is talk of action appearing to meet a standard of practical reason. Hence I will use 'the good', 'value', and 'evaluative' (without asterisk) to refer to the standards of practical reason.

3.3 Intelligibility

Many GG theorists argue from an alleged constraint on action, namely that it must be intelligible to its agent, to the conclusion that action must appear good to its agent. However, the argument is rarely developed and defended at length, and although there are commonalities among GG theorists' understanding of the intelligibility motivation, there is yet little uniformity. In this section I offer a brief tour of the supposed intelligibility of action with the aim of presenting a coherent and unified GG theory of the intelligibility of action. The account culminates in §§3.3.3 and 3.3.4, in which I propose accounts of the relevant sense of the intelligibility of actions with respect to agents in both the primary and critical standpoints (respectively) — that is to say, I offer an account of what it is to find one's own prospective action intelligible as well as what it is to find another's action intelligible.

3.3.1 In what way intelligible?

The intelligibility constraint (IC) holds that an action is performed for a reason only if it is *intelligible* to its agent, or that an action that is truly unintelligible to its agent is not an action for a

reason.[43] IC is thought to be a pre-theoretic datum in need of explanation, which is what enables it to motivate GG: a GG theorist will hold that acting for a reason must involve seeing the action as good, and that it is in virtue of this latter that the action is intelligible.[44]

There are a few not-so-trivial differences in the way the thesis is presented in the literature, however. Often GG is offered as an explanation of why desires rationalize action (Quinn 1993, 28-29; Gregory 2016). This might seem to invite quite different issues in light of the fact that IC says nothing explicitly about desires or rationality, but there is a common thread here. The use of "rationalize" dates back to Davidson (1963, 3), who stipulated that *r* rationalizes *A*'s φing just in case *r* explains *A*'s φing by giving the reason for which *A* φed, and Quinn (*ibid.*) seems to follow Davidson's usage.[45] And both of these theorists have often been accused of using "desire" in a "placeholder" sense in which it is analytic that any state that motivates an action that is done for a reason is a desire.[46]

According to this version, what GG is supposed to account for is just how it is that we can explain an agent's action in terms of what gives the agent's reason for it — or, to put it in a less stilted form, how we can explain action in terms of the agent's reason for it. Arguments for GG in this vein proceed by attempting to establish that actions which are not seen by their agent as good are not interpretable as motivated by a desire, since the relevant motivational state in virtue of which they are held to act cannot be seen as *giving a reason* for that action. But invariably the intuition that the state does not give a reason depends on the apparent unintelligibility of acting on the basis of that state.[47] That is what is so convincing about Quinn's infamous Radio Man, who is in a bizarre functional state that causes him to turn on any radio at hand — though he does not turn them on in order to hear anything, or indeed in order that anything else happen (*op. cit.*, 32).[48]

Another common formulation is that statements giving the reason for an agent's action are

[43] IC is supposed to articulate a deep, robust truth about the nature of human action, so the claim is not that actions *just happen* to satisfy it, but discussion of the particular grade of necessity with which this constraint holds would take us too far afield at this point. As discussed in the conclusion (Chapter 5), given the particular theory of intelligibility developed in this chapter and the commitment to a non-necessary form of GG in the next, I am committed to there being some slippage between intelligibility and GG: either IC is not *necessarily* true or it does not entail GG.

[44] As it happens it is sometimes the *converse* of IC that is at issue in discussions of intelligibility and GG — see for instance Stocker (2004, 303) — but this version is not important to our purposes.

[45] In that same essay Davidson famously argued that explanations in terms of the *gives reason for* relation were causal explanations and furthermore assumed that the first argument of that relation was occupied by the agent's motivating reason itself. These two theses encourage a psychologism about motivating reasons, since it is plausible that what explains actions are psychological states. However, since we will go on to reject psychologism in Chapter 4, it bears mentioning that there is another plausible interpretation of the *gives reason for* relation: *p* gives the reason for *A*'s φing only if *p* is a psychological state that *represents A*'s motivating reason for φing.

[46] See for instance Schueler (1996, 33ff); Schapiro (2014).

[47] Cf. Gregory (2016, 12 n. 9) who, although he treats *rationalization* as an intuitive concept, also glosses it in terms of what actions will make sense to an agent in the light of a certain state.

[48] See also Tenenbaum (2007, 237), who argues that it is the "perplexing character" of the compulsive hand-washer's hand-washing that prevents us from judging that she *wants* to wash her hands.

required to be intelligible if they are true.[49] This is different in two respects: (a) it is *statements* which are held to be intelligible (b) *absolutely*. To take (b) first, it is an interesting feature that actions do not seem to be intelligible to one agent and not in principle to another, and we will return to this below (p. 71). One possible explanation is indeed that *intelligibility* is primarily an absolute notion — statements or actions simply are intelligible or not — and they become intelligible *to* someone when she grasps the features in virtue of which the statement or action is simply intelligible.[50] But for the time being we can be neutral on the relationship between intelligibility and intelligibility-to, since IC merely makes a claim about what is intelligible *to* the agent of an action, and thus what is intelligible to her from the primary standpoint. It is also important to keep in mind that the word 'intelligible' is slightly misleading. IC makes a claim about how the action is *actually cognized* by its agent, not whether or not it *could possibly* make sense to him. An action that could have been made intelligible to an agent but was not is still thought to have a surd quality.

Given that IC is concerned with intelligibility to the agent, (a) might seem especially strange, but there is a good explanation of it. For it seems that we assess the intelligibility of others' behavior, and often our own past behavior, by making it intelligible to us now. And that process will often involve observing (or remembering, or hearing recounted . . .) behavior and assessing a proffered explanation of it. But it is important to keep in mind that language may not play a constitutive role in making actions intelligible. Anticipating the pleasure of an activity is often held to be a paradigmatic way of making it intelligible, but that need not take the form of thinking to oneself, "This will be pleasant." It can simply *seem fun*.[51] (Do note that although IC concerns what is intelligible to the agent of an action, we can make others' actions intelligible to us from the critical point of view, and it is desirable that a theory of intelligibility offer an account of what makes an action intelligible to an agent in both the primary and critical standpoints. We will return to this project below.)

Even all this ground work we've just done, however, hardly suffices for a positive characterization of the precise sense in which IC holds actions to be intelligible. IC is not concerned with every sense in which an action may truly be called "intelligible". The sense in question, which I'll denote with 'intelligibility$_A$', is often conveyed by referring to familiar Anscombe-inspired examples. We have already seen one in the form of Quinn's Radio Man, and here is another:

> Suppose, for example, that you notice me spray painting my shoe. You ask why I am doing that, and I reply that this way my left shoe will weigh a little more than my right. You ask why I want the left shoe to weigh a little more. Now suppose I just look at you

[49] See Anscombe (1963, 26-28) *et passim*; Raz (1997a, 55).

[50] MacIntyre (1986) holds a view of this sort.

[51] Of course, a major influence on Anscombe's adoption of this formulation of IC was the intellectual atmosphere at Oxford at the time, especially post-*Investigations*.

> blankly and say, "That's it." I seem not to understand your puzzlement. You grasp for straws. "Is this some sort of performance art, on the theme of asymmetry?" "No." "Is someone going to weigh your shoes as part of some game?" "No. Why do you ask?" (Clark 2010, 234-235)

Another suitable example is Anscombe's own of a neighbor who one day takes all the green books from his house and spreads them carefully upon his roof, and who when prompted for an explanation offers merely that he feels like it, or is doing it for no particular reason.[52] There is also Raz's contention that one cannot drink coffee for love of Sophocles.[53] All can be used to pump the intuition that the proffered explanation utterly fails to give proper insight into the agent's reason for acting. Without that insight the action is unintelligible$_A$ to us, and we cannot confirm that it is intelligible$_A$ to the agent either.

We can also glean some insight into intelligibility$_A$ by collecting some of the *orbiter dicta* philosophers have used to characterize it. As Anscombe would say, to judge another's action intelligible$_A$ we need to see the *point* the agent saw in so acting, and not just anything the agent could, conceivably, sincerely say about her reasons or motives would do. We might also say that an action is intelligible$_A$ to someone just in case they see *why* it was done. But not just any mechanistic explanation would so, since "blind urges" may explain actions even if there is no positive answer to the question of why such an action was done.[54] Philip Clark (2010, 234) characterizes intelligibility$_A$ as requiring "a particular kind of thought that is neither value judgment per se nor evaluatively neutral factual belief."[55]

It is important to differentiate this notion of intelligibility from others. Note that in some cases showing an action to be intelligible$_A$ to its agent may also show her to be utterly delusional, so that insanity and intelligibility$_A$ are not mutually exclusive. If we learn that the man organizing the green books on his roof is convinced that Satan will take his soul unless he places them *just so*, we do see the point in his acting. An action can be intelligible$_A$ even if a decent person, or a person of sound mind, would find it unimaginable or unthinkable to do. Uncharacteristic actions can also be intelligible$_A$: suppose Scrooge were to have a momentary change of heart that leads him to appreciate Bob Cratchit's hard work, and for that reason gives him a Christmas bonus before the spirits even get to him. Scrooge's business associates might find his action absolutely baffling, unintelligible, but only in the sense that they cannot explain it in terms of his character. Familiar actions, such as those resulting from a recurring compulsive urge, can also be unintelligible$_A$.[56] Akrasia, too, is generally thought to be unintelligible in some way if it is not altogether

[52] Anscombe (1963, 26-27).
[53] Raz (1997b, 8).
[54] Cf. Schapiro (2014, 143).
[55] [For this note see Appendix A.]
[56] Clark (*op. cit.*, 235-236) makes this point.

impossible, though it is in most cases easy enough to ascertain the point the agent saw in it.[57]

Distinguishing these distinct senses also has the effect of undercutting an important source of skepticism about IC. Kieran Setiya writes that "[i]n the ordinary sense of the word ['intelligible'], acting for a reason—even a reason one sees as good—is not always intelligible", and gives in support of this claim an example quite similar to that of Scrooge above.[58] But once we separate this notion of intelligibility from the one relevant to GG, we see the objection is aimed at the wrong notion. There is no such thing as *the* ordinary sense of the word, and the sense at issue in debates over GG is drawn from a unique class of cases.

3.3.1.1 Substantive intelligibility

It seems clear that the intelligibility$_A$ of an action to an agent depends on that agent's mental states and actions, and that IC could be formulated as a constraint on an agent's mental states and actions in the case that she acts for a reason. However, we must be careful to avoid two deflationary understandings of this constraint.

One is a *structural* reading on which IC rules out certain combinations of mental states and actions. On this view, an action is intelligible$_A$ to an agent if the agent's performing the action does not depend on the agent instantiating (or performing) a combination of mental states and actions it forbids. A natural candidate for this version of intelligibility$_A$ would be a wide-scope instrumental principle: on this view an action φ is unintelligible$_A$ to an agent just in case he intends by φing to achieve a certain end E, believes that by φing he will not bring about E, and φs nevertheless; φing is intelligible$_A$ to him otherwise.[59] But this version of intelligibility makes for a bad fit with our previous characterization of intelligibility$_A$. There are, on this version of the instrumental principle, three ways of conforming to it, one of which includes *not acting*. The mere fact of

[57] Davidson (1970a, 42) makes the same point. Do note that some GG theorists hold that one acts for a reason only if one sees *sufficient normative reason* for so acting, so that a certain kind of akrasia is impossible. (Thanks to Sarah Buss for reminding me of this.) I say "a certain kind" because this version only rules out acting for a reason *without seeing one's action as having sufficient justification*. It does not by itself rule out a common formulation of akrasia on which akrasia is acting *against one's better judgment*, where one's better judgment may be a belief that one should not so act. In other words, even this strong form of GG admits of the possibility of agents who do what *appears* to them to be supported by sufficient reason even though they believe that they should not do it, and thus are committed to believing that their appearance is inaccurate.

[58] Setiya (2007, 63). There is one subtle but crucial difference between the two cases, however, and it is that Setiya's Scrooge stand-in is depicted as *merely believing* that his action is one he ought to perform, without appreciating why. (The same kind of example is used for the same purpose in Stocker 2004, 315.) But this introduces a second potential source of unintelligibility: is Scrooge supposed to be unintelligible because he is acting uncharacteristically, or because there is no particular *kind* of good he sees in acting? As I explain below, GG theorists can and should say that the latter kind of action is unintelligible$_A$.

[59] Of course I do not mean to endorse this as the only or best way of formulating a wide-scope instrumental principle. It is only offered by way of example. A more common example is that found in Bratman (2009): if I intend end E, believe that M is a necessary means to E, and that M will occur only if I intend E, then barring any change in my beliefs, rationality requires that I either intend M or give up E.

conformity to that standard, therefore, does not seem to give us any insight into the agent's point in acting. And while there does seem to be a kind of unintelligibility in violating the instrumental principle, especially in synchronic violations, it is more akin to that found in akratic behavior, for we generally can appreciate the point the agent saw in acting when he is akratic or means-end irrational. In instrumentally irrational behavior it is given by the end of the agent's action; it just is the case that the means taken to the end is not the one the agent deems necessary.

This might lead us to a *weak* view of intelligibility$_A$ on which making an action intelligible$_A$ does require certain mental states, but nothing more than that an agent act with *some* end or other in mind, or perhaps simply that she act under a weak normative light. James Lenman (2005, 39) advocates such a view when he writes, "[f]or bare intelligibility we don't need a desirability characteristic, just a[n] explicit specification of propositional content."

Much as with the overly-weak reading of GG (WNL, p. 35), that action must be intelligible in one of these weak senses seems to be almost a trivial truth. For that reason, such a view might seem to be an obvious non-starter for someone hoping to use IC to motivate GG. But that would be a little too quick. Within the GG camp the purest development of an idea of this form is found in Sussman (2009). Sussman seems to put the locus of intentionality in being guided by a concern — any concern. He conceives of such guidance as a dynamic process involving awareness of what one is doing, having a (relatively) clear idea of a standard for what one ought to be doing, and calibrating one's actions to the standard in light of what one knows oneself to be doing. He goes on to argue that, because of the complexity and richness of the good, actions which are relatively more intentional will be guided by the good, whereas perverse actions are limited in their complexity and thus in their intentionality.[60]

I do not doubt that failures of Sussmanian calibration, like failures to conform to the instrumental principle, make for some kind of unintelligibility, and given our characterization it seems that a criterion like Sussman's describes a necessary condition on intelligibility$_A$. After all, seeing the agent's point in acting entails that the agent has some point or other in acting. However, this does not entail that just any point will do for intelligibility$_A$, provided it is sufficiently complex. Note that while the Green Book Man's actions may demonstrate considerable conformity to various (pointless) standards and means-end efficiency in meeting them, his action is still unintelligible$_A$. And because the Green Book Man is a paradigm of unintelligibility$_A$, it seems that Sussmanian intelligibility is insufficient for intelligibility$_A$.

[60] Although Sussman officially takes the complexity of the guidance task and one's success in it to be together proportional to *intentionality* (*ibid.*, 620), one of his main theses is that "intentional action must be seen in terms of some point or guiding concern, where the clearest way something can have a point is to be thought good in some way" (*ibid.*, 623), and he takes for granted that his view better explains why perverse action (that is, action for the bad as such) is intelligible than does Raz's (1999). To that extent I take Sussman also to be offering an account of intelligibility$_A$.

Altogether, these reflections point in a single direction: IC requires rather *substantive* kinds of mental states and actions from agents. It is not a permissive constraint, as the structural and weak understandings of it would have. Of course, this is not to say much about what substantive restrictions would do, nor about the distinction between the substantive and the structural as it applies here, but the examples — actions must appear to be healthy, pleasant, demonstrating intact liberty in the unsubmissiveness of one's will, etc. — are highly suggestive of a core feature of the very idea of intelligibility$_A$: whether an action is intelligible$_A$ to its agent depends on the description of it under which the agent acts, or on the content of his mental states — his emotions, desires, intentions, beliefs, etc. Certain descriptions and contents make some actions intelligible$_A$, and others do not. Given that actions done for a reason will themselves be represented by their agents, at least in the sense that no one acts for a reason if she is in every sense unaware of the action, then we can simplify the formulation: whether an action is intelligible$_A$ to its agent depends on the content of her mental states.

3.3.2 Appearances and the forms of the good

According to the intelligibility motivation for GG, what explains IC is that actions are intelligible$_A$ to agents in virtue of appearing good to them. According to the proponents of this view, therefore, the restriction on content we just offered is more precisely characterized by the thesis that actions must appear *good* to their agents.[61] Such claims are often put forward as an analysis or constitutive account of intelligibility$_A$: an action is intelligible to *A* just in case, and because, it appears good to *A* in some way.[62] But in *what* way must it appear good?

A full answer to this question would require considerable space to articulate and defend. Fortunately, most aspects of the general question are not directly relevant to showing that proponents of the intelligibility motivation for GG are committed to a public conception of the good, and so, as noted above (p. 29) we can follow Sergio Tenenbaum in giving a rather minimal characterization of such appearance-states. The word 'appearance' and its cognates as used here can take, at the very least, any of the senses expressed by its uses in the following list (reproduced from Tenenbaum 2007, 39):

[61] In Chapter 4 we will see some grounds for resistance to the thesis that actions are intelligible$_A$ to an agent in virtue of the *content* of the agent's mental states. Indeed, Tenenbaum (2008) holds that desires are appearances of the good not in the sense that the content of a desire implicates the good, but that the very attitude of desire implicates the good. A desire for *x*, Desire(*x*), is according to him better construed as Appears-Good(*x*) than Appears(*x* is good). A fuller reply to this line of thought will need to await that chapter, but until then it helps to note that I am working with a broader notion of content: if a notion figures in the complement of 'appears that', then it figures in the content of the corresponding appearance-state.

[62] Note, however, that I am not here endorsing the account of intelligibility$_A$ under development *as* an account of the very nature of intelligibility$_A$. In fact, we will see in the concluding chapter (Chapter 5) that the argument in Chapter 4 gives us reason to think that this account is true only of the intelligibility$_A$ of human action.

- From far above, the car appears very small.
- Looking only at the evidence you gathered, it appears that she is not guilty.
- It appears red to me, but you had better ask someone else.
- The raccoon appears to be dead.
- Presented this way, the argument appears to be valid, but when we formalize it, we see that it is not.

Note that while some of these senses of 'appears' are perceptual, others are better characterized (as Tenenbaum puts it, *ibid.*) as an inclination to judge. Do note as well that an appearance is a mental state, not the object of such a state (as we might use it to mean when we talk of the appearance of the Duchess at the ball).

Still, pressing questions remain about these appearances of the good. Consider the following two claims, close analogues of which we have argued for:

10. There is a set Q of properties such that an action must appear to have a property in Q if it is to be intelligible$_A$ to its agent
11. An action must appear good if it is to be intelligible$_A$ to its agent.

It does not follow from 10 and 11 that $Q = V$, where V is the set of substantive values (i.e. standards of practical reason under substantive descriptions). That would give us:

12. An action must appear substantively good if it is to be intelligible$_A$,

where an action appears substantively good just in case it appears to meet a substantive standard of practical reason. But 12 is a natural way of respecting the constraint we argued for in §3.3.1.1, and it moreover has the benefit of neatly sidestepping an otherwise worrisome objection raised by both Michael Stocker (2004) and Kieran Setiya. The objection centers around the fact that merely noting that someone's action was performed under the guise of the good does not help make the action intelligible$_A$ to anyone else. To paraphrase a dialogue in Setiya (2010, 97):

> "She is drinking coffee because she loves Sophocles."
>
> "What? That makes no sense at all."
>
> "Oh yes it does. She thinks that makes it *good* to drink coffee."

What has gone wrong here is that we are not given any substantive good, no particular value, under whose guise the coffee-drinker drinks. If we were told that she's drinking coffee because she thinks this *honors* Sophocles, we may be confused as to why she'd think that, but we would then appreciate her goal of honoring a literary figure.

The move from 10 and 11 to 12 is not completely unproblematic, however, for 12 requires the appearance of a *genuine value*. In requiring this 12 cannot be said to mischaracterize the commitments of GG theorists, and especially of those inclined to press the intelligibility constraint: Anscombe (1963, 76-77), for instance, holds that "the good (perhaps falsely) conceived by the agent to characterize the thing must *really* be one of the many forms of the good". Others are keen to press that a gap between what is good and what attracts an agent can arise only when an agent mistakes an action as having genuinely good-making properties when it in fact does not have them (Raz 1999, 27), or that it is impossible for agents to be attracted to what is merely apparently good, at least insofar as they are rational (Lawrence 1995, 129-130).

Two serious objections have been raised against this move. First, why couldn't an action be intelligible$_A$ in virtue of, say, possessing a very specific property which the agent wrongly believes to be a good-making feature (Bond 1983, 54)? This proposal may satisfy both 10 and 11 even if there is no substantive value, no real "form of the good", the action would be presented as having since the agent could deploy the generic concept *good* in such a belief. (A generic concept would in this sense be a thin normative concept, such as *good**, *ought*, or *right*, which is not conceptualized any further: it does not analytically entail any any more determinate normative concepts of that kind, such as the moral good, goodness of a person, legal rightness, etc.) Moreover there is precedent for a version of GG that allows for action to be guided by such generic evaluative concepts (see Clark 2001).

A second worry is that an agent could act on the basis of an action's appearing to him to be, not quite *good* in some substantive way, but substantively near-good — that is, the action appears to him to meet some standard that is a twisted variant of a genuine standard of practical reason.[63] The two objections seem to be strongest together: one might imagine an antebellum Southerner surveying his plantation who takes stock of the racial purity its organization exhibits, and judges that it is (generically) good. Presumably, maintaining racial purity is not sanctioned by the standards of practical reason, but it may very well be that in the farmer's mind it is difficult to disentangle from more benign standards that very well may be, such as maintaining an efficient organization of one's estate.

The first objection is easier to deal with than the second. Its possibility is excluded by the intelligibility requirement, and we can appreciate why by considering an example. Suppose that Ziyad is at a loss for what direction to take his career in and goes to the Oracle to ask for advice. Her answer only puzzles him, however, for she tells him to sweep his bedroom three times by 5 PM today. "Why?", he asks. "Never you mind!", comes the reply. "It'll be good, that's all you need to know." Still, because Ziyad trusts the Oracle completely he does as she directs.

Isn't Ziyad's action intelligible$_A$ *enough* to him, the objector might press? And it does seem

[63] This objection derives from Stocker (2004).

difficult to demur on this point. But I want to suggest that it is intelligible$_A$ in virtue of its similarity to more quotidian cases of advice-seeking. When we were children and didn't know what to do we may have followed our mother's advice even when we couldn't fathom her reasoning. But there it's clear what specific value we found in following, for the world was less predictable to us then and we not as well-equipped to make wise choices and handle misfortune. It's *safer* to follow the advice of a caring adult. Since Ziyad trusts the Oracle, he sees following her as the safe choice.

This contrasts with a case in which an agent really does not see a substantive value in her action. Imagine that you were suddenly taken with the idea that the *best* thing to do is to lie on your stomach in the street, or stand on your toes and touch your nose to the wall. If it is really *only* under this generic description that you act — you don't do it for spontaneity's sake (which arguably falls under the umbrella of creativity), or so as to appear unbound by social norms, or to prove a point — then it does not seem that this is a description which renders your action intelligible$_A$ to you. And that is as it should be, since it does not seem you do it for a reason either. It seems to me to be more easily interpreted as a case of compulsion.[64] On the whole it seems that we have reason to think that appearances of merely generic goods will not make action intelligible, and that any action that is intelligible$_A$ to its agent will be so in virtue of it's appearing good to its agent in some substantive way.

The second objection is more difficult to dismiss. Examples such as that of the Green Book Man may show that not just any substantive "point" the agent sees in acting will suffice to make that action intelligible$_A$ to her, but that is not the same as showing that *only* substantive goods will do. I confess I have no *a priori* argument that rules out, in a principled way, as unintelligible$_A$ all the kinds of actions this objection envisages. But I also suspect that the reason I do not have one is that none can be found. While it cannot be that our actions are intelligible$_A$ to us any time we act under a weak normative light or a generic normative concept because, it seems, this would allow actions that were too "blind" to be action for a reason,[65] the fact that only substantive standards of practical reason make our actions intelligible$_A$ to us may require an explanation in terms of distinctive human capacities. (Indeed, project of Chapter 4 contributes to such an explanation.) We should therefore distinguish between two theses: (a) that only the appearance of an action as being substantively good could possibly make it intelligible$_A$ to an agent, and (b) whenever our actions are intelligible$_A$ to us, they are so in virtue of appearing to be substantively good. My suspicion is that (a) is false while (b) is true.[66]

[64]For the sake of the discussion in Chapter 4, note that "acting under a description", as I use it here, does not necessarily require the agent to exercise, or even possess, the concepts expressed by that description. One may act under the guise of a non-conceptual representation of a good. Note as well how this case is not as extreme as that of Radio Man, who is not described as acting under any evaluative description at all.

[65]For more on the contrast between blind behavior and action for a reason, see Appendix A.

[66]Note that even if the good is the formal aim of action or desire — that is, if the concept of the good, as it figures in GG, just is *that at which action aims* — it would not help secure (a) without a further argument that action necessarily

If my suspicion is on track, a defense of 12 would require an inductive or abductive argument from (possibly contingent) facts and principles we have discovered concerning human cognition and action. I think such a defense is in principle possible, for we do need some way of accounting for the fact that even sorely mistaken acts and acts of great evil can be quite intelligible$_A$ to their agents. There is obviously no space to mount that defense here, but I can offer two strategies for assuaging worries that might arise from apparent counterexamples. Both solutions depend on the fact that intelligibility$_A$ only requires that an agent's action must appear to her to have *some* point or other and that there can be, as it were, a great deal of chaff with the wheat.

The second version of the Green Book Man illustrates this possibility well, and the intuitive analysis of why his action is intelligible$_A$ to him serves as paradigm for the first model. Recall that on this version the man was laboring under the delusion that Satan would take his soul unless he arranged all the green books out onto the roof of his house *just so*. Plainly, the substantive good that his action appears to him to secure is his own self-preservation, but he is also confused about a number of non-normative considerations, among them that Satan exists and has peculiar preferences about the placement of his green books, that he has a soul, and that the loss of his soul would mean the loss of his self (or perhaps that part of himself which matters morally). So, it seems that if we can conclude upon bracketing an agent's non-normative confusions that there is a genuine substantive good that his action appears to him to achieve, we will have vindicated 12, for in that case we will have shown that the agent is at least acting under the guise of *some* good.[67]

However, one might legitimately worry that this model cannot be extended to all worrisome cases. That would depend upon the possibility of finding, for any case in which an agent seems to act for a reason though not under the guise of a genuine standard of practical reason, a convincing candidate for a *substantive* standard of practical reason that appears to the agent once we identify and bracket any non-normative confusions upon which his action depends. We have no reason to think this will generally be much easier than the task we began with. Take the case of the antebellum plantation owner above: even setting aside certain of his beliefs about the differential natural aptitudes and characters of certain races, and about race as a biological category, we may find certain of his fundamental ideals, or perhaps his conception of his duty, to be unpalatable as candidates for standards of practical reason. His grounds for his stance against miscegenation, for instance, may simply be that people of different backgrounds and skin colors should not mix, a principle he takes as fundamental.

---

aims at the substantively good.

[67] There are additional complications. Is it intelligible$_A$ to me to prefer the destruction of the rest of the world to the scratching of my finger? Perhaps, one might think, self-preservation can only go so far to make protective actions intelligible$_A$ when other considerations prevail. I acknowledge the intuition that there is something incomprehensible to decent people in this, but I do not think it is necessarily unintelligible$_A$ to any human agent. We need not posit an abnormally strong preference for the one's own physical integrity, for all we need to imagine is a person who utterly lacks concern for the rest of the world.

This brings us to the second strategy, which can be seen as a variant of one proposed by Joseph Raz (1999, 32ff.). Raz holds that perverse or defiant actions, those done by an agent for the reason (as they see it) that they are *bad*, are to be understood as actions intended to meet *inversions* of genuine standards of practical reason. Unfortunately, it is not clear how this proposal is supposed to vindicate GG, or *if* it is supposed to.[68] How would this show agents to act under the guise of the good when their aim is quite deliberately to foil it?[69] But there may be something right about the general idea that such deviant cases are to be understood in terms of deviance *from* the standards of practical reason. More precisely, a GG theorist motivated by IC may succeed in attempting to show that apparent cases of agents acting without their actions appearing substantively good to them are to be understood as agents acting under a deviant appearance of the good.

For comparison, suppose that there appears to me to be a moose before me, but that I am also slightly confused about moose and I begin to wonder whether this moose would make a good candidate for Santa's sleigh team. I have confused moose with reindeer. However, my confusion does not entail that the content of my appearance is not after all *that there is a moose in front of me* and is instead *that there is a moose-reindeer hybrid*. For my appearance is correct in case there is a moose in front of me, not in case there is a moose-reindeer hybrid. Nevertheless we might say that my appearance as of a moose is defective since it is also an appearance to me as of an animal that is fit to drive Santa's sleigh, and because of it I am disposed to draw inferences to some false conclusions. Intuitively, the moose appears to me *as such* even when I have not *quite* mastered the concept of a moose, or have significant deficiencies in that concept, or simply have a range of false beliefs about moose.

The point generalizes to the normative domain: a standard of practical reason can appear to an agent as such even when she has not quite mastered the concept, has an underdeveloped capacity to recognize instances of actions that would meet the standard, or simply has a number of false beliefs about what the standard entails. In the case of the plantation owner, it is plausible that he is simply *misapplying* a genuine good, such as that people should take care of those close to them and not meddle in others' affairs. Indeed, it is a point in favor of GG that it can diagnose what the plantation owner gets wrong about miscegenation: it appears to him to violate a standard of practical reason when it in fact does not.[70]

[68] Raz (*ibid.*) describes defiant action as "an exception that proves the rule", which seems to entail that it is an exception to GG.

[69] As it happens I believe the solution to the problem of defiant action is simple: GG holds that actions must appear good to their agents. It does not hold that they may not also appear bad. In the terms of Chapter 4, agents may affectively enrich their conceptual judgments about the badness of certain actions and thereby represent there as being reason to do what is bad.

[70] There are other worries one might have about 12; one might worry that some paradigmatically intelligible$_A$-making motivations do not constitutive an appearance of anything genuinely good, or that what they are appearances of is too defective in some way to be intelligible$_A$ in virtue of what it is an appearance of. I take Martha Nussbaum's (2015) recent work on anger to lie in this vein. She argues that there is something necessarily irrational or ethically

Of course, this is all by way of suggestion, and it would require significant work to turn it into a genuine defense of claim 12.[71] Nevertheless, we have made a preliminary case for an account of intelligibility$_A$ on the basis of 12, recalling that GG theorists are best interpreted as offering a constitutive account of intelligibility$_A$:

Primary Intelligibility$_A$* *S*'s action φ in circumstances *C* is intelligible$_A$ to *S* just in case, and because, φing appears to *S* to be substantively good.

3.3.3 Appearances and rational force

In this subsection I will consider a different aspect of how appearances of the good lend intelligibility$_A$ to actions. The common perceptual metaphors of "seeing good" in one's action and its "appearing good" suggest that the goodness-appearances relevant to the intelligibility constraint have *rational force*, and for good reason — but as it turns out, they do not have the rational force of perception.

The idea of the *force* of a representation derives ultimately from Frege,[72] and in its broadest usage it concerns any aspect of the representation apart from is content. Here I am concerned only with the *rational* force of appearances of the good, with their power to defeasibly make certain other attitudes or actions rationally permissible to hold or perform.[73] Perception and imagination, for instance, differ in rational force. It is one thing to see Barack Obama in the doorway and quite another to vividly imagine him there. The former state defeasibly licenses the belief that Obama is standing in the doorway while the latter does not, even if the content of both states is the same. On the other hand, for a subject who know that *p* entails *q*, a perception that *p* and a belief that *p* might

inappropriate about genuine anger. She notes that it could be taken to motivate us to *get payback* for harm that was done to us, or that motivates us to restore our own social standing by laying our aggressor low. But, she argues, it is not true that an aggressor has incurred a debt that must be exacted in harm, and our concern for our own social standing does not give us reason to mete out the kind of harm anger motivates us to perform, for that would be too narcissistic.

Again, a full response to worries of this sort is not possible here, but I do want to suggest a line of response. *Even if* Nussbaum is right in her analysis of the phenomenology of anger and the ethics of acting on it, that would not count against Primary Intelligibility$_A$* giving a good account of why it makes action intelligible$_A$. The problem with anger, Nussbaum should say, is that it incoherently combines two genuine standards of practical reason into a single appearance: it attempts to add the normative weight of debts and obligations to the relatively unweighty matter of maintaining our social status.

[71] One possibility that deserves exploration is whether the second strategy can be vindicated by an *a priori* argument. It has been a familiar refrain since the work of Donald Davidson that there may be *a priori* limits on just how irrational an agent can be, and that irrationality in agents can only be explained in terms of deviations from the norms of rationality to which they must subscribe. (See for instance Davidson 1985, 195-196.) However, it is not clear that a constraint of this sort can support thesis 12. It seems to me that the constitutive, rational norms of agency are most likely to be, and only be, norms of coherence: there are limits to just how many inconsistent beliefs and intransitive preferences an agent can hold in her head, and how baldly she can hold them. But thesis 12 is not a coherence requirement of this sort.

[72] See for instance Frege (1918, 294).

[73] The notion of rational force I derive from Schafer (2013).

be said to have the same rational force with respect to the belief that q, for both provide grounds for it. We can then say (to regiment our terms) that φing appears to me with rational force if I have an appearance of φing that represents it with rational force, or that the appearance rationally supports φing — and similarly with respect to attitudes.

Now, must an appearance of φ's goodness have rational force with respect to some attitude or action in order to make φing intelligible$_A$, according to GG? That is, if an action is to be intelligible$_A$ to me, must it appear to me in a way that rationally supports my believing that it's good, or my doing it? The question has received little direct attention in the literature. But generally speaking, theorists have held that according to GG, motivations to act rationally support evaluative perceptions and beliefs concerning that action, which in turn rationally support the action so evaluated. Davidson (1978, 86) saw pro-attitudes as expressing value judgments, for instance, and judgments that φing is good do indeed rationally support φing. At least one prominent critic agrees with this interpretation of GG. Velleman (1992a, 115-116), in criticizing GG, argues that GG must show not only that desires present their content as to-be-brought-about but that they must furthermore "aim to get it right" as to whether it really is to-be-brought-about. He holds that this feature is what would enable desire to entail an action-guiding judgment about what really is to-be-brought-about.

The main questions GG theorists need to address in this regard are, first, with respect to what do appearances of the good have rational force? And second, why must they have rational force with respect to that? Davidson and Velleman suggest that the answer to the first question is *evaluative judgments and beliefs*. But judgment and belief imply more commitment on the agent's part than is entailed by generic motivation, and most theorists have sought to preserve the intuition that one can desire something, such as an attractive but poorly-baked cake, that one does not judge or believe to be good.[74] For that reason most theorists have emphasized the quasi-perceptual nature of some motivations. Perhaps desires to φ are sometimes like perceptions of the good in that, under suitable conditions, they license the agent to take an all-out attitude like believing that φing is good even though they do not commit her to it, just as a visual perception that P licenses belief in P under certain conditions without committing the agent to it.[75]

Let us call 'perceptualism' the view that when an action of ours is intelligible$_A$ to us — and thus when it appears good to us, according to the GG view on offer — it always appears to us in such as way as to have the rational force of an evaluative perception. In the remainder of this

[74] One recent exception to this trend is Gregory (2013, 2016), who holds that desires are beliefs about reasons.

[75] See Stampe (1987); Tenenbaum (2007); Schafer (2013); Saemi (2015), among others. Note that I am not attributing to any of these authors the view that desires, or any other kind of motivation, *are* perceptions of values. Recall that Tenenbaum, for instance, thinks that some desires may be appearances of the good in the sense of "appears" in the sentence "Presented this way, the argument appears to be valid, but when we formalize it, we see that it is not." (See above p. 56.) Whether or not this appearance is perceptual, what matters here is that it has the rational force of perception.

subsection I will argue that perceptualism about intelligibility$_A$ faces difficulties accounting for pretense and expressive actions. In its place I will propose *actism*, the view that when an action of ours is intelligible$_A$ to us, it always appears in such a way as to rationally support our performing that action. Actism has a major implication for our understanding of intelligibility$_A$, for it holds that to find an action intelligible$_A$ is already to *favor* it to some extent.

First, the trouble cases for perceptualism:

Lava Duane, a small boy, is being very careful not to step on the grass between flagstones. When asked why he's doing this he replies, "Because if I step on the lava I'll get burned!"

Computer My computer's processing speed has slowed precipitously, and out of frustration I slap the computer. I do not stand back from my frustration and consciously decide that it is better to alleviate it by slapping the computer. I simply act out my frustration.[76]

Intuitively, Duane acts for a reason. What is interesting about his case is that we know the description under which he is acting and which makes his action intelligible$_A$ to him, which is also the reason he gives for avoiding the grass: it is *that if he steps on the lava he would get burned*. How odd this is! Duane presumably doesn't believe that he actually might be burned, nor would it be rational of him to. But then, if perceptualism is true, how could he find avoiding the grass to be intelligible$_A$?

The second case is more contentious. Many philosophers would balk at the idea that I slap the computer *for a reason*.[77] Still, my action is intelligible$_A$ to me, it seems, for in this case as well we can find the description under which my action is intelligible$_A$ to me. If I were to reflect on and articulate my motivation I might say that the computer *deserved a beating* for being so recalcitrant, which would uncontentiously give my reason for slapping the computer if I really believed that the computer deserved a slapping. But of course I do not believe that, and my frustration does not give me good reason to believe it either. Nor can my frustration easily be construed as giving me a defeated reason to believe it. Slapping a computer out of frustration is not, or at least need not be, like falling for a mirage or being overwhelmed by fear on the Plexiglas viewing platform at the Grand Canyon, cases in which it is more plausible that I have a perception-like representation of the good. I am not momentarily taken in by the delusion that my computer is an animate object that is deliberately preventing me from completing my dissertation. It just does not seem to me as if my

[76]Thanks to Daniel Drucker for pressing me to consider this kind of case, and to Sarah Buss for a suggestion that improved the example.

[77]However, we will see in Chapter 4 why this is indeed a case of acting for a reason, or more cautiously (if somewhat unidiomatically) a case of acting *for reason*: one's frustration, as an affective state, represents there as being reason to punish or correct the computer. This is consistent with Hursthouse's (1991, 61-62) contention that such expressive actions are not done for any *further* reason. I take that to mean that in such cases agents should not be understood as responding to a consideration which is their reason for acting. That may be so even if their frustration represents punishing as being *reason-supported*.

computer is *actually* recalcitrant. Perceptualism thus has difficulty explaining the intelligibility$_A$ of this action too.

One option for dealing with the clear case of pretense in Lava is to say that it is a case of acting for a reason, and so GG does apply to it, but that necessarily one sees some *further good* in pretending as one does, perhaps because playacting itself is good (Raz 2010, 115). This seems rather too complicated to me. What makes pretending to avoid the lava intelligible$_A$ to Duane is, as he would put it, *that I'd die if I stepped in it*. It is not the higher-order thought that playacting is itself good because, perhaps, it is essential to one's development as a practical agent (as it likely is, in fact). This is not to deny that children see value in playing as such. The point is rather that the *kinds* of values in virtue of which *particular* games are intelligible$_A$ to them have a great deal to do with the content of the game.

The options for dealing with expressive actions available to perceptualists are not particularly palatable either. One could hold that expressive actions are intentional but not done for a reason, and because of that GG does not apply to them (Raz 1999). It seems to me that a GG theorist who endorses the intelligibility motivation cannot make this reply, however, since the action clearly is intelligible$_A$ to me and thus deserves an account from the GG theorist of the way in which it appears good to me. One could also go in the other direction and attribute a slightly more sophisticated motivating reason to me. Döring (2003), for instance, holds that such actions are intended to vent emotion in symbolic displays, though she downplays the extent to which such expressive actions are rational (in the success sense) and holds that they are not rationalized by the emotion's representational content (*op. cit.,* pp. 224, 227). Perhaps, then, in slapping the computer I am moved by a perception-like appearance of the good of symbolically venting my frustration. Yet this move risks denying the phenomenon since expressive actions are hypothesized not to be done *so as to do* something else. It also leaves Döring with a need to explain the sense in which symbolic actions are intelligible$_A$ to their agents. Why is it that, out of the many actions available to me, I sometimes settle for symbolism when I am frustrated?

I think the best response to Lava and Computer is to treat them as species of the same genus but to deny perceptualism. These agents do not have perception-like appearances of their actions as good. Instead they have *pretense*-like appearances of their actions as good. Duane is pretending that the grass is lava, and he also pretends that it's good to avoid the grass. In slapping the computer I am engaging in the *pretense* that it is recalcitrant and deserves punishment. Of course, this is not to say that I am reflexively conscious of the fact that I am engaging in pretense. The sense in which it appears to me that the computer deserves a slapping is rather more like the sense in which it appears to me that a cloud is a giant, puffy giraffe.

Furthermore, we can after all make sense of Duane's response as giving his reason for avoiding the grass, so long as we recall that his assertion occurs within the context of a game and as such

falls under the scope of a fiction operator (so to speak). For *in the fiction* that he is acting out, the lava at his feet is a very good reason to be careful on the flagstones. Similarly, in the fiction that the computer is recalcitrant, there is good reason to punish it, namely that it deserves a slapping. In this way we can explain the sense in which in Computer, I do have a motivating reason for slapping the computer, although I cannot felicitously cite it as my reason. For if I were to assert that the computer deserves a slapping it would license my audience to infer that I believe that it deserves a slapping, which I do not, for of course the pretense that X is F does not rationally support the belief that X is F.[78]

But crucially, pretense can indeed rationally support action. Indeed, according to the pretense account of these cases each agent is acting intentionally under two descriptions, and the account can explain why both are rationally supported. Duane, for instance, is *pretending to avoid the lava* and the same time he is *avoiding the grass*; in fact, his pretending to avoid the lava is realized by his (actually) avoiding the grass. Similarly, according to the pretense account of expressive action I pretend to punish the computer by slapping it. But note that the *pretense* that φing is good does rationally support *pretending* to φ, e.g. the pretense that there is lava between the stones rationally supports pretending not to step on the lava. For in *the fiction* there does very much appear to be something good about avoiding the lava, namely that one would perish if one did not. (Of course the rational support here is merely *pro tanto* and will need to be weighed against the real-world effects of acting out the pretense.) But because pretending to avoid the lava in this case consists in avoiding the grass between the flagstones, Duane's pretense also rationally supports his avoiding the grass. That, in turn, is because rational support for one action transfers (*ceteris paribus*) to the intentional actions that constitute it, e.g. if it is rational for me to clean the house, then it is rational for me to mop the floors, clean the windows, etc.[79]

So, to sum up the results thus far: it seems we have found in Lava and Computer counterexamples to perceptualism. But these same cases are not counterexamples to actism. Understanding both of these cases in terms of their agents taking up an attitude of pretense allow us to extend GG's account of intelligibility$_A$ and to neatly explain the sense in which these actions are intelligible$_A$ to their agents. And according to the actist understanding of intelligibility$_A$ that results, when agents find an action φ intelligible$_A$, that action appears good to that agent in a way that rationally supports their φing. This leads to the following account of intelligibility (of an action from the primary standpoint):

[78]This might raise the question of what one's reason for engaging in pretense in the first place might be. The answer to this question is that one can have all kinds of reasons for engaging in pretense. But engaging in pretense is not always an action, and nowhere is this more true than in expressive action. I have a reason for slapping the computer, namely that according to the pretense I am engaged in, it deserves it. But in expressive action I do not *choose* to engage in this pretense any more than I choose to dream a daydream I simply fall into.

[79]Note, however, that pretending to believe something does not generally consist in actually believing anything. This helps explain why pretense can rationally support action and not belief.

Primary Intelligibility$_A$ *S*'s action φ in circumstances *C* is intelligible$_A$ to *S* just in case, and because, φing appears to *S* to be substantively good, where this appearance has rational force with respect to φing.

On this actist interpretation of intelligibility$_A$, it is not that such appearances of the good rationally support action in virtue of *first* supporting an evaluative judgment. Rather, such appearances *directly* rationally support that action. (It bears reminding that the rational support involved here is merely *pro tanto*, and can be defeated.)

This result transforms our understanding of intelligibility$_A$. On the present view, there are close connections among acting for a reason, the intelligibility$_A$ of action, rational support for action, and that action's appearing good (at least in the sense of "appearing good" relevant to GG). To find an action intelligible$_A$ is to see the point in acting — which GG interprets as the action's appearing good to one. This in turn is to see an action in a way that rationally supports one's performing that action. And that action's being rationally supported is what *enables* one to perform that action for a reason — namely, the reason in virtue of which the action is intelligible$_A$ to one. Thus, to find an action intelligible$_A$ is not to make a merely passive judgment. It is an inherently active, *practical* stance, as it involves taking up a position of *favoring* one's own acting in light of the substantive considerations by which the action is intelligible$_A$ to one — just as having a visual experience involves taking up a point of view and having a belief involves taking up a position on how the world is.

3.3.4 Intelligibility of others' actions

In order to give a complete account of intelligibility$_A$ we should also consider what it is to make another's action intelligible$_A$ to oneself. It would be useful, however, to clarify at the outset just what what the target of the account is. If my action is intelligible$_A$ to me, then it is the token action itself that is intelligible$_A$: it is (my own) φing here and now, as considered from the primary standpoint. We make someone else *S*'s action intelligible$_A$ to us from the critical standpoint, and are thus considering the intelligibility$_A$ of *S*'s φing. But because of psychological differences between people, often the best that we can do to make someone else's action intelligible$_A$ to us, supposing we do not make an extraordinary effort to match our capacities and experiences to theirs, is to make that general *type* of action intelligible$_A$.

Suppose that my preferred diet is burritos and pancakes, that I have known little else for a meal, and that I have practically no appetite for discovery of new food. At a party I am introduced to a great gourmand, one M. Mirande, who waxes poetic about the meal he had last night: a generous portion of *cassoulet*, filleted pike with *sauce Nantua*, and artichokes on a bed of *foie gras*, all washed down with Bordeaux and capped off by a glass of champagne. As he details how each aspect of the meal contributes to a veritable symphony of flavors, my mind is mostly blank. I can

think of a really great burrito I had once; it was very fresh, and the salsa really complimented the sour cream. It seems that I can appreciate the general kind of aesthetic value that the gourmand engaged in, but I am at present unable to appreciate the full complexity of the gourmand's meal. Making someone's token action (or activity, in this case) intelligible$_A$ to oneself is better thought of as an ideal achieved by those who have shared backgrounds and experiences. Nevertheless, for the sake of analysis I will concentrate on what it would be to make someone's token action intelligible$_A$ to oneself.[80]

This brings us to another set of distinctions. In talking of making someone else's (token) action φ intelligible$_A$ to oneself, we could be speaking of:

13. Understanding *that S*'s φing is intelligible$_A$ to *S*;

14. Understanding *that* and *why S*'s φing is intelligible$_A$ to *S*; or

15. Making *S*'s φing intelligible$_A$ to oneself.

13 and 14 have analogues in the domain of non-normative thought. Suppose you take Little Freddie to his Grandma Norma Jean's house. She's taken her oak table out of storage, and with one look at it Freddie exclaims, "Wow, that's big!" You remember it looking big to you when you were a kid too, and you think that Freddie is right on his part to call it big given his size, but you would not yourself call it big now. Here you understand *that* Freddie thinks of the table as big (the proof of which is that you understand him as asserting just that), and you also understand *why*, in a certain sense, because you recognize his assertion as appropriate given his circumstances, or perhaps because you judge that you would judge the same were you in his shoes.

After looking at this theoretical analogy, it might seem that the analogy is a complete one in that there is no gap between 14 and 15. To see what such a theory would mean in practice, recall Ayn and George (p. 49). Doesn't George "get" exactly why Ayn shortchanged him, and doesn't that mean her action is intelligible$_A$ to him?

It is true that George can see *that* Ayn sees a point in shortchanging him, and he also knows just what that point is — her own self-interest. And given that he too is an egoist, if he is rational he will recognize that he would do the same in her shoes. Moreover, it does seem to be the case that 15 entails 14. Making another's token action intelligible$_A$ to you presupposes that it is intelligible$_A$ to them, and surely involves understanding in virtue of what it is intelligible$_A$ to them. But there are serious problems for the view that 14 and the kind of intelligibility involved in it suffices for 15.

According to the account we have developed so far, making one's own action intelligible$_A$ to oneself involves seeing it as good, that is, as meeting a standard of practical reason. But someone S_1

[80]Thanks to Peter Railton for discussion on this point.

seeing the action of another agent S_2 as good in *this* way clearly is not necessary for S_1's making it intelligible$_A$ to herself, for what matters is not whether S_2's action actually met a standard of practical reason applying to S_2's circumstances but whether it *appeared* to her to meet such a standard. So, one might think, perhaps the sufficient condition is that S_1 see S_2's action as *seen* by S_2 as endorsed by a particular standard of practical reason. We can also easily imagine how S_1 might do this: to revert to our earlier example, George (as S_1) might imagine himself in Ayn's shoes as an agent in the primary standpoint (S_2) about to shortchange George and then consider whether shortchanging appears good to him then.

The flaw in this proposal is that this kind of understanding of another's action does not involve taking up a practical standpoint with respect to it. And as shown in the previous subsection, intelligibility$_A$ requires taking up this kind of standpoint: it involves an appearance with the force to rationalize one's own action. So as we have described him, it does not seem that George finds Ayn's action intelligible$_A$ after all. Consider what we are inclined to say about him: intelligibility$_A$ is a matter of seeing the point in an agent's action, yet as we have described him George does not himself *see the point* in Ayn's shortchanging him. He only sees *that* Ayn sees a point in it, and what point she sees. Her action does not appear good to him, and indeed it appears quite bad to him.

But what would it be to take a practical stance with respect to someone else's action? A good model for one agent's finding another's action intelligible$_A$ is the way in which I find most of my immediately past self's actions intelligible$_A$. Suppose I have just gotten on the bus to go to the grocery store. Unless my information or preferences have suddenly changed, I will not revisit my decision moments ago to have gotten on the bus. And that is because I not only see what point I saw in getting on, *that I get food for myself*, but that very point still has a kind of normative pull on me. Out of the respect for the good I saw then, I am inclined to take a tolerant stance towards my having gotten on the bus, which in this case means I am inclined not to remonstrate with myself for having done so. We can then extend this same line of thought to George: for George to find Ayn's shortchanging him intelligible$_A$ would require him to genuinely *appreciate* the point she sees in shortchanging him, and thus for him to favor tolerating her shortchanging him out of respect for that point.

So it seems that finding another's action intelligible is slightly more involved than finding one's own action intelligible, because it comprises not only an understanding of what good appears to the agent and that it so appears, but also an appearance of tolerating that action's being good insofar as it respects that same good:

Critical Intelligibility$_A$ S_2's φing is intelligible$_A$ to S_1 just in case, and because,

16. S_1 understands the substantive good that is apparent to S_2 *as* apparent to S_2, and

17. tolerating S_2's φing appears to S_1 to be good, out of respect for the good apparent to S_2, where this appearance has rational force with respect to tolerating S_2's φing.

A few notes: GG does not hold that awareness that another person is performing a certain action entails that you make their action intelligible$_A$ to you. And in most cases, only the general type of action the agent performed, and the kinds of reasons for which they might have performed it, will be intelligible$_A$ to an agent. Finally, it must be remembered that the good of tolerating S_2's φing is *pro tanto* and, since it is out of respect for the good S_2 sees in φing, conditional on φ's achieving the good in question. For that reason it may not be all-things-considered rational of S_1 to tolerate S_2's φing if S_1 has sufficient reason to believe that S_2 's φing will not lead to that good. In other words, S_1's satisfying condition 17 does not entail that S_2's φing appears to be *justified* to S_1.

Suppose I see my partner putting on her parka to go outside, and I know exactly why she's doing that, for it's been cold all week. But I've just come from outdoors and I know that it's a surprisingly warm spring day, and there's no point to wearing a parka. In this case Critical Intelligibility$_A$ does not entail that the intelligibility$_A$ of her action to me must rationally support my tolerating her putting on her parka — though no doubt a different argument from a moral requirement to respect persons will enjoin me to be careful in how I stop her from going out with it on. Or for a more extreme example, if I am sufficiently imaginative I can appreciate why religious flagellants do what they do — they think an infallible judge will punish them if they do not expiate their sins in this way — but because I do not share their theistic presuppositions, that appreciation's ability to give me justification to help them whip themselves is defeated.

3.4 From intelligibility to publicity

In this section I will trace a route from the account of intelligibility we have just offered to the thesis that any value that can appear to an agent, and hence (according to GG) any value that can motivate them, is a public value. My purpose is not to defend the route but to demonstrate by limning it just how close we are already to a conception of all such values as public. The route can be summarized quite briefly: Plausibly, it is a fact that we are quite generally capable of making others' actions intelligible$_A$ to us. According to Critical Intelligibility$_A$, rendering them intelligble$_A$ requires an appearance that it is good to tolerate them out of respect for the good the agent sees in acting. The best explanation of these systematic appearances is that it simply is good of any agent in the critical standpoint to tolerate someone's valuable pursuit.

I close the section by considering some implications of this publicity thesis and some reasons one might resist it.

3.4.1 Agency and intelligibility

The first step on the route to public values is the following thesis:

Inter-Intelligibility (INI) If a token action is intelligible$_A$ to its agent, then it is in principle possible for any other agent to make it intelligible$_A$ to herself.

"In principle possible" here means "possible given the right concepts and sufficient imagination to understand the agent's point of view", or in terms of the account of intelligibility$_A$ we have been developing, "possible given that one satisfies the first condition (16) of Critical Intelligibility$_A$". Roughly speaking, INI claims that if S_2 sees a point to her φing, then S_1 could, through understanding why φing is intelligible$_A$ to S_2, come to appreciate that point as well — even if S_1 disagrees with S_2 whether φing would actually further that point given her circumstances.

INI is quite a substantive thesis when it is interpreted in light of the account of intelligibility$_A$ we have been developing. It is perfectly coherent to suppose that someone may be fully cognizant of the state of mind in virtue of which S finds φing intelligible$_A$ but not see any point to S's φing herself. Indeed the example of Ayn and George was intended to illustrate this possibility. Given the coherence of that example it is extremely doubtful that INI is a necessary truth, much less a conceptual one.[81] But I do not think this makes INI philosophically uninteresting, for it may nevertheless be a robust truth about agents like us.

The question of the extent to which an action that is intelligible$_A$ to its own agent must be intelligible$_A$ to other agents has received scant attention in the literature. MacIntyre (1986) is plausibly committed to it as a result of his objective conception of intelligibility, but as we mentioned above (p. 52) that very objectivity also made it unsuitable as an account of intelligibility$_A$. Anscombe (1963) seems committed to INI for roughly similar reasons: for her it is technically reports and attributions of someone's wanting something that are intelligible$_A$ or unintelligible$_A$, and she holds that they are (objectively) intelligible$_A$ if they characterize the object of desire as good (where she also deploys a non-relativistic conception of goodness). Christine Korsgaard has long argued for the thesis that an agent's motivating reasons — which on her view bear a tight connection to normative reasons since, roughly speaking, according to her a consideration or principle becomes a motivating reason for an agent by being reflectively endorsed by that agent — must be "in principle shareable" (Korsgaard 1993, 289), meaning that we must be able to communicate and share in each others' ends (*ibid.*, 299 *et passim*). It seems that as she uses the term, the "communicability" of a reason to someone entails that it is in principle intelligible$_A$ to them. According to a line of thought in Korsgaard (1996) and Korsgaard (2009), the shareability of reasons is a presupposition of the possibility of genuinely inter*acting* with others.

[81] [For this note, see Appendix B.]

However, I do not think the case for INI need rest on claims about the nature and possibility of shared agency, for a strong presumptive case can be based on robust intuitions about our ability to recognize action for a reason. It seems that we are quite generally capable of recognizing action for a reason when it occurs, and that we are not comfortable in judging that an action was done for a reason unless we find at least that general type of action intelligible$_A$ ourselves. And once we provide the case for this conclusion we already have a strong case for INI, since the difference between finding a general type of action and a token action intelligible$_A$ is merely in our grasp of the agent's state of mind, and INI only makes a claim about what we find intelligible$_A$ when we have a full grasp of the agent's state of mind. It is quite reasonable to suppose that giving one agent a fuller understanding of another's state of mind would not make the action unintelligible$_A$ to the first, especially on the supposition that the action is intelligible$_A$ to its agent.

To make the case, consider Anscombe's collector of bits of bone three inches long (Anscombe 1963, 75). Suppose you see her pull a bit of bone out of the dirt and say, "This is a fine one here!", measuring it against her handy ruler. It's obvious what her reason for doing that would be, if she had one: it's *that this piece of bone is three inches long*. But I do not think we are comfortable simply with judging that that is her reason, and taking the issue as settled. Her psychology remains surd in a crucial way, and so we suspend judgment on whether that really is *her* reason for acting while we await more information. Furthermore, no amount of protestations from the collector that she does indeed find her action intelligible$_A$ will assuage us, for it seems that what we need is evidence that her action fits into a type that is intelligible$_A$ to *us*. We need to hear that only bits of bone three inches long will fit into the arc she needs for her art piece, or that such bits make for excellent soup stock — all projects that we can recognize as worthwhile. An action must be intelligible$_A$ to its agent if it is to be done for a reason, but it seems that we ourselves are suspicious that a token action is intelligible$_A$ to its agent unless we think it can be made intelligible$_A$ to us.

It may seem that there are obvious ways to resist this thesis, but they close up on further examination. Everyone has his fair share of aesthetic distastes, and it is easy to look at a certain style of dress or cut of hair and say that you cannot fathom what people fancy about it. But then again each of us has our own peculiar predilections, and we don't think these may not be intelligible$_A$ even in principle to others these just because we don't expect them to actually share our tastes. Indeed we generally think that we can justify such choices to others in terms of their desirable features, even if we think others wouldn't be wrong in not finding it as engaging or intuitive to grasp.[82]

[82]Examples are legion but perhaps not helpful in establishing the general claim that is my target. Compare Korsgaard (1993, 289): "Someone who says "I *just* want to" is not offering you his reason; he is setting up a bulwark against incomprehension. You may be the problem, or he may feel himself inarticulate: many people do. But listen to the articulate talk about their projects and you hear the familiar voice of humanity, not the voice of alien idiosyncrasies."

The same goes for deep moral differences. We may have difficulty fully understanding a white supremacist's motives, and we may be resistant to exploring such an unattractive mindset, but that hardly means it could not be made clear to us if we tried. It's easy to say what motivated Dylann Roof, who murdered nine black Charleston churchgoers on June 17, 2015. He identified strongly with a certain group of people — in this case the "white race" — and felt it was under attack.[83] What we find objectionable is his racialized conception of social identity, and it strikes us as plain false that the white race is under attack (perhaps not least because 'white race' may fail to refer). But the urge to develop a social identity and to protect a social group in which one feels oneself to belong is something we all find intelligible$_A$.

3.4.2 The transcendental step

The next and final step of the route to public values takes us from claims about what values appear to agents to a claim about the nature of the values themselves, and so for that reason it might be called the 'transcendental step'. Call a substantive good G, which does appear to some actual agents,[84] and of which the following claim is true an *apparently public good*:

18. For any agents S_2, S_1 such that φing appears G to S_2 and tolerating S_2's φing appears good to S_1 out of respect for G: if S_2's appearance of φ's being G is accurate, then so is S_1's appearance of it being good to tolerate S_2's φing out of respect for G.

An appearance to an agent that an action is G is accurate just in case it is G, that is, it meets the standard of practical reason that is G.

The transcendental step is taken by combining the following two claims:

19. Any good that can render an action of at least one agent intelligible$_A$ to herself is an apparently public good.

20. Apparently public goods are public goods.

Now, even if it is not obvious that 19 must be true, it seems clear why it might be true: the good that one agent sees in φing and the good of tolerating that action out of respect for that good are not separate goods. Rather they are the same good considered from different perspectives. Such a

[83] One potential misunderstanding can be laid aside here: one might wonder whether the present account can capture what's genuinely unattractive about a white supremacist's mindset if we *could* appreciate something of his motives if we tried. The response is simply that we can rightly find it unattractive to entertain what we know to be misleading appearances of the good, much as a dieter can be sickened by the thought of walking past a pastry shop. (Indeed, Ch. 1 gives one mechanism for how this can occur.)

For news coverage on Roof, see `http://www.nytimes.com/2015/07/17/us/charleston-shooting-dylann-roof-troubled-past.html?_r=0`.

[84] This stipulation merely ensures that the condition expressed by 18 is not trivially satisfied by goods that happen not to appear to any agents (if such there be).

good is shared among agents just as some goods are shared intertemporally within agents (pp. 49, 69 above).

This same explanation also functions to support claim 20. As mentioned in the previous subsection, making another's action fully intelligible$_A$ to oneself may be cognitively demanding, and enabling this in some agents may require endowing them with improved capacities and new experiences. In working with such idealizations of agents one runs the risk that the standards of practical reason that apply to the ideal agent are importantly different from those that apply to the non-ideal agent. That raises the possibility that even if the idealized agent who makes someone else S's action intelligible$_A$ to herself should tolerate S's action, the agent of whom she is the idealized version, in her non-idealized circumstances, is under no obligation to even consider tolerating S's action. But if the good S sees in acting and the good the idealized agent sees in tolerating it are the *same* good, the same standard of practical reason, then the worry is less acute. Such a good would enjoin an action of a particular agent (in the primary standpoint) and toleration of that action of anyone else (in the critical standpoint), making it a public good.

Now we are in a position to derive the publicity thesis. Take an arbitrary agent S_2's φing for a reason. IC entails that φing was intelligible$_A$ to S_2, and Primary Intelligibility$_A$ entails that S_2 saw some substantive good G in φing. Given the intelligibility of S_2's action, INI entails that it is possible for an arbitrary agent, S_1, to render S_2's φing intelligible$_A$ to herself. Critical Intelligibility$_A$ entails that when S_1 does this, she sees some good in tolerating S_2's φing out of respect for G. Then, given that S_2's φing is intelligible$_A$ to herself, 19 entails that G is an apparently public good, and 20 that G is a public good. Since the argument would go through no matter what G motivated S_2, we have derived the conclusion that any good that can motivate some agent is a public good. That was what was to be shown, the publicity thesis.

Note that a good deal of the heavy lifting in this argument is done by INI and the account of intelligibility$_A$ we have offered. Those premises are what enable any agent to appreciate *the very good* that a given agent sees in acting by seeing toleration of that good as itself good. From there it is quite plausible that such a structure of appearances reflects the structure of that underlying good, i.e. that it does enjoin the respect of other agents just as a public good does.

According to a common definition, the argument I have just given does not technically qualify as a transcendental argument. A transcendental argument attempts to derive propositions about the external world by arguing from the the fact that subjective experience exists and from certain premises about the conditions of its very possibility. Here I have given an argument concerning what *best explains* certain appearance-states, not what makes them possible. My use of 'transcendental' is intended rather to indicate an affinity with a long tradition of arguing from premises about the nature of rational agency or practical thought to a conclusion about the standards of practical

reason.[85] In the remainder of this subsection, I will give a sense of where the present argument fits within that tradition.

There are at least three important ways in which this tradition has been pursued, and individual attempts often incorporate elements of all three to greater or lesser degrees. All share the basic premise that the nature of practical thought or of rational agency imposes some constraints on the standards of practical reason, and they go on to propose different accounts of agency or practical thought. Some appeal to limits on what can be rationally endorsed from a common standpoint that everyone can occupy;[86] others to consistency or non-arbitrariness principles;[87] and others to proposed formal features of thought or action.[88]

The argument in this chapter does not make use of non-trivial consistency or non-arbitrariness arguments. Although it does make use of a special standpoint that we can occupy when we take a third-personal perspective on ourselves, it is not really a *common* perspective, and at any rate the thesis that values are public was not made to hang on special features of this perspective. Instead it makes a very general proposal about what kinds of goods must appear to certain agents, and with what rational force. To that extent it falls in the third category, though unlike previous attempts it does not claim that the constraints are imposed by thought or action *as such*. By far it is closest to Thomas Nagel's argument in *The Possibility of Altruism* (TPA), and so I will contrast it to that argument and show how it can succeed where Nagel's failed.[89]

Nagel's argument depends crucially on two premises. The first is the *completeness* of what can be judged from the impersonal standpoint, the standpoint that "provides a view of the world without giving one's location in it" (101). More precisely, any significant judgment from the personal standpoint — any sense-making judgment from the standpoint typically expressed using indexicals — commits its subject to two further judgments: an impersonal judgment about the same scene and characters that shares its content with the personal judgment, and a self-locating judgment that identifies oneself as one of the characters in the scene (102-104).

[85] Do note that constructivism, which *analyzes* claims about the standards of practical reason in terms of the judgments or concerns of rational agents, is but one form of this tradition.

[86] Nagel (1970); Rawls (1999); Darwall (1983).

[87] Hare (1963); Gewirth (1978); Smith (2015); Korsgaard (1996, 143).

[88] Nagel (1970); Korsgaard (1996, 132-142); Kant, especially in his justification of the universal law formulation of the Categorical Imperative. Another strategy commonly associated with the forgoing three is the attempt to show that agents occupying quite different standpoints converge in their reasons. This is most closely associated with Hobbes, though Smith (2015) also falls in this category. However I have not included it in the list above because generally these arguments attempt to show how agents' reasons are more convergent than we thought, once we look more closely at their circumstances or at the nature of reasons. They do not attempt to draw conclusions about reasons from the nature of thought or action.

[89] All page references to Nagel in this subsection are to Nagel (1970).

One minor respect in which the arguments are similar is that both appeal to a parallel to the normative relations that tie one agent across time before making more controversial claims about normative relations across agents. But the parallel is less strong and less prominent here than in Nagel's argument.

The second is that first-personal judgments about normative reasons for action have what he calls a *motivational content*, which he characterizes as "the *acceptance of a justification* for doing or wanting something".[90] He further asserts that "[f]irst person judgments about reasons are inherently relevant to decisions about what to do, and they provide the basis for justification and criticism of *action*, and *desire* — not just of judgments *about* action and desire" (110). For me to judge that the gouty toes of the man to my left are a reason for me not to step on them is already for me to take up a practical stance, capable of producing action, in favor of my not stepping on the toes; it is not *merely* to note a curious fact about my relation to the toes. But according to Nagel's first premise, this first-personal judgment commits me to an impersonal judgment with the *same content* as the first, including its motivational content. I am committed to a claim such as "The gouty toes of the man to P.B.'s left are a reason for P.B. not to step on them", and furthermore to being in *favor* of P.B.'s not stepping on them. The result seems to be that I cannot make judgments about what *I* have reason to do without committing myself to being in favor of *people* doing what they have reason to do. This supports a public conception of reasons since reasons' being public would vindicate these commitments.

The problem with Nagel's argument is in his implicit move from (what he calls) motivational content's being an aspect of first-personal judgments about reasons, and even an *essential* aspect (113), to its being part of the *content* of those judgments. In talking of a judgment we could be talking either about a certain mental state — an act of judging — or about *what* is thereby judged to be the case, its content. There is indeed a traditional understanding of the content of judgments according to which it is either a proposition or it shares its truth-conditions with a proposition, which is a perspectiveless entity. But Nagelian motivational content does not look to be content in *that* sense. It is an aspect of the mental state by which we apprehend such content: it is an *acceptance* of its relevance for one's own action. And the properties of the mental states by which two differently-situated people judge a certain proposition to be true may rationally differ to a significant degree.

For illustration: I, standing in Ann Arbor, judge that *the Limmat River is far away*. Nina in Zurich agrees: the distance between the Limmat River and Ann Arbor is quite large. But her judgment (a mental state) does not involve a tokening of the concept *far* as mine does, or at the very least it differs significantly in the way it deploys the concept. Nor is it a commitment of my judgment that Nina in agreeing with me must use the concept *far* in the way I do by applying it to what I apply it to. Thus Nagel's argument gives us no reason to think that the element of acceptance involved in my (act of) judging that I have a reason not to step on the man's toes commits me to thinking that any impersonal assessment of P.B.'s reasons must contain that same acceptance, if it is to judge what I judge.

[90] P. 109, emphasis in original; see also 65.

Now, there is a sense in which a normative judgment from the impersonal standpoint to the effect that someone, Hortense, has a reason to φ does commit one to a kind of acceptance or favoring, but it is only a commitment to favor acting that way *if one were Hortense*.[91] It is consistent with this commitment that one *actually* not favor Hortense's doing what one judges she has a reason to do, given one believes that one is not Hortense.[92] So it seems that we cannot find in Nagel's argument support for the publicity of reasons.

In this chapter there is no attempt to draw conclusions from the thesis that what can be judged from the impersonal point of view exhausts the truths that can be judged from any point of view. The argument I have developed here is able to draw a similar conclusion only because it begins with a stronger assumption, the intelligibility constraint. Given that, the linchpin of the present account is the connection between an agent's making an action intelligible$_A$ to himself — whether that is his own action or someone else's — and his having an appearance with the force to rationally support an action of his own, especially as expressed in Critical Intelligibility$_A$. Rational force does indeed play a role analogous to Nagel's motivational content since they both involve an element of acceptance of an action, but it is not subject to the equivocation above. For the rational force of an appearance is a matter of what one may be rationally entitled to infer from an appearance with a given content, and it is not part of the content itself. The role of rational force in the present argument is also more limited, since it only captures the way in which making an action intelligible$_A$ is not a practically neutral stance towards that action but instead always involves an element of favoring it.

In a nutshell: Where Nagel held that endorsement of someone's action was a commitment of judging that they had a reason to do what they did, the present account holds that making someone's action intelligible$_A$ to oneself — really *seeing the point* of what they did — requires seeing the good of tolerating their action out of respect for the value that they themselves see in it. It then holds, not that we must in fact make others' actions intelligible$_A$, nor that we are rationally committed to doing so, but that we must be *able* to make them intelligible$_A$ to us (INI). From there it argues that the structure of appearances this requirement generates — i.e. that a good that one agent finds in acting is one that is apt to appear to others to be deserving of toleration — is most likely a reflection of the structure of the standards of practical reason (theses 19-20).

[91] Expressivists may hold that a judgment that Hortense has reason to φ may commit one to (actually) favoring Hortense 's φing, but they can nevertheless get a handle on the kind of judgment I have in mind here: it is a *conditional* normative judgment, a plan about what to do *in the case that I am Hortense.*

[92] Of course the point I am making is not new; see Williams (1985, 61).

3.4.3 What publicity entails

We have throughout been working with very abstract conceptions of the standards of practical reason and of public standards of practical reasons, so it would be good to pause before closing this chapter to consider what the publicity thesis does and does not entail. I will also catalog a few of the most promising ways to resist this implication of the intelligibility requirement and the costs of doing so.

Kieran Setiya (2007, 14) calls "ethical rationalism" the thesis that the standards of practical reason can be derived, at least in outline, from the nature of agency or practical thought. In that sense of the term, this chapter has indeed been friendly to rationalism — depending on what an outline of the standards of practical reason is and what suffices as an account of the nature of agency or practical thought. At most the strategy can be used to derive a formal constraint on the standards of practical reason. It does not, even in outline, give us any insight into what the substantive standards of practical reason are. It also is not offered as an implication of the essence of practical thought or agency, since it seems to me that a crucial step in the derivation, INI, could only be justified by appeal to the distinctive capacities of human agents. (What Setiya calls 'rationalism' is closer to what we will call 'constitutivism' in the next chapter. The main differences between the two are, first, that constitutivists may have more limited ambitions than the rationalists, aiming as they sometimes do to derive only a few standards of practical reason. And second, as the name suggests, constitutivists hold that it is *constitutive* of being an agent that these standards apply to agents. It is this second feature that makes the present chapter less helpful to constitutivists.)

The publicity thesis entails that practically speaking, there is one domain of values within which we all live and which all touch us to some extent. I cannot correctly regard someone's projects as having *no* bearing on what I should do except insofar as they further or hinder my own. To that extent the publicity thesis is inconsistent with current formulations of desire-based theories of reasons, many decision-theoretic principles of choice, and universal egoism, all of which allow (or even require) the rejection of that thesis. For instance, absent special modifications, these theories entail that an ideally coherent misanthrope and cinema-hater has no reason not to burn rare movie reels if he can get away with it; in fact it may be a requirement of practical reason that he burn them. Given that other people find value in watching those films, the publicity thesis entails that it cannot be true that for him there is *nothing* to be said against burning, since burning would prevent the realization of the aesthetic value of others' watching the film.

Still, the publicity thesis is relatively weak in other respects. We have said nothing about the normative strength of the injunction on third parties to tolerate action that public goods create; all we have required is that the standard specifies that tolerating the action is to be supported, to some extent, out of respect for the good that its agent sees in acting. The most natural way of

interpreting this is that the agent in the critical standpoint has a *reason* to tolerate the agent's acting, but stronger and weaker formulations are possible: one could hold that third parties are *required* to tolerate others' valuable pursuits, or merely that they are *pro tanto* permitted to tolerate them. (However, it must be said that a requirement to tolerate others' valuable pursuits seems rather implausible given that in driving, consuming rival goods*, etc., I can hardly avoid interfering with others. Hence it seems better to interpret the standard as specifying that third parties have some reason to tolerate others' valuable pursuits.)

Furthermore, as we mentioned above, tolerating an action out of respect for the good it produces may not itself be good if the agent is acting under an illusory appearance that their action will produce that good. So the thesis is also weak in the sense that it does not entail that *every* time an agent acts for a reason, third parties have a reason (or requirement, etc.) to tolerate the action.[93]

Finally, the thesis that we have reason to tolerate others' pursuits may look to be a moralistic one, but it must be remembered that it is a *formal* constraint on standards of practical reason. If there are inherently immoral but nonetheless genuine values, the thesis applies equally well to them. Nor is the publicity of values best thought of as sufficient to explain our moral reasons not to interfere with other agents, as R. Jay Wallace (2009) points out. The main reason not to board up one's neighbor's house and lock her inside is not that I would prevent her from doing her errands, but that the value of her autonomy demands my respect.

The practical dynamics of competitive games may seem to cause difficulties for the thesis that all goods are public. Suppose I am an especially empathetic goalie in a soccer match, and as the striker of the opposing team winds up for a kick aimed at my net, I come to understand just how much winning the game means to her. I find her attempt to win intelligible$_A$. Does this mean that I have cottoned on to a reason to *let* her win? It seems not.

One way of accommodating such cases is to note that the nature of toleration or non-interference is structured by the nature of he activity tolerated, and then to hold that tolerating the moves of an opposing player in a game we are both willingly playing is consistent with attempts to prevent them from winning. A good weight trainer does not help a client by lightening loads but by adding resistance, after all; perhaps the suggestion is that by swatting away the striker's attempt I am helping her play a good game of soccer. But so far this response fails to address the *specific* good that makes her action intelligible$_A$ to me. She wants *to win*, and perhaps to win even against an opponent who is not trying particularly hard. It still seems wrong to say that I have some reason to play leniently out of respect for the good of her winning. (We will reconsider this response to the problem in a moment.)

[93] For the sake of averting a potential misunderstanding, note as well that the argument for the publicity of values is consistent with agents having to particular reason to render other agents' actions intelligible$_A$ to them. The derivation above depends on the *possibility* of agents finding others' actions intelligible$_A$, not on the supposition that agents generally do or should make them intelligible$_A$.

Another response is to say that the good of playing the game itself provides an (exclusionary) reason to players not to consider reasons in favor of tolerating their opponents' attempts to win (Nagel 1970, 131-132), or that my intention to play the game provides reasons not to take certain considerations (such as my opponent's feelings) to be relevant to my game play (Scanlon 1998, 51ff.). One potential problem with this line of response is that it allows or even takes for granted that the good of the opponents' winning *is* a reason for me to let them win. It just isn't one I should take into account, according to the response. But some might object that I have *no* reason at all to allow a goal in soccer.

I myself am not entirely sure I share this intuition, but I wish to note that our loose characterization of toleration allows us to make principled distinctions concerning what toleration requires in particular contexts. It may turn out that in the context of a game the kind of toleration that is required by the good of one team's winning is not that one allow them to win but that one be willing to accept the benefits and burdens of their winning: proper respect for the good of one's opponents' win entails that one should let them have the trophy. I do not mean to endorse one of these responses at the exclusion of the other here; I wish merely to note two ways in which public values are plausibly consistent with robust competition.

One reason to favor the publicity thesis is that it address the inertness worry. As we have just seen, we *can* learn something about the nature of practical reason just by establishing GG via the intelligibility motivation. Some GG theorists, however, may want to resist starting down the road to a fully public conception of reasons, and I will close out this subsection by pointing out a few promising places in the argument above where they might turn back along with the costs of doing so.

Perhaps the most obvious claim to reject is the account of finding another agent's action intelligible$_A$, Critical Intelligibility$_A$, according to which two conditions are each necessary and jointly sufficient for finding another's action intelligible$_A$. The first (16) seems innocuous: if GG is right that an action is intelligible$_A$ to its agent in virtue of seeming good to that agent, surely in order to make your action intelligible$_A$ to me I must understand what good appears to you and that it so appears. But, one might object, why should making your action intelligible$_A$ to me — which sounds very much like a cognitive feat — require furthermore *appreciating* the good of your action in a way that inclines me to tolerate it, as 17 requires?

The difficulty with this response is that it creates a problematic bifurcation in the GG theorist's account of intelligibility$_A$. As argued in §3.3.3, intelligible$_A$ appearances of the good must have the force to rationally support action. Otherwise, it seems, they cannot play the role in rationalizing explanations which GG requires them to play. But if an action, even another agent's action, could be intelligible$_A$ to me in way that does *not* require an appearance with the force to rationally support action, then it is not clear why I could not use the same method to make my own actions

intelligible$_A$ to myself. Perhaps just as George sees it as a merely curious feature of Ayn's short-changing him that it is aimed at her own good, which he (let us suppose) rationally does not see as having any practical significance for him, I could make my own actions "intelligible$_A$" by making a similar judgment that they are aimed at some good or other, while rationally not taking this fact to have any practical significance for myself. But this would leave us without a good account of intelligibility$_A$. For, making one's own action intelligible$_A$, seeing the point in it, necessarily involves seeing it as practically significant. We would furthermore deprive ourselves of explanation of how seeing an action as good can provide one's reason for acting, which GG ultimately requires. So those who are doubtful of Critical Intelligibility$_A$ would do best to also reject the project of explaining intelligibility$_A$ in terms of the guise of the good, or to reject GG itself.

INI also looks vulnerable — or better, it seems innocuous if Critical Intelligibility$_A$ must be rejected for the reason just given, but not obviously true if it must not. As I indicated above, I am not inclined to defend INI as a necessary truth, nor even as an exceptionless truth about actual human agents. However, I think it is quite plausible that we are generally capable of appreciating the point in nearly any agent's intentional action. And even if INI is merely a generic truth about human action, in conjunction with the rest of the argument of this chapter it still would deliver a proportionally strong result about the degree to which the good is shared.

Finally, some might characterize the transcendental step in §3.4.2 as a *leap*. Admittedly, it received little defense. But even if it covers rocky terrain quickly, it is quite a reasonable explanation of the structure of appearances that the argument to that point had developed, and to that extent it functions to shift the burden of argument to those who would deny it. And even if it fails, it leaves in place something still rather interesting, which is that in making others' actions intelligible$_A$ to us we must to some extent respond just as if all values were public.

3.5 Conclusion

We began with a problem for the Guise of the Good: lest GG be too weak or theoretically inert, we needed to find some way of restricting the notion of good at GG's core. A conception of the good in terms of the standards of practical reason solved the weakness problem. However it does not solve the inertness problem since the notion of the standards of practical reason is itself a rather formal one about which there is much dispute. If GG is to make a contribution to the theory of practical reason, we need to argue for further restrictions on the notion of good.

The very name of the thesis, the Guise of the *Good*, and the way it has occasionally been interpreted by its critics led us to look for a solution to this problem in terms of *shared* standards of practical reason. These standards favor not only one agent's action but other agents' tolerating that action. And it turns out, I argued, that the requirement that actions must be intelligible$_A$ to their

agent allows us to see a route from GG to a conception of the good on which the good is shared with *everyone*, making it public. The publicity thesis does indeed constitute an important move in theorizing about practical reason, so the route just traced is one way of solving the inertness problem.

Before we move on to Chapter 4, it would be good to make two notes. First, the value of the argument I have given in this chapter does not depend on the truth of IC. It is enough to point out that it is a popular motivation for GG and that there are strong reasons to think it leads to what is currently a controversial conception of action and value. It is open to those who reject this conception to use the foregoing argument to reject IC or GG. Second, I have not given any indication so far as to *how* it may be that agents in the critical perspective making sense of others' actions can see their actions as good in the required sense. In the next chapter we will made some headway on this question by considering how affective states, as representations of the good, can play the role of appearances of the good.[94]

[94]This paper was presented at a workshop on the Guise of the Good in Paris in May 2016, and it benefited from the audience's comments. Thanks also to Peter Railton, Sarah Buss, Uriah Kriegel, and Benjamin Wald for comments that greatly improved the paper.

CHAPTER 4

Affective Content and the Guise of the Good

4.1 Introduction

Recall the formulation of the Guise of the Good from Chapter 3:

GG Human action or motivation to act, of some special kind or another, is only possible insofar as the agent performs or is motivated to perform the act because of the *good* she sees in so acting.

As before, my primary concern is with action for a reason. In this chapter I raise two related problems for the most popular model of GG, what I call *the Standard View*:

21. Necessarily, if A φs for a reason, then A φs because she desires to φ; and

22. A desire to φ essentially represents φing as good.

Here I take "essentially" to imply "necessarily". It is important to the standard view that its two constituent theses are claims about practical thought and activity *as such*, so that the thesis expresses a metaphysical necessity. The same carries over to the version of GG that the Standard View of desire entails, which is correspondingly strong.

As I explain below, I take it that a significant upshot of GG as implemented by this model, and thus a major motivation for the model itself, is a strong kind of internalism about practical normativity, namely that, necessarily, the standards that guide one in acting are standards of practical reason. In this chapter, I argue that the Standard View and the strong form of internalism to which it leads are false. I then propose a novel affect-based theory of action for a reason that vindicates GG, and I show how it satisfactorily avoids both of these issues while still preserving a more limited form of internalism. Along the way I also argue that a standard understanding of the *guise* of GG, what I call the "attitude/aim" strategy, in its best version collapses to the understanding that I favor, the content-based strategy.

As such this chapter is divided into two parts. The first (§§4.2-4.4) presents the negative case. After some preliminary clarification of GG (§4.2), I argue *contra* thesis 22 that desires are not

essentially evaluations (§4.3). I then offer an argument against the conjunction of 21 and 22 (§4.4). Neither of these arguments is intended to offer incontrovertible proof that the Standard View is false, and both rely upon intuitions that supporters of the Standard View could coherently reject (though, of course, I believe that they should not). The main role of these arguments is to set up the positive view to follow in the remainder of the chapter: I aim to show that even if one is inclined to reject the Standard View for the reasons I offer, one need not reject GG wholesale.

A note about nomenclature before moving on. As in Chapter 3, 'good' has the schematic function of ranging over standards of practical reason. Talk of the good is intended to be neutral over talk about what one has *reason* to do, what one *ought* to do, or what is *good** to do (where the asterisk indicates the word takes its normal, evaluative sense). If an agent sees some good in acting then she has a seeming — an appearance — of there being some good in so acting, and so "appearance" can take on a similarly schematic role of ranging over whatever states in virtue of which an action seems good to an agent. However, since most GG theorists have concentrated on the thesis that *desires* are the relevant appearances of the good, and indeed many have interpreted it as *primarily* about the nature of desire, the reader should recall during the discussion below that I take the view to be broader.

4.2 Guises of the Guise of the Good

GG is a family of resembling views rather than a unified doctrine, as some commentators have noted,[1] so it is all the more surprising that the literature of the past 25 years has focused on the Standard View. But then again it is common for a view to be shaped by one of its critics, and I believe that is what has happened here.

4.2.1 Background

In an influential critique of GG, David Velleman (1992a) first clears the ground of some unstable versions of the thesis before setting his sights on one he regards as worthy of a sustained assault. One version he rejects, the *content-based view*, "incorporate[s] the valence of desire into its content, by describing a desire, not as a favorable attitude toward the representation of some outcome, but rather as an attitude toward a favorable representation of the outcome" (*ibid.*, 103).[2] For instance, consider the Davidsonian (1963) theory of action for a reason as action φ caused by a desire that P and a belief that by φ-ing one will bring about that P. One subtype of the view

[1] See e.g. Orsi (2015).

[2] Proponents of this view include Goldman (1970, 94), Davidson (1978) (on one interpretation). Brewer (2009, 26, 46-47) espouses a variant on which, at the deepest level, one wants certain *good objects* and *activities*.

makes the further claim that a desire that *P* may also be construed as a representation of a different proposition, roughly of the form *it would be good that* P.

This subtype has the distinct advantage of preserving the parallel between theoretical reasoning and practical that we mentioned in Chapter 3, at least in simple cases of each. The beliefs that *R* and that if *R* then *Q* altogether represent the world as such as to entail the truth of *Q*, which in turn makes belief in *Q* correct. The desire-representation of its being good if *P*, together with the belief that φing will bring it about that *P*, represent the world as such that φing would lead to something good. Since *leads to something good* is plausibly itself a good-making feature, and since it is (*pro tanto*) correct to perform good actions, in each case the world is represented as such as would make the "inferred" attitude or action as would make it correct.[3]

However, Velleman goes on to argue that an attitude towards a proposition requires the possession of the concepts used to express it (*ibid.*, 104). If, moreover, *any* desire is also an attitude toward some proposition concerning what's good as such, and if small children and animals do not possess the concept of the good, then this implies that they do not have desires — and indeed if the Davidsonian theory is right, that they do not act for a reason. It is reasoning of this sort that is responsible for the low regard in which content-based theories of GG are held today.

A nearby position that preserves the connection with practical inference requires only that the agent's desire be *expressible* in terms of an evaluative proposition, or that a belief that there is something good in the action be properly *attributable* to the agent. Donald Davidson (1978, 86) is representative of the former while the latter is one way of reading Joseph Raz. Raz is quoted in Stocker (2008, 125) as writing in a lecture transcript that he does not

> attribute a kind of high order reflectiveness to people: for example that they not only think of their actions as pleasurable or thrilling or beneficial to X or Y, etc. but also think of them as good in virtue of possessing those properties. All I ever claimed is that people act for considerations which *we classify* as a belief in the possession of a good making property.[4]

But as Velleman (1992a, 104-105; 109-110) notes, it is unclear whether these should count as versions of the Guise of the Good since they render doubtful that the agent must have "mental access" to a justification for acting (*ibid.*, 105). The question is not whether *we* can see the agent as seeing their action as good, nor whether the agent is behaving just *as if* they saw it as good, but whether they *themselves really see* it as good. It could be the the proper expression of my thirst is

[3] This is simply a sketch of one advantage the theory *might* have, if it could be further developed. Note, for instance, that as of yet we have not earned the right to speak of the action φ as presented to the agent as good *as such*, as GG requires.

[4] My emphasis. There is another way of reading the second sentence here on which the view is that action for a reason requires a *de re* belief in the action's possession of a good-making property. But this view is incredibly implausible; surely we can make mistakes about what properties are good-making and act for a reason nonetheless. See also Raz (2010, 114).

"it is good that I drink", but it is unclear what to make of my own intentional relation to the good if I never think to express it. Of course, we can apply the same Vellemanian point to Raz. There is a sense in which we can properly attribute an attitude to an agent even in conditions where they do not literally possess the attitude, as when we are engaging in the fiction that our pets have a level of psychological sophistication they do not literally possess. In such cases we can use the fiction to explain their behavior since the pet's actual attitudes no doubt approximate at some level the ones we attribute to them. But Fluffy is not literally *resentful* of us for having left her at home all day. GG, however, is interested in the literal truth about our state of mind.

Velleman then settles on an interpretation of GG in which the good figures in some feature of an attitude *apart* from its content: on his view, in the constitutive aim of desire.[5] On this construal desire is thought to involve a particular way of regarding its content, a regarding it *as good*. Just as a belief that P regards P as true with the aim of getting it right as to whether P is true, desire is thought to regard its object as good with the aim of getting things right as to whether it really is (Velleman 1992a, 116). This version of the standard view of GG, which I will call the "attitude/aim" strategy, holds that although (a) the good need not figure into the content of an appearance of the good, (b) there are appearances of the good as required by GG, and they are such in virtue of the fact that the goodness of their content is an intrinsic standard of their correctness, much as the intrinsic standard of belief is the truth of its propositional content. That is, a goodness-appearance is correct only if its content is good, and this standard of correctness applies to it in virtue of its intrinsic properties.[6] (Note that we will find reason to modify this formulation slightly below in §4.2.2.)

Versions of this strategy are now predominant.[7] Tenenbaum (2007, 2008) argues that to desire X is for X to appear good to one, and to intend that X is to hold X to be good. Schafer (2013)

[5]Note that the aim here is one *of desire*, not necessarily one of the agent. Indeed, talk of desires or attitudes of the agent as having aims is best understood as metaphorical.

[6]The fact that the standard of correctness must be intrinsic to the appearance goes some way towards explaining why GG proponents and critics are often comfortable in transitioning from talk of desire's being an *appearance* of the good to an agent to talk of desire's *aiming* at the good. One might at first think that there is a gap between a state's representing the good and its aiming at the good. As we will see in the next subsection, the best way of closing the gap is to interpret the aiming as aiming to get things *right* about what's good. Note as well that a thing's having an aim depends to a considerable extent on its actual, intrinsic dispositions: I can't be said to aim at what I am not disposed to try to get, nor to plan to get, nor to even think about getting. So, it seems, a state's aiming to get things right about what is good is to a considerable extent grounded in its intrinsic properties. If furthermore a desire's status as an appearance of the good, and thus as a representation of the good to which a standard of correctness applies, amounts to its aiming in this way, then the application of that standard of correctness is also grounded to a considerable extent in those intrinsic properties.

[7]This is not to deny that there adherents of other versions. Aside from Raz's view, Boyle & Lavin (2010) have a GG theory in which the good plays an essential role in the explanations rational agents are thought to proffer for their actions; it is unclear to what extent this view can be assimilated to an attitude/aim theory. Buss (1999, 413) holds that action for a reason must be motivated by a belief with evaluative content and Hawkins (2008) has a content-based view which is in many respects similar to the one I will go on to offer.

holds that the good figures into the rational force of desire, where the rational force of an attitude determines what further mental states it makes rational.[8] On his account, a desire to φ presents φ as something that ought to be done (*ibid.,* 277). Saemi (2015) holds that the good is part of the form of intentions and tryings, not their content, and gives a teleological account of that form. [9]

Velleman ultimately rejects the attitude/aim theory because he thinks that attitudes cannot take objects whose adoption would transparently undermine their constitutive aim. Beliefs cannot take contents that present themselves as false and so undermine belief's aim of believing truly: I cannot really believe *I am five inches taller than I actually am*. But it appears that one *can* desire what's bad as such. The young Augustine famously took pleasure not only in thieving a pear, but in the sinfulness of it.[10] Unless we radically and implausibly reinterpret motivations like Augustine's, it seems the good isn't the constitutive aim of desire after all.

Unfortunately for Velleman, what may explain the impossibility of my believing *that I am five inches taller than I actually am* may be the peculiarity of belief that it is not possible to have a belief with transparently inconsistent content.[11] For it is easy to find an example of a possible attitude-content pair where the latter transparently defeats the aim of the former. It is quite plausible that we can regard a proposition as not only true but as *objectively* true: we aim to get at what's true independent of any perspective. But an overzealous and insufficiently reflective undergraduate might take this kind of attitude towards an unrestricted relativism about truth. The fact that she does this can be evidenced by the fact that she spends a week trying to convince everyone she knows of its truth, views those who deny it as *disagreeing* with her and *wrong*, etc. There is some incoherence in her state of mind, but there is nothing impossible about it. The same may very well go for wanting the bad as such.

Note as well a second strategy for dealing with this problem available to some GG theorists. Most of the notions of good with which we are dealing, such as *good** and *reason*, are graded or weighted concepts. One event may be overall bad* yet good* to some degree, or one could have some reason for an action without having sufficient reason for it. It is entirely consistent with GG that the constitutive aim of desire is only that its content be *good to some degree*. Desiring that which one on reflection takes to be overall bad need not defeat this weaker aim, for on this view the desire only presents it as to some degree good, which it very well may be. Moreover, this view

[8] For instance, perceiving that *P* and imagining *P* share the same content but differ in rational force, for the former and not the latter makes belief that *P* rational. For more on the notion of rational force see Schafer *op. cit.* pp. 270-273 and Chapter 3 above.

[9] Note that on some views practical reason is held to aim at the good in a way analogous to belief's aiming at truth even if the theory is silent on or even skeptical of actions' necessarily *appearing* good to their agents, and are thus not GG theories. See Velleman (1996) on full-blooded action and Wedgwood (2003, 205).

[10] *Confessions*, Book II.

[11] Do note that we must be careful in specifying the content of the belief. Since it is possible that I could have been five inches taller than I actually am, and hence not inconsistent, the inconsistency is apparent only when we take the content of the belief to be *that actually I am five inches taller than I actually am*.

is less vulnerable to an iteration of Velleman's worry. It is one thing to desire the bad as such and quite another to desire what is to no degree good as such, and it seems that GG proponents are in a stronger position to defend the claim that it is impossible to desire what is to no degree good as such.

So, the attitude/aim theory survives Velleman's critique. Yet care must be taken in developing the attitude/aim strategy if it is to succeed. In the next subsection I aim to get a bit clearer on the ways in which this version of GG can be said to unify practical and theoretical reason, thereby clarifying the thesis itself.

4.2.2 Belief : true :: desire : good

Only propositions and certain sentences and mental states are truth-apt; sentences inherit their truth-values from the proposition they express, and the truth-apt mental states from their propositional contents. The idea that "belief aims at truth" is then, at a first pass, the idea that belief or the process by which it is formed is by nature governed normatively and descriptively by norms that bear on the truth of the contents of those beliefs, and thus on the truth of the beliefs themselves. Beliefs tend to go out of existence, and *correctly* go out of existence, when they are confronted with evidence that their propositional content is false, and tend to and correctly come into existence in light of evidence of the truth of their content. "Belief aims at truth" means not that beliefs aim to *produce truths* but that beliefs aim to *believe what's true*, or to *be true beliefs*. The aim is thus self-directed. Hence there is an interesting question whether these regulative norms of belief are such that beliefs merely aim to be true, or whether they aim more precisely to be knowledge.[12] Similar remarks apply to the aim of theoretical reasoning generally, although insofar as it involves free use of supposition and imagination the connection between theoretical reason and the truth is more flexible than that between belief and the truth.

However there must be quite a different relation between appearances of the good, as construed by the attitude/aim theory, and the good. Generally speaking, one *proposition* isn't better or more normative than any other — though of course the same is not true of the actions, states of affairs, and (perhaps) the principles those propositions are concerned with. It's important to be clear on the contrast. Truth is a property of representational entities, if it is a property at all, and it is the same for beliefs as for propositions.[13] But when we consider the properties associated with the standards of practical reason — being good*, being supported by reasons, being something one ought to do — we immediately see that these are not generally properties of representational

[12]For a defense of the latter claim, see Smithies (2012).

[13]For the sake of simplifying prose I will write throughout of the property of truth, but I do not mean to take a stand on the nature of truth. Deflationary translations can be applied to all such phrases but at the cost of clunkiness.

entities.[14] For instance, the goodness of ice cream consists in its tastiness, but the goodness of an appearance of ice cream does not consist in the *appearance's* tastiness. Perhaps the thought animating the attitude/aim theory is that the goodness of an appearance of ice cream consists in the ice cream's tastiness; but how is that a goodness of *the appearance?* If there is a way that appearances can be good it seems it would be the good of faithful or accurate representation, not tastiness.

In short, the idea that actions appear good to their agents seems to be in tension with the analogy to truth. Actions are good in all sorts of ways: they lead to tasty ice cream, are fulfillments of duties, etc. But it makes no sense to say that desires present their contents as good in the way that beliefs present their contents as true, for propositions are neither good nor bad. And to say that appearances aim to be good appearances loses sight of all the different goods that make our actions intelligible$_A$, collapsing them into the good of faithful representation. The analogy with truth forces upon us what we can call a *wrong kind of goodness* (WKG) problem.

A cousin of the WKG problem is the *non-inheritance* problem. Though beliefs inherit their truth-values from that of their propositional contents, appearances do not generally inherit their goodness from the goodness of the things, states of affairs, etc. that their contents represent. We saw one example of this in Chapter 2, in the case of the politician's adviser. The polls made the politician's situation seem bad* to the advisor, but he knew the appearance was misleading. In this case we would not say that the adviser's appearance state was itself bad*. It moreover seems we should say the same even if the appearance given by the polls were not misleading and the politician's candidacy really were in bad* shape, for this appearance is not a messenger that needs shooting. Or suppose that the variety of goodness that interests us is *being supported by reasons*. I see a man frantically twirling his hand in the air in front of me on the street, and I suddenly get the impression that I've reason to run from him. And in fact I do have reason, since as it turns out he is thereby signaling to hidden confederates to start the attack. But I am completely unaware of this; I react simply because I am irrationally afraid of men making hand gestures. I am aware that I have this phobia as well, and so it is highly intuitive that I have no reason to think I have reason to run. Hence the instantiation of the property *being supported by reasons* does not transmit from the object of the appearance state, the action of running, to the appearance state itself.

One response to the WKG problem would be to hold that the many forms of the good with which GG is concerned all qualify actions primarily and that actions are always partly representational. It is common to hold that in acting intentionally or for a reason I must know, or at least believe, that I am performing that action. One could expand one's notion of an action to include this representation as a part, and thus deny that the goodness of GG is not a property of represen-

[14] Anscombe (1963, 76) briefly makes the same point.

tational entities.[15] But what needs to be established is that *actions* are representational, not that actions contain representational parts. Moreover, the latter does not entail the former. A car is full of representational parts — gauges and blinkers — but cars are not themselves representational entities. And at any rate, is hard to see how this maneuver would address the non-inheritance problem.

Moreover, there is a simpler solution at hand, and one that even better respects GG's theoretical aim. Instead of seeing belief and desire (more broadly, appearances of of the good) as each constitutively governed by distinct kinds of properties, the true and the good respectively, we could hold that each has the self-directed aim to *be correct*, where correctness for belief is truth and for desire, goodness. Given this a parallel can be quickly reestablished. A belief that X is correct iff it is true; a belief is true iff its propositional content is true; and it is true that X iff X. That gives:

23. A belief that X is correct iff X.

And according to the attitude/aim theory, the correctness of an appearance of the good is not given by the truth of its content but by the goodness of the action under consideration:

24. An appearance of φing as good is correct iff φing is good.[16]

What this means is that we should not understand appearances of the good to present their contents as good in just the same way that beliefs present their contents as true. The goodness of the content — if that notion even makes sense — is irrelevant. Instead each attitude represents its respective correctness conditions as satisfied, where the respective conditions are given by 23 and 24. Thus we have arrived at our improved version of the attitude/aim theory as initially characterized on p. 86: appearances of the good and beliefs are best thought of as intrinsically governed by their respective correctness conditions, not by properties of their contents.[17]

Of course, the aim of practical reasoning is not exhausted by a condition like 24; the point of practical reasoning is not to reflect the normative landscape so much as to effect or respect it. But neither is 23 thought to give the full aim of theoretical reasoning. A person who regularly

[15] Thanks to Sergio Tenenbaum (personal communication) for this response.

[16] Of course, many who hold GG also hold that appearances of the good (desires, generally) are *propositional* attitudes. We could easily modify 24 to accommodate this, at the cost of a slightly uglier condition: a desire that P is correct iff it would be good if P. However, since GG is a thesis about *practical* reason, it is reasonable to suppose that the content of any desire to which it applies must include an action as its subject, so it is unlikely that 24 is misleading. Furthermore, I argue in §4.7 that the right-hand side of 24 is the content of the appearance of the good, and it is propositional. So there is no reason for propositionalists to be alarmed.

[17] The reader may wonder how the claim in this chapter that appearances of the good are governed by correctness conditions coheres with the argument in §3.3.3 that in the case of pretense an appearance of the good does not have rational force with respect to belief. If a state is governed by the correctness condition X, isn't it in a position to license the belief that X? The answer is that it may not be, and pretense is a case in which it is not. For, the pretense that X is governed not by the truth of X but by the truth of X *in the fiction.*

affirms the consequent could not be said to reason well even if he only draws trivially true, and thus correct, conclusions thereby.

Before moving on, we should note an additional way of restoring the analogy. One could maintain that appearances of the good and beliefs are each intrinsically governed by properties of their content but reinterpret slightly the kind of property thought to govern desire. One could, that is, hold that whereas beliefs are governed by whether their contents are true, desires are governed by whether their contents are good to *make true*.[18]

However there is a certain artificiality in the proposal, and that I think is reason enough to prefer my solution. To see the problem, suppose that the good is value*. Things, states of affairs, people, and actions are valuable*, and it's the value* of these that appearances present. It is not the value* of making a certain representational entity, a proposition, instantiate the property of truth. If, *per impossibile*, the God of Philosophy decided to delay making the proposition *John gave his mother a Mother's Day present in 2016* true until an hour after John does give his mother a Mother's Day present in 2016, in that hour John still would have successfully produced exactly the good that appeared to him. Hence it seems that appearances are better thought of as appearances of what would be good, not what would be good to make true; we are concerned with reality, not the match between representations and reality.[19] Note as well that this proposal also suffers from the WKG problem: what is the extra value of *making it true* that John gave his mother a Mother's Day present in 2016 over and above the value of John's giving his mother a Mother's Day present? Once the question is put it is no longer clear what intrinsic value there is in making a proposition true.

4.3 Against evaluative desire

That clarification made, we press on to an objection to claim 22, and thus to the Standard View of GG. That claim makes the Standard View a version of *evaluativism*, the thesis desires are essentially evaluative.[20] In this section I argue that even if desires are evaluative, they are not evaluative states *qua* desires and are thus not essentially evaluative.

One important dimension along which theories of desire vary is *complexity*, for instance in how many psychological states and relations a theory posits. Among the simplest is Brentano's (1889) view, according to which desire is a psychologically primitive (albeit gradable) attitude. Slightly more complex are the popular simple dispositionalist or functionalist theories according

[18]This move takes inspiration from Shah and Velleman's (2005) distinction between cognitive attitudes, which treat their contents as true, and conative ones, which treat their contents as to be made true.

[19]I thank Sarah Buss for this phrase.

[20]This is not to be confused with *evaluativism*, the view that pains are representations as bodily events as bad. The nomenclature is regrettable.

to which a desire to φ is a disposition (or its categorical basis) to φ in circumstances where (a) one believes one can φ and (b) has no stronger, contrary dispositions, or is a state that has the function of causing one to φ in such circumstances.[21] More complicated are holistic theories on which to desire is to exhibit enough of a set of characteristic states and dispositions: a person who desires a sandwich tends to have her attention drawn to opportunities for obtaining a sandwich, think favorably of getting a sandwich, be pleased by eating one, and make plans to obtain one, among other things.[22] Learning-based theories further require that such dispositions are integrated into a feedback-feedforward system that constitutes the desired object as a reward and which is itself the desire.[23]

In light of this diversity of views of desire it is important to note two things. First, all relevant parties to the dispute agree that desires essentially have satisfaction conditions and that there is a tight connection between a desire's satisfaction conditions and the motivation it involves. Desire is said to have a world-to-mind direction of fit (Anscombe 1963) or is satisfied only if its object exists or comes to pass (Searle 1983). It is difficult to analyze the notion of a direction of fit or to say what separates desire's satisfaction conditions from belief's correctness conditions, but it is generally accepted that when one desires that P, and yet $\neg P$, it is in some sense *the world* that "must" change — or perhaps, to desire that P is in part to regard the world as "needing" to conform to P. However if one believes that P, and yet $\neg P$, it is one's *belief* that must change — or perhaps, to believe that P is in part to regard oneself as needing to not believe P if $\neg P$. Either way, it is essential to desire that it involves a standard the world is to meet; any desire, *qua* desire, is evaluable with respect to its satisfaction condition, which is that it be successful or effective.[24]

That desire has an essential connection with motivation or action (hereafter shortened to *motivation*) is somewhat more contentious. Galen Strawson (1994), for instance, holds that beings congenitally incapable of any sort of behavior can be said to desire nice weather in virtue of hoping for it and being pleased by it. But views that disconnect desire from motivation do not sit well with the Standard View, which is generally concerned to preserve a necessary connection between desire and *practical* thought.

To see this, consider a natural explanation of accedie that is available to GG theorists. Why is it that depression and exhaustion can keep one from deciding to get out of bed, which one knows one

[21] See Smith (1994) for a canonical defense of the dispositional view and Millikan (1984, 140) and Papineau (1987) for functionalist ones.

[22] Theories of this sort include Lewis (1972) and, in an interpretationist vein, Davidson (1974).

[23] Adherents of learning-based views include Schroeder (2004); Railton (2012); Arpaly & Schroeder (2014).

[24] If there are desires that fundamentally do not have propositional contents, there is a question of how to apply the notion of satisfaction conditions to them. If I desire *Sally*, but not in a way that is equivalent to desiring *that I be Sally's partner* or *that I be around Sally* or any other set of desires with propositional contents, it is less than clear how to satisfy this desire. But these desires are nevertheless thought to have a world-to-mind direction of fit, so the issue is not of present concern. For objections to propositionalism about attitudes, see Montague (2007) and Brewer (2009); for a defense, Sinhababu (2015).

ought and even must do? Perhaps it is because in these conditions, though one *believes* that one must φ, one does not at all desire to. One has the wrong kind of appearance of the good to motivate action.[25] But this would hardly be a good explanation if desires comprised ineffectual hopes. After all the typical accidic may very well hope, in a listless way, that he finds the motivation to get out of bed. Since the project of this chapter is to take part in an intramural GG dispute we can without loss of generality set aside non-motivational views of desire: as far as we are concerned desires must *be* motivations, or be dispositions to be motivated, or have the function of motivating, etc. And of course, the motivation must tend in a certain direction: it must tend towards the satisfaction of the desire. Desires are *regulated* by their satisfaction conditions.

It is also important to note is that evaluativist treatments of desire are not best understood as *competing* with the views sketched above that tie desire closely to action or motivation. Indeed, as just mentioned, one of the guiding concerns of GG is to explain how reason can be practical, and the Standard View proposes to do this by holding that evaluations are necessarily implicated in motivation and the production of action.

4.3.1 Initial skepticism about evaluative desire

In the next subsection I will raise a problem for conceiving of desires as generally evaluative, but in this subsection I wish first to set aside two other sources of skepticism. All three worries can be seen as developments of the general idea that desire cannot serve two masters, motivation and evaluation, but the first two aim for a stronger disconnection between the two roles — that motivation and evaluation must *always* be separate or separable — and for that reason miss their mark.

If desires were evaluations then they would have not only satisfaction conditions but correctness conditions. A typical proposal would be that whereas a desire that P is satisfied only if P, a desire that P is correct only if P is good, or if it would be good if P. And the fact that an evaluation has mind-to-world direction of fit has been taken to be incompatible with desire, given its mind-to-world direction of fit: Hulse et al. (2004), for instance, argue that there can be no felt desires on the basis that feelings have mind-to-world direction of fit and desires have world-to-mind direction of fit.

But it does not follow from the fact that an entity has one direction of fit in virtue of being a feeling that it *cannot* have another direction of fit in virtue of being a desire. Things can have multiple roles and functions. Indeed it is entirely possible that an entity can be in a bind where meeting its correctness condition would mean not meeting its satisfaction condition, or vice versa.

[25]This is not to say that this explanation is available to every exponent of the Standard View. Sergio Tenenbaum (2007), for instance, also holds the strong view that the motivational force of a mental state must match its evaluative content (see *op. cit.*, 3). His own explanation of accedie can be found in *op. cit.,* Chapter 8.

Suppose, for instance, that you are a *Signaler* in a game. A Signaler has two roles, and a good Signaler executes both well: a Signaler should faithfully relay the Decider's intentions (Attack or Defend) to the Executor when the Executor is near, and the Signaler should dissemble when any information is likely to be picked up by the enemy. That is, when the Enemy is near the Signaler ensures that the Enemy has a false belief about the Decider's intention. Now, suppose the Enemy and the Executor are near and the Decider intends to Attack. A problem arises, for the Signaler would meet its satisfaction condition if it confuses the Enemy, which requires that it signals Defend. But doing so would prevent it from fulfilling its role of successfully reflecting the Executor's intention. This is unfortunate for the Signaler, and if she is strongly disposed to fulfill both roles she may feel a considerable amount of consternation. But there is nothing impossible about being subject to and regulated by conflicting norms, or frankly even uncommon. So it looks like we will not find here an argument against the thesis that desires are evaluations, even in the case where desiring something puts us in a *bind*: that is, a situation in which we desire to φ, and thus represent φing as good, even though it is not good.

A second source of skepticism begins with the *Humean modal separability thesis* that whenever one is in a state representing that φing is good, it is yet always possible to fail to be motivated to φ.[26] Indeed, one might reach for the stronger thesis that any given mental representation could fail to engage one's motivation. But if representational and motivational states have such different modal profiles, the argument continues, they must have been two different states all along. Since desires are essentially motivational and representational states are not necessarily motivational, and hence not essentially representational, it follows from Leibniz's Law that desires are not representational states and hence not evaluative.[27] But in spite of its intuitive appeal, anti-Humeanism today remains popular.[28] Indeed, Chapter 2 proposed one way of rejecting the thesis: affective

[26]Smith (1994, 118-120). Another Humean worry comes from a decision-theoretic argument by Lewis (1988, 1996) against identifying desires with beliefs. However, Lewis's argument has a serious limitation: it only rules out the possibility of an agent whose degree of desire for a proposition is necessarily identical to his credence that it would be good. No GG theorist is committed to the existence of such agents. Sergio Tenenbaum (2007), for instance, who holds the strong view that there is no separation between motivation and evaluation, nevertheless holds that it is always possible for an agent to have a desire that is more like an evaluative *perception* than a *belief* that the proposition would be good. For further criticism of Lewis, see Hájek & Pettit (2004); Bradley & List (2009).

[27]This last inference may seem inconsistent with ethical expressivism, which is usually taken to deny that evaluative states of mind are representational. However, the notion of representation under discussion here is not the kind of "natural", teleofunctional or tracking notion which expressivists are most concerned to deny for evaluative states (Gibbard 1990, Ch. 6). Here a state is representational if it has a certain kind of *correctness* condition, and correctness is (according to most) a normative matter. Hence expressivists can parse what it is to accept that the evaluative state of mind M is correct iff C: it is to plan to be in state M just in case C.

Not that expressivists should endorse the conclusion of the Humean argument. Rather, expressivists should accept that there are evaluative representations (understood as above) but deny the modal separability thesis, for on their view such representations are constituted by motivational states.

[28]See for instance McDowell (1979); Dancy (1993b); Millikan (1995); Gibbard (2003); Döring (2007); and many more.

states can be representations of the good even if they can be reduced to a motivational role.

4.3.2 Desires as such are not evaluations

In this subsection I grant for the sake of argument that desires necessarily are evaluations. Still, I argue, it is not *qua* desires that they are evaluative. Since it is quite plausible that if desires are essentially evaluative, it is *qua* desires that they are evaluations, it follows that desires are not essentially evaluative. By this argument I mean to offer a conception of the nature of desire that is more compelling than the evaluativist's.

Suppose a man, call him *Determined*, takes it into his head to pet a hungry tiger in the zoo. He just really wants to; *desperately* wants to, even. His desire is not motivated by any other desire and is non-instrumental, so that Determined does not want to pet the tiger because he wants some further thing which it will constitute or cause to obtain. And he can, let us say, articulate what it is about petting the tiger that he finds attractive. He says he wants to make friends with a big and dangerous cat, he often speaks admiringly of the cat's regal bearing, he frequently wonders aloud how nice it would be to touch the tiger's muscled contours, etc. And this burning and persistent desire keeps him planning the big event for months. If ever his courage or his intention to make it into the tiger's cage flags for a moment, the desire is there to buttress it, right until the very end — or so we read in his obituary.

Contrast Determined with *Ditherer*, a woman with a similar affliction. Ditherer desires the same thing as Determined, and she would characterize her desire in the same way as Determined does. But at crucial moments, as when the time comes to survey the zoo attendants' rounds a week before the attempt, and again just before she is to jump over the fence into the tiger cage on the appointed day, her desire suddenly falters. At those points she simply no longer desires to be in the tiger cage. And, let us stipulate, her desire falters precisely because she appreciates, if only momentarily, what a dumb thing she is about to do. It is not that her prudence overpowers a persisting desire to be with the tiger; the desire simply disappears for a moment before roaring back. But the second time the desire falters, it comes back too late. Her awkward attempt at scaling the fence is noticed by the zoo attendants and she is spirited away from danger.

Now as we noted above it is agreed on all sides of the dispute that desires are essentially motivational and that they can be evaluated *qua* desires in terms of the extent to which they tend to be successful or effective: a *good* desire tends to be an effective one.[29] And it is clear that

[29] Of course, talk of "good desire" is not a natural phrase in English as "good toaster" is. But here I follow Thomson (2008, 20) in holding that we can sensibly talk of the goodness of a kind when that kind has a function or an intrinsic standard. (However I would at the same time like to distance myself from some of the other theses Thomson holds, such as that goodness of a kind is fixed or determined by the kind of thing it is. It is hard to believe that once we have an account of the nature of persons we will be in a position to determine what a good person would do in any situation.)

Determined's desire is good as a desire in this respect. It not only meets its success conditions but is quite resilient to distractions, sensitive to opportunities for fulfillment, and non-deviantly causes its satisfaction conditions to obtain. It is a desire that is *well-regulated* for fulfillment, and it is successful as a result of being regulated in this way.[30] Ditherer's desire, however, does not score so well along this dimension. It does not achieve success, and its failure is due to its own inconstancy.

We are also supposing that these desires are also evaluations, and thus that they represent petting the tiger as good. So, we can also score these desires according to how well they function as evaluations. Along this dimension the desires score quite differently: even admitting that there is *some* value to petting the tiger for both Determined and Ditherer, of the sort found in sky-diving perhaps, Determined's desire is in its strength and persistence far out of proportion to the value the act actually has. Ditherer's desire, at least, is at times sensitive to the fact that it is incorrect.

But now we can return to our question, as applied to Determined's and Ditherer's desires: are these desires evaluations *qua* desires? We can answer this question by considering how they score *overall* as desires. It is quite plausible that if they were evaluations *qua* desires, then being better evaluations — being more sensitive to the evaluative facts — would tend to make them better *qua* desire, or more simply put, better desires. They would be better at doing what essentially is one of their functions to do. So here is a test: if desires are *not* evaluations *qua* desires, then when we ask whether Determined's desire is overall better as a desire then Ditherer's, the answer should be a clear *yes*. Whether or not the desire accurately reflects the evaluative landscape will not count in determining whether it is a good desire (good *qua* desire), just as Ben Roethlisberger's moral qualities off the field do not properly affect our judgment about whether he is a good football player. If, on the other hand, they are evaluations *qua* desires, then in asking whether Determined's desire is overall better than Ditherer's, the answer should be less clear, and indeed we should be inclined to judge that Ditherer has the better desire. After all Ditherer's desire gets her close to fulfilling the desire, and unlike Determined's it is at least at times sensitive to the fact that it is wildly incorrect.

I think it is clear that Ditherer's desire is not better as a desire than Determined's is. Note that on the assumption that desires are evaluations, Determined and Ditherer's desires are in a bind, as we defined the term above. The object of the desire is not worth pursuit, and the desire (by hypothesis) represents that it is. Accuracy demands the desire go out of existence and the desire's satisfaction condition demands that it remain. If there were any intuitive pressure to include the desire's function as an evaluation in our assessment of whether it is good as a desire, it seems we should react to it as we do to the Signaler, for whom it is simply unclear what to do in a bind. But

[30] Here I take for granted that incorrectness of a desire — mere false representation — does not itself make a desire defective. A representation can be incorrect just by being unlucky, and being unlucky is not a defect in a thing.

I think we feel considerably less hesitation. Determined's desire stays true to the very end (in the sense that a well-aimed arrow stays on target) and is not inconstant like Ditherer's, and it is really the extent to which a desire is well-regulated to achieve its object that determines whether it is a good desire. Thus, even if desires are evaluations, they are not evaluations *qua* desires.

This is not to say that a desire is a better desire the stronger it is, for the extent to which a state or process is well-regulated to achieve an outcome does not correlate with the extent to which it can overpower any obstacles. A modest desire for chocolate is good as a desire to the extent that it reliably tends to procure one a modest amount of chocolate, and there is certainly nothing wrong with a desire that one for the moment does not intend to satisfy. Furthermore, being well-regulated in this way does not amount to having an intrinsic standard of persistence. A desire to φ functions to lead one to φ; carrying out this function typically involves the desire's persistence, but the standard here is directed towards the world and one's behavior, not the persistence of the desire itself.

Nor need we deny that desires need not be sensitive to a range of factors including whether the desire can be satisfied at all, whether the possibility of its satisfaction is too far into the future to bear even present planning, and even whether the thing desired is any good. The point here is that there is a distinction between it being better to have desires for good objects and a desire's being better *qua* desire for desiring a better object. Desiring better objects does not necessarily lead to desires that are *better as such*. It is just better to desire what is good, for the obvious reason that desiring better things tends to lead to better things. (Or at least that is so if one's desires are good enough *qua* desires, otherwise one might lose one's appetite at crucial moments like Ditherer does.) The faculty of desire is like a hunting dog that can be equally well-trained on a variety of prey: the goodness of a hunting dog is determined by how well it hunts, not *what* it hunts.

Someone may also object that although a desire functions better as a desire to the extent that it responds appropriately to the evidence of the goodness of its content, its effect on scoring Ditherer's desire is masked by the desire's inconstancy. On this view inconstancy is such a bad fault in a desire that it is difficult to see that Ditherer's desire is to some extent a good desire insofar as it is responsive to some degree to the evidence that it is incorrect.[31]

To respond to this objection, consider another example. Suppose I hear on Monday that *A Touch of Evil* is playing at the Desperado Theater this Thursday. This is great for me since the Desperado always has free popcorn, and I think it worthwhile to watch a film only if Orson Welles is both directing and acting in it and if I can eat free popcorn during the show. Let us further suppose that I *know* it is worthwhile to watch a film only under these conditions, for they reflect deep and long-standing values of mine, and what makes it worthwhile to watch a film depends on one's cinematic and concession values.

[31] Thanks to Rohan Sud for discussion on this point.

As it happens, on Wednesday I hear that the Rialto Theater, also known for its free popcorn, is playing *Citizen Kane* on Thursday. By my accounting of things the two options are equally good, but by this point on Wednesday my heart is set on *A Touch of Evil*. My desire to see it is considerably stronger than my desire to see *Citizen Kane*. So, I form the intention to see *A Touch of Evil* tomorrow. Certainly it is common and appropriate for one's preferences to become settled in this way;[32] we are not rationally required to desire options exactly in proportion to our judgment of their merits. There is nothing defective about my desiring to see *A Touch of Evil* more.

But now suppose that when I arrive at the box office of the Desperado I discover to my chagrin that the popcorn machine is broken. This makes it the worse option in my view; I judge that it would be best to go across the street to the Rialto. Suppose that I now reopen the question of whether to continue buying a ticket to see *A Touch of Evil*, and I find that I still *want to see* A Touch of Evil *more*. It is not that I am too lazy to go across the street; it is just that learning this piece of information has not changed my preferences at all. We can all admit that if I were to *decide* to see *A Touch of Evil* after all my choice would be defective, since it would be akratic, i.e. it would be a choice contrary to my better judgment. For that reason it would not be a good choice, *qua* choice. But it doesn't seem that my *desire* to see *A Touch of Evil* is now defective. It was not defective before, and it is not defective now. That is a problem for the evaluativist, who it seems is committed to judging that my desire is defective on account of its not being responsive to the evidence. It is functioning poorly as an evaluation of the value of its object.

No doubt the effectiveness of such appeals to intuition as I have been making will depend on one's favored theory of desire. Certain theories of desire which view them as more complex states will tend to view evaluations as constituent components of desire. But we can use the same kind of considerations to press the point that even in cases of desire that best support these theories, a more accurate orientation towards the good need not come with improvements in desire.

Take Talbot Brewer's (2009, Ch. 2) understanding of desires for what he calls *dialectical activities*, which begin with a desire to engage in a good activity (as such) and constitutively involve a continual process of refinement or evolution on our part. As Jill began graduate school she transitioned from merely thinking that it would be good to be a good philosopher to actively desiring to be a good philosopher, but she began with only a vague idea of what that involved. She knew that clarity was a constitutive ideal of good philosophy but often thought that clarity meant formal, technical work. After years of actively trying to become a good philosopher she was forced to recognize that this conception of clarity needed refinement, for sometimes formalism can obscure a theory without adding to its explanatory power. Heedless pursuit of formalism conflicts with other ideals of good philosophy too, since sometimes the precision it lends to one's ideas is merely artificial. Hence, the style of philosophy she aims to produce has changed over the years. She now

[32]Here I use 'preference' merely to compare desires: I prefer A to B if and only if I desire A more than B.

has a more accurate conception of what good philosophy is and a more refined ability to recognize it.

But has her *desire* to be a good philosopher improved *as a desire* by her having a more accurate conception of philosophy and a better ability to recognize it? In many respects it need not have improved. Over the same time she could have become tired of philosophy or jaded with it, or wavered in her commitment to it, or come to prioritize other projects. Granted, the desire does seem to have improved as a desire by becoming more refined. Jill has after all become more discriminating in her pursuit of good philosophy. However, the sense in which refinements of desires are improvements of desires is easily explained by the refinement's contribution to the effectiveness of the desire. We began with a non-instrumental desire with normative content, to be a *good* philosopher, so it is no surprise that a more accurate and discriminating conception of what it is to be a good philosopher enables Jill to become a better philosopher by acting on that desire. But we could have seen the same kind of improvement if we had started with a desire with non-normative content: suppose Jill is instead a budding artist who wants to make a genuinely *new* or *original* kind of art, but begins with little idea of what counts as new or original in the context of art and refines her concept of it through the practice of making art and studying its history. What is doing the work here is the fact that the refinement or specification of an original desire contributes to its effectiveness, not the supposed fact that desires are better just for being better aimed at the good.[33]

Now, one could take issue with my proposed test for determining whether a desire's evaluation is essential to it in the following way. The test depends on scoring desires by whether they are successful and evaluations by whether they are accurate, and one might think that what matters for whether a desire is a good desire is not directly whether it is successful or correct but whether it responds as it *ought* to. Plausibly, when we regard a desire's motivational dimension we are inclined to say that it ought to respond in ways likely to bring about its fulfillment, as Determined's does and Ditherer's does not. But, the objection continues, things may be different when we consider its evaluative dimension. A well-functioning visual system *ought* to misrepresent an oar in the water as broken. Perhaps something similar obtains for Determined's desire: it ought to misrepresent the tiger as extremely good to pet. Or perhaps, to try a slightly different strategy, it is evaluating the tiger just as it should *given* his longstanding overall attitudes towards tigers and his other desires and attitudes: perhaps he has long been enamored of tigers, has clear-headedly

[33]One could reject this line of argument by asserting that originality is an aesthetic value, and so accuse me of not having started with a *non-normative* content to the desire. But this misunderstands the force of the argument. What works of art *are* original will surely be affected by human interests, which in turn are (especially if GG is true) directed by values. But the concept *original* is not a normative concept, and the example goes through so long as we can find a desire with non-normative content that undergoes a similar process of refinement. Jill could want to be an astronaut — not even a good astronaut, just an astronaut *simpliciter* — and improve her desire by refining her beliefs about what is required to become an astronaut.

accepted the risks associated with petting them and has shunned starting a family for that reason, etc. One might think that, given those conditions, his desire evaluates the tiger just as it ought. And if that is so it seems the proponent of the standard view can also explain the intuition that Determined's desire is better *qua* desire.[34]

Or so the objection goes. However, I do not see how it can succeed. There is really no sense in which Determined's or Ditherer's desires are well-functioning as evaluations of any kind of good, nor that they are operating just as an evaluation of that sort ought, precisely because they are so far out of line with normative reality. Determined's desire is more analogous to persistently seeing the oar in the water not as broken but as bent in on itself. And although we can make judgments about how an representation ought to respond to certain evidence while bracketing others, as when we say that a coherent dinosaur-denier ought to believe that there are no dinosaur fossils, conforming to an "ought" of this bracketed or conditional sort does not make a representation a good representation *simpliciter*. The most we could say of Determined's desire, as an evaluation, is that it is consilient with his other desires and consistent with this past evaluations of the value of tiger-petting — but from this it hardly follows that his desire is a *good evaluation* of the value of tiger-petting.

So, I conclude that even if a desire is an evaluation, it is not an evaluation *qua* desire, and therefore that desire is not essentially evaluative as claim 22 asserts. And thus, it seems, if GG is true it is not in virtue of the nature of desire . We await an explanation of the thesis.

4.4 Purely instrumental practical thought

In this section I press an objection against the Standard View as a whole and to the strength of GG thesis it entails, namely that it is a metaphysically necessary truth. Because the success of the objection would weaken the strength of the internalist thesis GG may be used to explain or derive, it is to the internalist project I first turn.

As we characterized the standards of practical reason in the previous chapter, they are necessarily normative for agents. Thus the fact that *the standards of practical reason* are necessarily normative for agents admits of a trivial explanation when the italicized phrase is given a *de dicto* reading. But it is *not* so obvious that it admits of a trivial explanation when it is given a *de re* reading. To see this, note that the standards must have a substantive characterization if they are to function as guides or standards for conduct: they must have a characterization on which they rule out, permit, or suggest particular courses of action for particular agents. To deny this is to make of practical reason a game with no way to win, much as if one were hunting what could only be described as *the quarry* (to use an example made famous by Velleman 1996).

[34]Both suggestions are drawn from Tenenbaum (2007, 153-155).

But once given such a characterization there is a substantive and interesting question as to why those standards and not some others are the standards of practical reason, and for any candidate set of standards one can coherently imagine a skeptic about them. Indeed the skeptic can press the same question even if we are unsure of what exactly the standards are, since the reasoning above guarantees us that if there are any standards of practical reason they will be characterizable in a way for which the question makes sense. This is one way of understanding the constitutivist's project: in the face of a skeptical challenge she attempts to explain both the application of and to vindicate the normative force behind some or all standards of practical reason in terms of abstract, essential features of agency.[35]

GG and constitutivism are similar in that proponents of both frequently hold that agents aim to meet (some) standards of practical reason, but it is important to be clear about their differences. The good of GG has historically been connected to value*, especially the moral good*, well-being, and goods* that provide agent-neutral reasons, such as privacy. To be a constitutivist, by contrast, one need only aim to derive *some* standards of practical reason from the nature of agency. But given our characterization of the good of GG in the previous chapter in terms of the standards of practical reason, this difference is more sociological than essential to the positions.

Other differences are of substance. GG holds that one must in some sense *see* one's action as good, as meeting a standard of practical reason. Constitutivists are not committed to the view that the standards which they aim to derive are ones the agent must see as good. On Michael Smith's (2015) version of constitutivism, for instance, agency is held to be a "goodness-fixing kind" in Thomson's (2008) sense of the term, and in particular he holds that a good agent satisfies its desires. This standard, that a good agent satisfies its desires, is thus held to be constitutive of agency. This enables us to derive further standards for agents: good ones are unified and coherent in their desires (lest the satisfaction of one frustrate another), sufficiently knowledgeable about the world, and instrumentally rational. Smith then argues that the *best* agents, the ideal ones, must moreover have dominant desires to be coherent. But note that it is possible for one to be an agent and yet sufficiently non-ideal that one not actually desire coherence. One need not "see" the fact that forming a desire would make one's desires more coherent as counting in favor of forming it. Perhaps one might need to *be* sufficiently coherent in order to be an agent at all, but one may be indifferent to the prospect of coherence to the point of ignoring it entirely in one's actions and deliberations. And if one ignores such considerations entirely then one does not "see" them in any relevant sense. So on Smith's view, it is true both that coherence is a standard for agents that is derivable from the nature of agency and that agents can fail to see coherence as good.

A second difference is that unlike constitutivism, GG need not be a thesis about what is es-

[35]Prominent constitutivist accounts include Korsgaard (1996, 2009); Velleman (1996); Railton (1997); Smith (2013). For a recent review, see Katsafanas (2016).

sential to or constitutive of agency as such. This is so in two respects. First, in this dissertation it is not interpreted to be about the essence of anything; GG could be a robust contingent truth. Second, GG is first and foremost a thesis about *motivation* or *action*, not agency. This fact largely explains why David Velleman can be a constitutivist who rejects GG. He holds (Velleman 1996, 180, 193-199) that it is constitutive of being an agent that one have an inclination to act out of a controlling consciousness of what one is doing, and that reasons for acting are to be explained in terms of this inclination. What is thus constitutive of agency on this view is a *particular* kind of motivation. But desires generally, he holds (Velleman 1992a, 117) need not aim at the good; they need only aim at the attainable.

And — to return to the main issue — a third difference between the two lies in just what GG and constitutivism attempt to explain. Recall from Chapter 3 that the standards of practical reason are held to have normative force for their agents. Constitutivists are not content to take the normative force of those standards as given. They tend to think that if the fact that I have a reason to do something has a "normative grip" on me, then there must be something about *me* that enables it to grab on. So, they seek an explanation of the normative authority of (some) standards of practical reason in terms of features of agency. The constitutivist thus aims to reply to a certain kind of skeptic, one who purports to be an agent and yet rejects the authority of certain norms.

Now, I did argue in the previous chapter that if GG is to avoid being weak, there must be some standards such that it rules out acting under their guise. And given an appropriate ought-implies-can principle, this would enable us to derive from GG a restriction on what standards *could* apply to agents. But this is not at all the same as aiming to derive substantive standards. GG is better seen as addressing a slightly different, complementary concern. Often, GG theorists aim to establish that there is a particular kind of mistake that agents *cannot* make: they cannot completely divorce themselves from the pursuit of good as they might from the pursuit of respecting the law or other standards. This is clearest in the work of Gavin Lawrence (1995), who holds that it is a hallmark of the "traditional conception of agency", which includes GG, that there is one sin against reason agents cannot commit: they cannot seek what is merely apparently good as such. Insofar as they seek anything they must seek what is *really* good, even if they are mistaken about what that is (*ibid.,* 129-130).[36] The good is the formal object of agency, according to many of these authors (Tenenbaum 2007, 2008); the good is constitutive of the defining question of practical agency, "What should I do?" or "What is best for me to do?" (Lawrence 1995; Watson 2003; Boyle & Lavin 2010). So whereas the constitutivist attempts to justify agents' conformity to standards of practical reason, the GG theorist aims to show that a certain kind of failure to conform to them — that embodied in wholly counter-normative agency, in which an agent is not in any way guided to

[36]Lawrence takes himself to follow Plato here; see *Republic* 505d-506a.

do what is good for her to do — is impossible.[37] He aims to destroy the very ground the skeptic needs in order to launch her attack.[38]

I am sympathetic to GG's internalist, anti-skeptical project, but it is too ambitious in aiming to show that this mistake is *metaphysically impossible* for an agent to make. In this section I make the case that GG is not a metaphysical necessity. In doing so I provide an argument against the Standard View, which entails a version of GG of this strength. The argument to follow does not by itself diagnose the problem with the Standard View. It could be that the creature below does not desire to do what it does (rendering thesis 21 false) or that its desires are not evaluations (22). (However, the conjunction of the argument in §4.3 and the added complications involved in interpreting a creature as acting for a reason but not as motivated by a desire together suggest that the creature is best read as acting on non-evaluative desires.)

Suppose that in venturing through a foreign country we come across an unfamiliar creature known as *the Wobbly*, which we first encounter engaged in a curious activity on the rocky cliffsides of this land. Perched on a rocky outcropping at a safe distance, we observe it from below. It scampers up them quite deftly, and upon reaching some kind of object T at the top — we cannot quite make out what it is from our vantage point — it inspects it, and if it is relatively large it dashes it to the rocks below and returns to a bottom ledge. But, crucially, we can see that before each run to the top the Wobbly spends a good deal of time sizing up the cliffside below. In fact, it even seems to pantomime some of the moves it later performs on the cliff during the run. Being good scientists we record each run as it occurs and later determine that the route the Wobbly takes to each T minimizes a weighted sum of horizontal distance traveled and total force exerted on the rock.

What is the Wobbly doing just before a run? Presumably it is thinking through various routes and the kinds of bodily motions involved in their traversal. This, I think, is already a kind of reasoning. But no doubt it thinks through these routes *with the aim of enacting one*, and so its reasoning is practical. How else does it get to the top? Indeed it seems that many of its actions are done for a reason, too: its reason for scampering up one rocky crag is *so that it can get to the T at the top*. We could be skeptics about its agency, of course; we could wonder if it is an Agent Zombie, a being that to all appearances behaves just like an agent, and perhaps even has an internal functional organization isomorphic to that of an agent, even though all is passive within. But defeating such skepticism about agency is beside the point here, for the Wobbly is just an

[37] Note that this project is not to be confused with moral rationalism, according to which immoral behavior entails some failure of practical reasoning on the part of the agent (perhaps in having chosen the objectively incorrect action). Vogler (2002) is a GG theorist who argues against moral rationalism.

[38] It bears mentioning that the skeptical ground is currently occupied by, among others, Wallace (2001) and Setiya (2007). See Watson (2003) for a general discussion of the possibility of counter-normative agency with special focus on the will.

illustration of a being that thinks through how to achieve a goal it has, and has no capacities other than what is necessary for that purpose. Certainly such are possible.

Now, if the Wobbly is an agent who reasons through its actions, then according to the strong version of GG, its T-throwing manifests one way of being aimed at the good. But *what* is the good at which it aims? Knowing what the T are does not help much; suppose they are towels. This narrows down the possible aims the Wobbly might have: perhaps it aims to skillfully reach and toss large towels, or get as many of them to the ground while expending the least energy. But the problem is that this gets us no nearer to a standard *of practical reason*. *What* is the reason the Wobbly has to throw down towels? What good does it achieve? Why ought it do that? We are no nearer to answers to these questions. Even worse, recall that towels were picked arbitrarily as a value for T, and that the Wobbly is intelligible as thinking practically no matter what the Ts are. How could the GG theorist find a kind of good the Wobbly is regulated by no matter *what* it aims to throw off the cliff without engaging in utter speculation about its motives? ("Perhaps it has long wanted revenge against the large towels.")

Here is one way of sharpening up this reasoning. The argument works just as well against a kind of constitutivism which holds that agents are necessarily guided toward the good, and although I will not discuss this version of the argument further, I have put it in square brackets for reference:

The Wobbly Argument

25. The Wobbly performs many actions for a reason, such as starting a run and going left rather than right at a certain point on the cliff. Furthermore, the Wobbly is not in any way misrepresenting what it is doing. Its action does not depend on a false appearance. [The Wobbly is not in any way unsuccessful or failing in any of its aims.]

26. The Wobbly achieves nothing good by throwing Ts off the cliff. (We can easily stipulate this; there must be *something* that a creature like the Wobbly has no reason to do, or more generally something that the standards of practical reason do not sanction, and it might as well be throwing towels off a cliff.)

27. If the Wobbly does not misrepresent what it is doing and achieves nothing good by what it is doing, then it does not see its action as good. [If the Wobbly is not in any way unsuccessful or failing in any of its aims, it does not act out of any sort of intrinsic guidance towards the good.]

It follows from 25-27 that the Wobbly does not see its action as good. This shows that GG is possibly false (in the metaphysical sense of possibility, not the epistemic sense), and since the Standard View implies that GG is necessarily true, it refutes the Standard View.

The justification of premise 27 is that "representation" is here being used in a theoretically lightweight sense. Throughout we have been supposing that an action appears good to an agent

only if the agent as an appearance of that action as being good, and an appearance is a representation.[39] So if an agent sees an action as good, she represents it as good. But if (as premise 26 states) her action is not good while she represents it as such, then she *mis*represents it as good. So if an agent does *not* misrepresent a certain action as good, then given that it is not good, it must be that she does not see it as good.

One might think that the reason there must be something the Wobbly can do that would not be good for it to do (premise 26) is that this is entailed by the fact that norms must preserve the possibility of error: perhaps for any norm applying to an agent there must be some possible situation in which the agent falls short of it. However, this kind of possibility-of-error principle has been questioned.[40] The better justification for 26 is simply that a conception of practical reason according to which *every* possible action for *every* possible creature in *any* circumstance has something to be said for it is extraordinarily implausible. This is especially so once we consider that a GG thesis which claimed merely that agents needed to see a minuscule amount of good in their actions would hardly be an improvement on the objectionably weak WNL version from the previous chapter. For small enough degrees of good there would hardly be a difference between seeing an option as good and seeing it as practically salient. A more sensible but still limited GG view might hold that agents must see their actions to be good enough that they would merit consideration if one were to deliberate about them, for instance. But then, once again, it is implausible that *every* possible action for *every* possible creature in *any* circumstance would merit consideration. And this gives us space for the argument: the Wobbly is to be understood as a creature that is doing what merits no consideration in its circumstances. Given that the Wobbly is a metaphysically possible creature, it follows from 25, 26, and 27 that the internalist thesis is false and that GG is not a metaphysically necessary truth.

I suspect that the premise opponents will need to reject is 25, and in particular that the Wobbly is not misrepresenting its action. Here is a plausible line of thought that would seem to guarantee GG: in order to act for a (motivating) reason, one must treat a consideration as a normative reason; and in order to treat something as a normative reason, one must see it as such.[41] Both premises can and have been attacked. We saw a version of this argument put forward by Tamar Schapiro (2014) in §3.2.1 of the previous chapter, and there (p. 33) we saw one reason to reject the second premise: the sense in which an an agent treats a consideration as a reason may amount to a matter of the agent's functional organization that does not also amount to its representing anything. Setiya

[39] In fact we have been supposing a little more than this. X is a representation only if X has correctness conditions, but the appearances relevant to GG are appearances *to the agent*. For that reason we have been working with a conception of representations as correctness conditions presented to a subject, but this further feature plays no role in the Wobbly Argument.

[40] Lavin (2004).

[41] This is a reconstruction of an argument mentioned in passing by Sarah Buss (personal communication), who is not to be blamed for any misinterpretation on my part.

(2007) rejects the first premise by completely divorcing motivating reasons from normativity, tying them instead to explanation. But the Wobbly invites us to see a different way of rejecting the first premise. We can grant that when the Wobbly's motivating reason for taking one route is *that it is short*, she in some sense treats that consideration as "favoring" taking that route and even sees it as "favoring" the route. But given that the Wobbly has *no reason* to throw towels off the cliff and no reason to take this route in order to do so, and more generally that she does *nothing good* by scampering up the cliffs, it seems far more plausible to interpret the Wobbly's "seeing as favoring" in terms of (what we called in Chapter 3) a *weak normative light*. She sees this route as practically salient in light of its shortness, or experiences the rock face in that region as soliciting her climbing. And if that is all the normativity she sees in her action, then as we established in Chapter 3, that is not enough for the Guise of the Good. Such views face the weakness worry.[42]

Most purported counterexamples to GG fail because they begin with a possible action and ask what good the agent saw in it, hoping to make it obvious that there could not have been any such good. It is relatively easy for the GG theorist to show in such cases — especially when they involve human agents in familiar social situations — that it is after all plausible that this or that value is guiding the agent in acting: either that she is guided accurately by some "small" value, such as in arranging the tea cups just so because it is so *dainty*, or because she misapplies some other value, as if she were to commit an evil act in the mistaken belief that the victim deserved it. Such maneuvers have caused at least one prominent skeptic of GG to cease searching for counterexamples to it.[43] Here the strategy is different: we stipulate that an agent achieves nothing good by its action and then point out that it is possible for an agent (though perhaps not a *human* agent) to be guided to produce just that effect, without having to further represent it as good.

Let us now consider two main classes of response to the Wobbly Argument. The first complains that the notion of *good* relevant to GG figures as a *formal object* or *aim* of action or desire: it is something that action or desire necessarily has as aim (Tenenbaum 2007, 6), or is a goal of action or desire stated in terms of (or in terms that depend on) that goal (Velleman 1996, 176). Relatedly, one might object that the Wobbly does aim at one good at least, for it aims to *act well*. But merely noting the intended necessity of the connection between action and the good is no help to the GG theorist who wants to hold on to the strong internalist thesis. Recall our argument that if GG is to

[42]Some may protest, following Scanlon (1998), that there is nothing more to the concept of a normative reason than that of a consideration's favoring or counting in favor of an action, so that once we admit that the Wobbly in some sense sees the shortness of the route as favoring climbing then we have already described her as acting under the guise of the good. But even if Scanlon is right in this respect, and even if he is furthermore right that the concept of a normative reason is primitive, reasons play a role within the theory of practical reasoning that affordances or mere practical salience do not play: among other things they justify action and weigh against each other to determine what an agent ought to do. Nevertheless it sounds correct to my ear to describe an experience of a doorknob's shape as favoring turning it. So perhaps the notion of a reason is that of a consideration's counting in favor — but not just ant kind of *counting in favor* will do.

[43]Setiya (2010).

avoid being too weak, it must admit of the possibility (for some agents in some circumstances) of doing nothing good.[44] So even if the good is the formal aim of action and there is nothing more to the concept of the good than that of acting well, there must be some way in which one can act poorly. What we are contemplating in the Wobbly is a being which aims at, and achieves, nothing but what is not good according to such a standard, and acts for a reason in furtherance of that aim. It forever acts poorly. The burden is such arguments for the good's being the formal aim of action or desire to show that the Wobbly *must* misrepresent its action as good. I cannot here hope to show that any such attempt is doomed to failure; each argument will need to be considered on its merits. But note that we could easily modify the example so that throughout its entire lifespan the Wobbly does nothing but cast towels off cliffs, and is a descendant of generations of towel-throwers. I find it incredibly implausible to suppose that beings could so persistently misrepresent the world in an important way.

Boyle & Lavin (2010) take a different, Aristotelian tack which would be instructive to consider. The relevant features of their view can be condensed into two theses:

28. A rational being, *qua* rational being, is guided in acting by knowledge of what it is doing and why (where "why" asks after an explanation of the action).

29. The form of an action explanation for a being is necessarily given in terms of the being's function, and thus in terms of the good of its kind.[45]

The result, according to Boyle & Lavin, is that

> [j]ust as a rational believer is one who can reflect on his grounds for belief by putting to himself the question "Why *p*?" so a rational agent is one that can reflect on his grounds for action by putting to himself the question "What speaks in favor of doing *A*?" (*ibid.,* 191).

This analogy too suffers from an equivocation on weaker and stronger senses of "counting" or "speaking in favor", and both premises of the argument are quite controversial. But the more fundamental problem behind Boyle & Lavin's approach is that the route from 28 and 29 to the Guise of the Good is harder than they let on because there is a considerable gap between goodness of a kind and the standards of practical reason. There are indeed defenders of the view that normativity in general is fixed or determined by goodness of a kind,[46] but the Wobbly can also serve as an argument against these views.

[44]Tenenbaum (2007, 32-33) complains that it is unfair to demand of the GG theorist a substantive conception of the good when there is not likely to be a substantive conception of truth. (For instance, it is unlikely that we can explain the concept of truth in terms of the concept of some concrete property.) But what's being demanded here is not a substantive *conception* of the good but a determination of *substantive goods*, i.e. a division of possible actions into those that do and those that do not meet a standard, where both categories are non-empty. And note that however deflationary we are about the concept of truth, there are substantive truths and untruths.

[45]Drawn from Boyle & Lavin (2010, 188-198).

[46]Notably Foot (2001); Thomson (2008).

Suppose that there is a standard of goodness associated with being a Wobbly and that it is fixed by the function of the Wobbly as an organism. Either it is part of the function of a Wobbly to throw towels off cliffs or not. If it is *not* a part of the Wobbly's function, then the Wobbly Argument gives us excellent reason to believe that the explanation of an action cannot always be given in terms of a being's function, and thus that 29 is false. Suppose then that it is a part of the Wobbly's function: when we see a Wobbly throw a towel we can truthfully say "There's a good Wobbly," meaning that this Wobbly is good-of-its-kind. But why should this lead us to think that the Wobbly is doing what it has *reason* to do, or what it really *ought* to do? Recall that towel-throwing was chosen as an activity that the Wobbly had no reason to engage in. (Perhaps Wobblies should not be expending their energy uselessly throwing towels and instead should be working to improve their habitat, even though none have reliably done so for a thousand generations, and that the only explanation of their continued existence as a group is their lack of predators and speed at which they reproduce.) We have an amount of freedom in defining kinds or in designing organisms (conceptually or through genetic engineering) to have certain functions, and this account of normativity seems to wrongly rule out the very possibility of designing them such that their function — what a good specimen of their kind would do — is to do what they have no reason so.

The second class of responses attempt to point out a value that the Wobbly can be interpreted as aiming to produce. Plausible candidates include that the Wobbly aims to act *skillfully* or *efficiently*, especially when this latter is interpreted in terms of means-end efficiency. It is indeed difficult *not* to make these plausible interpretations of the Wobbly's motivations because it is hard to give a convincing example of action for a reason that does not have a broadly instrumental or calculative structure.[47] But unfortunately for those hoping to hold onto the strong GG thesis, appeal to this value will not help. It clearly will not help those who accept the intelligibility motivation for GG, which as we saw in Chapter 3 requires a conception of the good, or a "point" in acting, that is not merely formal. And it seems that skillfulness and means-end efficiency are merely formal characterizations of a point in acting. *No matter what* one is doing one can ask whether it is done skillfully, just as one can be means-end efficient no matter what one's end.

Of course, not all GG theorists put much weight on the intelligibility motivation, and quite a few philosophers believe that a means-end principle of some form or another is a standard of practical reason.[48] It is hard to show that the Wobbly is not intrinsically guided to conform to such a standard as this, but it also seems that this kind of standard is not the kind of standard of practical reason that GG theorists have in mind.[49] Note, for instance, that few would think it holds up the

[47] See Vogler (2002) for an argument to this effect.

[48] See Railton (1997); Hubin (2001); Broome (2013); and many others.

[49] The one possible exception here is Vogler (2002), who counts *the useful* (relative to one's aims) as one of the basic kinds of practical good. However, it is unclear what restrictions Vogler believes apply to end-adoption, and in recent work she appears to hold that human action cannot be understood apart from a tendency to perform acts that

desired analogy between truth and goodness. Truth is the standard of correctness for belief, but few think that mere conformity to principles of instrumental reasoning is the standard of correctness for action or intention. Aiming only to desire (say) end-conductive means, or skillful ways of completing an action, is not the same as aiming to desire correctly. I do concede that the Wobbly argument cannot be used to rule out such versions of the Guise of the Good on which it suffices to act under the guise of the good that one see one's action as skillful or conducive to some end or other, where there is no further restriction on what ends one may adopt or skills one may develop. But I regard the territory as barren, and not worth attacking.[50]

The moral of this section, then, is that if we restrict our attention to the standards of practical reason relevant to GG, the internalist thesis is false: the Wobbly shows that it is metaphysically possible to act for a reason without aiming at or representing the good. This shows that the strong version of GG, which interprets it as metaphysically necessary, is false. If a version of GG is true, it will be one of less modal strength.

4.5 Hard-line affectivist GG

The central flaw in the Standard View that both of these arguments exploit is the strength of connection it attempts to forge between motivation (or action) and evaluation: in particular its problem seems to lie in the thesis that motivation of the kind relevant to action for a reason *necessarily* implies evaluation. The thought, of course, is that only a connection of this strength will vindicate a necessary internalist thesis. But if the necessary internalism is not to be had then the GG theorist should consider whether she can get more by aiming for less. Why, after all, should GG be thought to be a metaphysically necessary truth about practical reason if it is true at all? Would it not be just as interesting if it were only a deep truth about human (or more broadly, animal) action and practical reasoning?[51]

We might, then, hypothesize a human faculty for acting in light of evaluations on the model of other human faculties. Humans have a perceptual faculty for taking in vast amounts of rather detailed information about the surfaces of objects around them both near and far. As it happens this role is primarily filled in humans by vision, and it turns out to be a fairly deep fact that visual systems represent and process color-like surface-reflectance properties.[52] Nonetheless, it could

merit praise and avoid acts that merit blame (Vogler 2014, 63-66).

[50]Note that this version of GG is also subject to the inertness worry of Chapter 3: it is not clear that it has any implications for the theory of practical reason.

[51]Ultimately I am interested in the possibility of an *animal* faculty of this sort for action and practical reasoning, insofar as animals reason. However because it is the central case, and for the sake of familiarity, below I explicitly mention only the possibility of a *human* faculty.

[52]See Hardin (1988, Ch. 1). The relation between colors and the kinds of properties the visual system represents (here simplified as "surface reflectance properties") is — to put it mildly — under dispute. However, the exact nature

have been the case that humans had developed some other faculty in place of vision which had no deep relation to colors, such as echolocation. Could the human faculty for practical reasoning operate uniquely via the appreciation and processing of *values?*

Of course, given Chapter 2's defense of an intentionalist treatment of certain affective states — pain experiences — and their role in rationalizing action, it would be natural to look to affective states generally to fill this role. According to what I call "hard-line affectivism" about human action, affective states represent the existence of normative reasons for action and are necessary for human action for a reason.[53] More precisely:

Hard-line affectivism (HLA)

30. In order to act for a reason, one's action or intention[54] must be based on an affective state that represents there being some normative reason for that action.

31. Whenever an action is based on an affective state that represents there to be some reason for that action, one acts for that reason.

To these two theses which comprise HLA, I will add a third, the *judgment externalist module (E)*. Although HLA can be defended independently of the judgment externalist module, the two form a natural package (HLA+E).

Judgment externalist module (E)

32. Conceptual normative judgments do not necessarily motivate action. When they do, it is because they are affectively enriched.

The judgment externalist module allows us to offer a potential explanation of accedie, the failure to be motivated to do what one judges one ought to do: in many such cases one may lack an affective representation of one's action as good. The severely depressed person may recognize that she ought to get out of bed, but the thought leaves her cold. For her recognition to be motivating, she would need to *feel the need* to get out of bed in recognizing that she ought to. This is captured by the notion of an *affectively enriched* judgment, which is quite simply the notion of judgment with feeling. Note that an affectively enriched judgment is a judgment that is colored by affect. The mere co-presence of judgment and affect does not suffice for this, for the affect must (in the nomenclature of Chapter 2) be intentionally attached to the judgment.

---

of the properties represented by the visual system is not a matter of direct concern in this chapter.

[53] Recall the tripartite distinction among normative reasons, motivating reasons, and explanatory reasons. The "reason" of action for a reason is motivating: action for a reason is action for which there is a motivating reason. The reasons that are of primary concern to the normative theory of practical reason are normative reasons.

[54] Some may hold that only intentions to act, and not actions themselves, can be based on mental states. It is not necessary for me to take a side here, but I will below interpret the condition to apply to actions largely for the sake of simplicity.

In the following section (§4.6) I lay out the details of the view and differentiate it from extant views of the relation between emotion and action. But ahead of that it would be good to note a general feature of HLA.

Whereas thesis 31 is likely metaphysically necessary if true, 30 and 32 are only nomologically true if true at all, and in particular have the modal status of psychological laws. It likely falls out of the nature of acting for a reason that when one bases one's action on a state that represents there being some normative reason for that action, one acts for that reason. However, it would be extremely surprising if acting for a reason had an essentially affective nature. Indeed, insofar as the Wobbly can be conceived not to possess, let alone exercise, any affective states whatsoever when it acts for a reason, then acting for a reason does not necessarily implicate affective states.

This introduces two important caveats about the ambitions of HLA and the version of GG it entails. Just as it is metaphysically possible to modify a human to take in vast amounts of information about the surface properties of objects through a non-visual system, perhaps through echolocation or extendable tendrils, so it may very well be possible to modify a human to act for a reason on the basis of a non-affective reason-representing states, or, as with the Wobbly, not on the basis of an evaluation of the action at all. The claim is that human action that violates HLA violates a contingent psychological law and is thus functionally quite abnormal. *HLA is not an account of action for a reason as such*. The second caveat is that, although HLA depends on a contingent hypothesis which is in principle subject to empirical investigation — namely that human action for a reason is based on affective states — the present work does not attempt to evaluate any evidence for that hypothesis, nor will it design a method for doing so. At this stage of inquiry it suffices to outline the theory in the context of evidence in light of which it is a reasonable hypothesis and to give it a preliminary philosophical defense.

4.6 Hard-line affectivism: The details

4.6.1 How-possibly

But how could it be that in humans there is a robust contingent connection between action for a reason and value? We can lend plausibility to the account by seeing it as a likely implication of a very common theory of action for a reason and a certain anthropocentrism about value. Fully articulating and defending such a derivation would take us far beyond the more preliminary and exploratory purposes of the chapter, however, so the present subsection merely sketches how such a connection could go.

4.6.1.1 Action based on representations of reasons

The canonical and still most popular theory of action for a reason is Davidson's (1963) causal theory according to which an action φ for a reason is non-deviantly caused (Davidson 1978) by a pro-attitude towards acting in a certain way and by a belief that by φing one would act in that way. Note that it is quite reasonable to suppose that any pro-attitude towards X presents X in a weak normative light: it presents X as *to be made to obtain*. To hold furthermore that an action must be *based* on such a pro-attitude/belief pair would amount to a less committal version of the canonical theory, given that the most popular accounts of the basing relation are causal accounts. (I say a little more about how I interpret the basing relation below.)

But as Davidson himself rightly points out, more is needed. The account so far is not sufficient for action for a reason. "We cannot explain why someone did what he did simply by saying the particular action appealed to him", he writes; "we must indicate what it was about the action that appealed" (Davidson 1963, 3). If I could see the doorknob as simply to-be-turned, and was thereby (along with the belief that by turning the doorknob I could turn it) motivated to turn it, intuitively it would not be true that I turned it *for a reason*. There must be some further feature of the action, or more broadly some further consideration, in the light of which the action appears to me to be favored — which seems to me, in the weak normative sense, to count in favor of that action.[55]

Now, as it happens, Davidson took one's reason for acting — one's motivating reason, that is — to *be* the pro-attitude/belief pair. Contemporary action theorists are more likely to be "anti-psychologists" about motivating reasons who count the associated consideration as the motivating reason, not least because it better accords with folk judgments about reasons.[56] It is sensible to equate our motivating reasons with the grounds that we could in principle cite for our actions and intentions, and we cite our own mental states as our grounds only in the minority of cases.[57] So a fairly orthodox analysis or account of the nature of action for a reason might hold the following:

AFR Action for a reason just is action based on an appearance to the agent of a consideration as favoring — in the weak normative sense — the performance of that action.

Typically, accounts of acting (or intending) for a reason are not content to analyze it merely in terms of the basing relation presumably because the basing relation might be thought conceptually too close to the idea of acting for a reason.[58] My interest lies elsewhere, however, and so I defer

[55]The view that one's motivating reason is the feature in the light of which one acts derives from Dancy (2000, 128-129).

[56]Note that on another way of characterizing anti-psychologism about reasons, also due to Dancy (2000, 14), it is roughly the thesis that one cannot give a constitutive explanation of acting for a reason only in terms of agents' psychological states. AFR is not anti-psychologistic in this sense.

[57]Note how perfectly acceptable sentences like "I saved the cat *because it was drowning*" and "I took the 5:38 train *to get to Boston by 10*" are. The most recent "anti-psychologist" wave to wash over views of motivating reasons was initiated by Dancy (2000, Ch. 1); see also Alvarez (2010, Ch. 5).

[58]See for instance Wedgwood (2006); Arpaly & Schroeder (2015), who offer causal accounts of the basing relation.

to those who study the basing relation — though I do believe, as with most epistemologists,[59] that the basing relation is a kind of causal relation, and I will rely on this feature in responding to a potential objection below.

It is important to note, however, that moving to an anti-psychologism about reasons does force us to reinterpret slightly the basing relation, which is often defined in epistemology as the relation which holds between a reason and a belief just in case the reason is a reason for which the belief is held (Korcz 2015). Many of the same considerations that motivate anti-psychologism about reasons for action will also motivate the same about reasons for belief: when asked to justify my belief that Hamish is no true Scotsman I may explain my reasoning by pointing out that he just put sugar in his porridge, a thing that no true Scotsman does. Here the reasons I cite are not my own beliefs but facts, at least as I take them to be. But then it would be odd to hold that my reasons for this belief are bases for it, for bases of beliefs are supposed to explain them. I may well be wrong in thinking that no true Scotsman puts sugar in his porridge. But how, then, could a non-existent fact or a false proposition explain my believing that Hamish is no true Scotsman? Instead it seems that my bases for believing this about Hamish are my *beliefs* that Hamish just put sugar in his porridge and that no true Scotsman does this. Thus, the usual characterization of the basing relation was right about what states enter into the basing relation — they are psychological states (as well as, on the present view, actions) — but wrong to characterize bases as motivating reasons, that is, as reasons for which a belief is held or action performed. The motivating reasons are the considerations *represented* by the bases of (say) beliefs and actions.

4.6.1.2 Anthropocentrism

It is plausible that AFR follows from HLA, in which case it is a commitment of the present view. HLA further holds that in humans the appearance required by AFR for action for a reason must be an *affective* representation, and in particular an affective representation of a *normative reason*. Why might this be the case? Some aspects of this question will be addressed below, e.g. why we should think action in humans is based on a representation of a normative reason as opposed to a representation of a valuable* object or state of affairs (§4.6.3). Why it might be that affective representations are required can ultimately only be answered after a thorough investigation into our psychological architecture. But a more fundamental issue the question raises is why in humans, but not in other possible agents such as the Wobbly, there should be a special concern with normative reasons, or with practical reason generally.[60]

[59] See Korcz (2015) for an overview.

[60] One way of responding to this question is to hold that these other possible agents cannot *reason*. As I mentioned in my description of the Wobbly (p. 103), I think the Wobbly does engage in practical reasoning even though his reasoning is not about his normative reasons. One view of reasoning that would allow this, and to which I am partial, is that reasoning is an approximately rational transition from thought to thought with the aim of coming to a view of

One combination of views that could explain this is an anthropocentric picture of the nature of practical reason together with the view that, although AFR sets necessary and sufficient conditions for action for a reason, it is also true that agents act under the guise of the kind of normativity appropriate to them. One might reasonably hold that there is quite a difference between the standards of theoretical reason and those of practical reason. The standards of theoretical reason — common proposals include never to fully believe a proposition and its negation, always to apportion one's beliefs to one's evidence, always to update one's credences by conditionalization, and probabalism, the view that a rational agent should have probabilistically coherent credences — arguably reflect relatively fundamental features of the world's structure. One should not believe p and also believe $\neg p$ because it could not be that $p \wedge \neg p$. One should not have .6 credence that p and .5 credence that $\neg p$ because one is thereby guaranteed to have misallocated one's credences relative to what is possible.[61]

But, the story might continue, the standards of practical reason do not answer to the world's structure. The fundamental fabric of the world is bereft of value. On the anthropocentric view, practical reason answers to the needs and nature of *human* beings.[62] Non-human creatures may have only reason-like schnormative schmreasons, or nothing like practical norms applying to them at all, although they act for motivating reasons. Such a view might come in response-dependent and response-independent varieties. A response-dependent version would hold that the standards of practical reason are derivable from human attitudes under certain conditions. An example of a response-independent version would be the Aristotelian view that the standards of practical reason are grounded in the form of human agency or in the human function, which is in turn grounded in human nature. HLA could fall out of such an anthropocentrism about practical reason together with the general view of action for a reason that agents act only under the guise of what is derivable from their idealized attitudes, or of what is consistent with their nature.

For the purposes of illustrating a contrast with the Aristotelian defense of the Standard View in response to the Wobbly Argument (p. 107), consider what an Aristotelian anthropocentrism might look like. On this view the standards of practical reason are not fixed for any possible agent by *that agent's* function but instead always by reference to the *human* function. Even if the Wobbly's function is to cast towels off of cliffs, this is irrelevant to the determination of its

what to do or of how things are. This view leaves open that one can reason about what to do without that reasoning's taking on a particular content, e.g. that it be about one's normative reasons.

[61] Of course all manner of views of the nature of theoretical reason are defensible, including pragmatist and deflationary ones. I mention this relatively robust realist view simply to emphasize that anthrocentrism about practical reason attracts even if it is true.

[62] For comparison, note that some such picture motivates the sentimentalist project broached in D'Arms & Jacobson (2000). However their project could be more cautiously put as the thesis that *in humans,* normative judgment and values are grounded in the sentiments. It need not be interpreted as a claim about practical reason as such, as I intend the anthropomorphic thesis to be.

normative reasons. Either it has no reasons or what it has reason to do is determined by what a human in its circumstances would have reason to do. This would be a special fact about normative reasons, since on this theory the Wobbly does have genuine motivating reasons.

What might explain this difference between motivating and normative reasons? One response might be that motivating reasons belong to a theory of the *explanation* of action whereas normative reasons, in virtue of their normative force, belong to *ethical* theory. Offering a motivating reason for one's action to someone else is a matter of explaining one's action to them: it is to offer the consideration in light of which that action seemed to be a doable thing, and one only acts for a reason if having such a consideration in mind explains one's action. But offering a motivating reason that one also takes to be a normative reason for having so acted is an attempt to offer a justification for one's action, which is a fundamentally different kind of act. (Indeed I speculate that the publicity thesis of Chapter 3 would lend further coherence to such a view. If the standards of practical reason are publicly shared, then our judgment that an agent has a normative reason for acting commits us (*pro tanto*) to tolerating their action to some degree. If we think of a particular hypothetical agent — especially one with a psychology radically unlike ours, for instance one which is incapable of any affective states — that we are not committed to tolerating to any extent the actions they perform in furtherance of their projects because such robot-like beings do not matter ethically, then we should rationally think that the reasons for which they act are not normative reasons.)

Of course, laying out and defending any such broad metaethical view of action and normativity would require quite a substantial defense, and it is not my project to embark on that here.[63] The goal of this subsection was merely to shine a light ahead and to motivate HLA by situating it within a sketch of a view that may be attractive to many.

4.6.2 Affective states

The ideas that emotions are representations of values, or at the very least responsive to them, and that they have a prominent role in explaining action are not new. Indeed, these positions represent the new orthodoxy in thinking about the emotions in philosophy and psychology, though it is a broad theme with many variations. In philosophy emotions are held to be key to our value*- or reason-tracking capabilities (Benn 1988, 83; Jones 2003), to be essentially evaluative or involving

[63]To highlight one potential cost of this view, some might object that it amounts to a kind of species-centrism in which we project our own standards onto other agents and fail to recognize the possibility that the different natures of other possible beings can ground standards that have normative force for them. After all, we surely admit of interpersonal variation in what is good* for humans; why not admit of interspecies variation in what an agent has reason to do? This question may still resonate even when it is remembered that HLA is officially taken to be true not just of actual human action but of actual animal action, insofar as animals act for a reason. For a vivid depiction of a worry along these lines, see Railton (1998, 71-76).

appraisal (Goldie 2000; Helm 2001; Seager 2002; Deonna & Teroni 2012; Ellsworth 2013) and are sometimes thought to be components of a perceptual or perception-like faculty of sensitivity to values (McDowell 1985; Wiggins 1998; Tappolet 2000; Johnston 2001; Roberts 2003). Emotions are sometimes also thought of as action tendencies (Frijda 1986) or calls for reprioritization of goals and actions (Simon 1967; Carver & Scheier 2013; Scarantino 2014), and are often also thought to have a role in initiating, guiding, and explaining action (Zhu & Thagard 2002; Döring 2003, 2007). In psychology, affect is often seen as a representation of value in multiple domains, including epistemic uncertainty (Behrens et al. 2007) and trustworthiness (Behrens et al. 2008) as well as reward and punishment values of outcomes of actions (Rolls & Grabenhorst 2008).

Many questions remain about the nature of affect and its relation to value. Is the relationship between affect and motivation constitutive or causal? Do affective responses to stimuli often precede cognitive processing about their nature (the "affective primacy" hypothesis, Zajonc 1980, 2000)? Are affective states mere responses to values, trackers of value, representations of them, or perceptions of value? What is the relation between affect and emotion?

For the purposes of setting out HLA+E I need only address a few of these issues. The first issue that needs addressing is the meaning of 'affect'. There are roughly three different ways to characterize affective states. The dominant conception in psychology defines it with respect to the affective system: it is a state that excites a system of coordinated mental and behavioral responses that prepare us for action, or a state that plays a special role in such a system (Matsumoto 2009, definition 1; see also Zajonc 1980, 2000; Winkielman et al. 2005). Others see affect as a state that carries a certain kind of information (Neumann et al. 2003; Bargh & Morsella 2008), especially evaluative information (Schwarz & Clore 1983;Clore & Huntsinger 2007; Matsumoto 2009, definition 2); and occasionally its phenomenological aspect, on which affective states are feelings, receives recognition too (Panksepp 1998; Matsumoto 2009, definition 2).

These poles pull apart. It is commonly thought that there can be unconscious states, or at the very least causes, of the affective system, and it is sometimes thought that these states may be carriers of evaluative information (Clore & Huntsinger 2007). But phenomenal states are always conscious states, for phenomenal consciousness is a major notion of consciousness.[64] And while it is entirely possible that an affective state carries information in virtue of its causal connections within a motivational system, it is also possible that the fact that these states carry information is determined independently of their use in a motivational system, as on covariation theories of representation.[65] Or instead these states may function instead as inputs to that system (perhaps

[64]Other important notions of consciousness include reflexive consciousness and access consciousness. See Block (2003).

[65]An example of a covariation theory is Cutter & Tye's (2011, 91) tracking theory of intentionality: "Tokens of a state S in an individual x represent that p in virtue of the fact that: under optimal conditions, x tokens S iff p, and because p."

they inform and guide motivation) or instead as outputs (perhaps they are inputs into an interpreter module, as in Gazzaniga 2000, that makes sense of action after it has been selected).

If these three characterizations do not largely pick out the same states, the worry is that an attempt to cash out a GG thesis in terms of affect's being a guiding representation of the good illicitly trades on the ambiguity of the notion. The first characterization makes it easier to secure a connection to motivation, the second to representation, and the third makes it more plausible that neither of the first two connections are largely true by definition. But the advantages are illusory if it turns out that no one kind of state plays all three roles.

To avoid this problem, I will continue to employ the phenomenological notion of affect that I used in Chapter 2. A state is affective if it is phenomenologically valenced: it is a *good* feeling or a *bad* one, in the way that some visual sensations are *red* and others are *blue*. An *affect*, then, is a state from which non-affective components have been abstracted. Pride, for instance, combines a warm feeling directed at some person or event together with a belief about that object's connection to oneself, especially one's own efforts. Pride is an affective state, but since the belief it involves is not affective, the warm feeling, as considered apart from the belief, is an affect. But affective states need not be valenced. Surprise is generally considered affective even though it need have no determinately positive or negative valence. Intentionalism might have a distinct advantage in being able to offer a general phenomenological characterization of affect as a feeling *of significance*, with "significance" taking on its evaluative and not its semantic sense.[66]

This phenomenological characterization of the state on the basis of which we act also has the advantage of sidestepping questions about the nature of emotions and their relation to conscious feelings. Although of course many affective states are emotions, HLA makes a claim about *affective states*. This not only makes the thesis more precise than most extant proposals, which concentrate on the prospects for connections between emotion and action, but it opens space for action to be based on affective states which are at best ambiguously emotions, such as moods, the brief flashes of affect which are hypothesized to play an informational role in conscious reasoning (Dutton & Aron 1974; Schwarz & Clore 2003; Schwarz et al. 2005), temperament, and sensations such as the primary unpleasantness discussed in Chapter 2.

HLA does require that action be based in a conscious feeling — a *phenomenally* conscious feeling, not necessarily a reflexively conscious feeling, but a conscious feeling nonetheless. One might reasonably wonder whether it is wise to work with this particular notion of affect, for two reasons. First, it is often seen as a contingent, occasional, and possibly late-evolved feature of affective states that they are conscious or that they involve feelings. Appraisal theorists of emotions, for instance, hold that emotions are processes which begin with automatic appraisal of the situation

[66] Another interesting possibility, suggested to me by Uriah Kriegel, is that affective states could have a unitary characterization as experiences of *being moved*.

which in turn causes a particular "affect program" of nervous system changes, facial and skeletal muscle changes, action preparedness, and recall of appropriate memories. Conscious feeling only comes at the end of this process, if it comes at all.[67] From the perspective of the function of affective states in appraising situations and preparing the body for appropriate action, whether such states are conscious is not manifestly important.[68] Second, recall that intuitively, the Wobbly does act for a reason even if affective states were nowhere mentioned in its description and thus seem not to be required. How could it be that *conscious* affective states are required in human action for a reason even if affective states are not required for action for a reason as such?

The answer to both questions is that it is quite reasonable to think there is a tie between action for a reason as such and conscious states even if affective states are not necessarily conscious. As we saw above, AFR requires that in order for an action to be done for a reason it must be based on a state that presents some further condition to the agent, and completely unconscious states cannot be candidates for this state because they are not appearances *to the agent*. (The claim that human action must be based on phenomenally conscious affect is then licensed by the fact that it seems improbable that there are affective states that are conscious but not phenomenally conscious.) And AFR requires appearances of this sort for good reason, since an appearance that *I* do not have access to cannot be a basis for *my* action. Take a rather famous example from David Velleman (1992b, 126-127):

> Suppose that I have a long-anticipated meeting with an old friend for the purpose of resolving some minor difference; but that as we talk, his offhand comments provoke me to raise my voice in progressively sharper replies, until we part in anger. Later reflection leads me to realize that accumulated grievances had crystallized in my mind, during the weeks before our meeting, into a resolution to sever our friendship over the matter at hand, and that this resolution is what gave the hurtful edge to my remarks. In short, I may conclude that desires of mine caused a decision, which in turn caused the corresponding behavior ... But do I necessarily think that I made the decision or that I executed it?

Velleman leads us to believe that it was not him but his subconscious resentment who did these things, but we need not here take a stand on whether *he* severed the relationship; even if he did, the question that concerns us is whether he did it for a reason, and what that reason was. And it seems to me that if his resolution were *completely* unconscious, then it could not have been a basis for his outburst even if it were a cause of it, and even if its constituent utterances were intentional under some other description.

Contrast this with a case in which a person is motivated by a state with conscious manifestations even if they are not aware of it as such at the time. Lauri and Burnley have long been friends, but

[67] See Ekman (1977, 55-59); Griffiths (1997).

[68] LeDoux (1998, 2012) has been particularly effective in advancing this perspective.

lately Burnley's new girlfriend has caused a strain in their relationship. Lauri begins to resent Burnley's absences from their traditional pub trivia games and comes to resent the girlfriend too, thinking that she's no good for Burnley. So she begins a deliberate campaign to insert herself more into Burnley's life, just to show him what a good friend can do for him. It's only later that she realizes that all of this is explained by the fact that she loves Burnley. Here it is more intuitive to say that Lauri's love of Burnley is indeed a basis for her deliberate campaign of interference — she did it *for love of* Burnley — but note as well that her love has conscious manifestations in her resentment of the new girlfriend and her conscious feeling of attachment to Burnley. So affective states with conscious manifestations are indeed conscious enough to form bases of actions done for a reason.

I grant that intuitions may differ on the necessity of conscious access to our reasons for acting. There is, for instance, a long tradition of Freudian psychoanalysis that aims to uncover completely unconscious motives for many of our actions. My suspicion is that the clearest of such cases will indeed involve conscious manifestations of affective states but in considerably altered form, just as in the case of Lauri and Burnley. One may, for instance, displace one's conscious anger at one's father onto a friend, but the form that displacement will take is conscious anger, or some other conscious state. But if my suspicion is false, then a theory of action for a reason that allowed completely unconscious states to serve as bases for action would need to show that such states count as *one's own* motivations in the relevant sense. Surely we judge that someone driven to murder while sleepwalking didn't *really* murder anyone.[69] She can honestly claim to have had no agency in the deed. How can acting on completely unconscious motivations be different?

4.6.3 Reasons

Whereas most extant representational views of affective states hold that they represent values*, HLA holds that they represent reasons for action. There are exceptions to the dominant emphasis on values* (cf. Arpaly (2000), Jones (2003), and Brady (2013, 110), who emphasize the responsiveness of emotions to reasons for action), but among those who consider affective or emotional states to be representational, the view that they represent reasons has, as far as I am aware, only been defended by Jennifer Hawkins (2008).

I'll first present Hawkins' view, for the purpose of drawing a contrast with my own. From there I argue by analogy to vision that affective states should be construed as possessing refined non-conceptual content concerning reasons (§4.6.3.1). I go on to refine the view of unpleasant pain presented in Chapter 2 by arguing that unpleasant pain, too, should be understood as representing

[69]The most famous such case is that of Ken Parks, who while asleep drove 14 miles to his in-laws' house and stabbed his mother-in-law to death. See *R. v. Parks*, [1992] 2 S.C.R. 871.

reasons for action (§4.6.3.2), and close the section by showing why affective states are not best understood as having non-conceptual *evaluative** content (§4.6.3.3).

Hawkins considers desires to be one type of affective state and ultimately casts her view as holding that desires represent their content as good. Indeed her overall goal is to propose that the affective nature of desire is the key to understanding how the desires in both small children and adults can be evaluative*. But it turns out that the core of her view is that affect represents something a little more ought- or reason-like. On her view affect presents certain actions as *making sense* (*ibid.*, 259-260), or as *feeling right* (259). She also characterizes it as an experience as of there being a reason to act that way (257, 259).

But crucial to her account is how these actions are thought to make sense via an experience that is, in a way, primitive (258-259). Toddlers are not thought to desire to φ by making the explicit judgment "φing feels right", but by having a *proto-thought* employing *proto-capacities* roughly to that effect. Here she takes inspiration from the work of José Bermúdez (1998), who notes that although infants cannot be attributed anything like a mature concept of an object — they cannot think about objects in abstraction from current experience, nor can they stand back from a present experience and ask whether there *really* is an object as there seems to be — they do, as they mature, exhibit a growing sensitivity to object properties such as having a stable shape, being impenetrable, etc. Hawkins' proposal is that even as adults develop mature normative concepts they retain a semi-autonomous proto-conceptual affective system. "[A]ffect is primitively structured in ways that allow for affective presentations *as of* the world being a certain way", she writes. "These presentations would not be conceptual, but at most 'proto-conceptual'" (Hawkins 2008, 258). On Hawkins' view the content of affective states is also evaluative* because she also offers either a fitting-response analysis of value* or a buck-passing one on which an object is good* just in case there is a reason to respond positively to it in certain ways (*ibid.*, 257).[70]

4.6.3.1 Refined non-conceptual reasons content

I think Hawkins' view is an excellent start, but it needs amendment in crucial respects. In this subsection I argue that affective content should be understood as refined non-conceptual reasons-content, not proto-conceptual content about what responses make sense.

But we should first get clearer on the contrast I mean to draw here. A common way of defining non-conceptual content is Michael Tye's (2009, 103):

> [A] visual experience E has a nonconceptual content if and only if (i) *E* has correctness conditions; (ii) the subject of *E* need not possess the concepts used in a canonical specification of *E*'s correctness conditions.

[70] Hawkins does not seem to intend to offer a disjunctive theory, but she clearly mentions both possibilities. Obviously the two are equivalent in case one analyses normative reasons in terms of fittingness.

I would prefer to amend condition (i) to read: (i′) E has correctness conditions that are presented to or conveyed to the subject. Siegel (2010, 42-43) points out that (i) alone would not suffice for content. For any experience E, E is correct iff "The world corroborates E" is true. But nothing about corroboration need appear in the content of an experience. What is missing, it seems, is that the correctness conditions must be conveyed or presented to the subject. Condition (ii) also stands in need of clarification. As Tye uses it, it would better read: (ii′) it is not the case that the subject of E needs to possess all the representations (notably, concepts) used in a canonical specification of E's correctness condition. So put, (ii′) does not presuppose that if an experience has content then it will have a canonical *conceptual* specification of that content. This enables us to more cleanly make a distinction Tye himself wants to make (*op. cit.*, 104): we can say that an experience has *refined* non-conceptual content if the canonical specification of its content is completely conceptual and that it has *rough* non-conceptual content if it is not completely conceptual.[71]

As Hawkins understands proto-conceptual states, they have rough non-conceptual content. Of course in explaining it she deploys concepts (of fittingness, making sense, reasons, etc.), but since these proto-conceptual capacities are hypothesized to be severely limited and indeed hobbled in comparison to our conceptual capacities, it would seem that any attempt to describe their representational outputs by deploying concepts would merely be a best attempt to describe what is too primitive to be put into words. This, at least, seems to be the best way of understanding Bermúdez's proposal concerning infants' proto-conceptual capacities to recognize objects. It is not that babies recognize *objects* as such, although they have no (conceptual) thoughts about objects. Rather the proposal is that at a given point in their development they recognize some object-like features and not others. Once this is clarified we see that proto-conceptual content is not best understood as refined non-conceptual content, for there is no reason to hold that if we *could* find or devise a canonical specification of proto-conceptual content, its expression on our part would require capacities that infants do not possess.

And it seems that Hawkins means to follow Bermúdez here, for as she emphasizes she uses the word 'fit' to "gesture at a very simple kind of experience, a primitive feeling *as of* certain responses *making sense* or *feeling right*" (*op. cit.* 259). But note that such phrases could adequately well characterize a mere sense of a doorknob as to-be-turned, or as there for turning. It thus appears that Hawkins' proposal is that affect presents actions in what we have been calling a weak normative light, not that they represent normative reasons.

This raises a serious question about how the proto-conceptual content of an affective state interacts with its conceptual content, especially insofar as affective states can be the basis of normative beliefs. Most theories that hold that emotions are representations of values are also perceptual

[71] Tye distinguishes among three options of which one, which he calls "robustly non-conceptual content", corresponds most nearly to rough non-conceptual content.

or quasi-perceptual theories. A perceptual theory of emotions holds that they *are* perceptions. A quasi-perceptual theory of emotions holds that they are perception-like representations, or that there is often a close analogy between the justifying role played by perception in sensory judgment and a hypothesized justifying role played by emotions in normative judgment.[72] And one of the key advantages of quasi-perceptual theories of emotions is their ability to provide an epistemology for normative beliefs. If emotions are quasi-perceptual representations of value they can be non-inferential bases of beliefs about the values they present, but it is hard to see how a vague, proto-conceptual representation of something value-like can justify a belief about a value absent an inference from additional premises. But this problem is really symptomatic of a deeper issue. Many views of perceptual content hold that it is significantly more fine-grained than quotidian conceptual content, since we can discriminate among many more shades of color than we typically have concepts for; natural language certainly does not have terms for all the different hues among which we can discriminate.[73] Is affective experience then *less* fine-grained than our everyday evaluative concepts?

Even if it is admitted that our affective experiences are overall less determinate than our visual ones, some intuitive considerations point in the direction of a negative answer: though you may feel a relatively determinate amount of pain in your finger, you likely have no concept for the exact degree of badness your unpleasantness represents. Mark Johnston (2001, 181) forcefully argues that affective experiences are of "utterly determinate evaluative[*] properties". A child may be able to genuinely and fully appreciate the (pretend) *exquisite refinement* of her make-believe tea party without being able to conceptualize her experience as such until much later in life.

But we might do better by drawing a more explicit analogy to visual representation.[74] How might we argue that there is refined non-conceptual representational content in visual experience in the face of a challenge? We can begin with an influential thought experiment from Peacocke (1983, 12-13): Imagine standing on a road that stretches straight out in front of you to the horizon. Just off the right shoulder of the road ahead are two trees which form a line parallel to the road. Intuitively, your visual experience represents them as being the same height. After all, he writes,

[72]For perceptual and quasi-perceptual theories of emotions, see Tappolet (2000); Johnston (2001); Prinz (2004b); Döring (2009); Montague (2014), among others. Cf. Brady (2013, Ch. 2). Note that most criticism of these views has focused on the question of whether emotions are perceptions, not on the question of exactly what their representational content is. For criticism along the former lines, see Salmela (2011); Deonna & Teroni (2012, 2014); Dokic & Lemaire (2013, 2015); Brady (2013); though also see Schroeter et al. (2015, 371-372), who do raise the issue.

[73]See for instance Evans (1982, 229); Peacocke (1992, 111); Heck (2000, 489-490). Note that I am using "fine-grained content" in a slightly different way than Tye does. Tye (2009, 39) uses it to describe contents that differ even when they correspond to the same property or possible world, as with the contents of "coriander" and "cilantro". Here I use it merely to talk of a contextually-determined level of determinacy of content.

[74]Note that I am only insisting on an analogy between the contents of visual experience and affective states. I advocate quasi-perceptualism about affective states, and take no stand on whether emotions are in fact perceptions. Note as well that a quasi-perceptual view of emotions is consistent with denying perceptualism, the thesis rejected in the previous chapter that appearances of the good *always* have the force to rationally support evaluative beliefs.

that is what you'd be inclined to judge merely on the basis of having seen them. The nearer tree also takes up a little more of your visual field than the further one. But surely we do not want to say that your experience thereby represents the nearer tree as being bigger than the further one, for we do not want to say that your experience has inconsistent contents. Given this, is there any non-mental property that this latter aspect of your experience represents? If not it would be a problem for intentionalism, which is committed to there being no non-representational aspects of phenomenal experience. Michael Tye (2003, 15-16) responds that although the trees are represented as having the same *viewpoint-independent* size, the nearer tree is represented as being bigger *from the perceiver's viewpoint*. He holds that this latter amounts to the nearer tree's being represented as subtending a larger angle relative to the viewer.

I think it's possible to improve on Tye's response, since the angle subtended by a particular object does not mark it out as having a particular *size* relative to the viewer but as existing somewhere within a cone projecting out from the viewer.[75] But note that if the phenomenal feature in question does represent a subtended angle, the subject of the experience does not thereby need to possess an understanding of geometry. Thus this is an excellent candidate for a state with non-conceptual representational content, for there is a clear sense in which "angle subtended by the tree relative to here" is nevertheless an apt description of what is being presented to the subject. Conditions in which there turns out to be no tree at all within that cone correspond very well to conditions in which you, as the subject, would regard your experience as inaccurate. That is to say, a good explanation of the particular disposition to be surprised by various treeless conditions is that you possess a state with precisely the hypothesized representational content. Furthermore upon learning geometry you would, if the analysis is right, be inclined to judge explicitly that that aspect of your experience and "angle subtended by the tree" pick out roughly the same features of the environment, and that the experience is correct just in case the angle subtended by the tree is roughly so-and-so. This suggests the following rubric for when to postulate that C is the content of experience E:

33. Our dispositions and behavior can be well-explained by postulating that E represents C. In this case, relevant dispositions include our penchant to be surprised if there is no tree to be found within the cone and to unreflectively avoid walking into where the tree is represented as being.

34. That E represents C matches with our experience of it. Upon mastering the relevant concepts we would be inclined to judge that E picks out the C-features of the world, and that E is correct just in case C obtains or is instantiated.

[75] Another difficulty raised by his response is that it puts one's own point of view in the content of one's experience, which seems not to be the case. Thanks to Sarah Buss for pointing this out to me.

35. We would also be inclined to trust E, in predictable ways, in accordance with our overall evidence on whether C, and to downregulate E — that is, attempt to modulate or extinguish it — insofar as possible in the case that we distrust it.

Now we can apply this rubric to affective states. Looking at the swaying plank footbridge in front of him, Fearful Dan cannot help but feel something about the bridge that inclines him not to walk across it; he inwardly recoils. Dan, we might say, non-conceptually represents the (apparent) precariousness of the bridge as a reason not to trust it for walking. This explains why he's inclined to avoid the bridge, or at least to reopen the question of whether to cross it (33); supposing he knew that the cables holding up the bridge were very strong and extraordinarily unlikely to break, he would not only regard his feeling as an unfortunate hindrance but as *getting it wrong*, for in this case he knows better than his gut (34). And depending on what other evidence he had about the bridge he would be inclined not to walk across it, or to quash his fear as far as possible (35).

Furthermore there is reason to think that his fear represents a *reason* not to walk on the bridge, as opposed to something reflective only of weak normative light, such as a "default pathway" (Hawkins 2008, 260) of not walking on it, or the potential non-walkability of the bridge. Unlike affordances and other perceived pathways for action, reasons have a role in practical reasoning and the rational control of behavior. Reasons for action are considerations that practical reason *requires* one to take into account when deliberating, or at the very least positively *permits* one to. And it does seem that if Dan's experience is veridical, then its content is something he is rationally required to take into account. So it seems that if his fear is representational, it represents not only the possibility of not walking across the bridge, but a requiring reason not to walk across the bridge.[76] This, then, is excellent preliminary evidence that affective states can represent reasons for action.[77]

4.6.3.2 The truth that motivates imperativism about pain

We can present further considerations in support of affective states' representing reasons at the same time that we add a crucial refinement and clarification of the view of unpleasant pain presented earlier in this dissertation. Although Chapter 2 defended *evaluativism** about pain, as we

[76] This is not to say that it is a decisive reason. On the distinction between requiring and justifying reasons, see Gert (2007).

[77] It should be noted that some recent critics have held that perceptual theories of affect and the emotions need first to establish the explanatory necessity of postulating that emotions have evaluative content (Schroeter et al. 2015). Such a vindication could not feasibly be attempted here, and I am moreover skeptical that it is required because I doubt that the representational properties of any system are strictly necessary to explain its operations, when these operations are described without reference to representational properties. Here I have started with the position that affective states have representational content and asked what that content might be. But this is already a step forward with respect to current quasi-perceptualist theories of affect, which have thus far been content to raise the possibility that affective states have non-conceptual content (Tappolet 2000; Hawkins 2008; see also Oddie 2005, who raises the same possibility about desire).

noted in the Introduction, the hypothesis that unpleasant pain represented a specifically *evaluative** property was chosen largely out of convenience: the arguments given in its favor in that chapter would equally well support any broadly evaluative theory on which negative affective states represent some *pro tanto* justification for aversive behavior, but there already exists a literature on evaluativism* about pain. But the view that unpleasant pain represents *bad** bodily perturbations does have the unfortunate effect of obscuring the strong relation between pain and action. It would better fit the biological function and experience of unpleasant pain to say, not that it represents the monadic property of badness*, but that it represents a relation between a bodily event and a possible action.

Although I do not ultimately support imperativism about pain — the thesis that pain sensations are commands to perform certain actions — I think it is right to emphasize pain's relation to action. In a recent defense of imperativism Colin Klein (2015, Ch. 3), in large part taking after the clinical researcher Patrick Wall, is impressed by the heterogeneity of pain stimuli and the variation between stimulus intensity and the strength of pain sensation. These factors lead him to hold that pain sensations could not represent some context-independent property. But pain *does* seem to motivate the same kinds of behaviors across contexts. Pain episodes often begin with a sharp pain accompanied by a quick withdrawal reflex and and an aroused, defensive state. If the injury is deep into the tissue or serious it will often be followed by a dull, persisting pain that can last from minutes to months. This second stage is typically accompanied by listlessness, irritability, and even depression.[78] Therefore the function of pain, Wall (1979, 264) suggests, is to motivate the agent to avoid further harm (especially in the first stage, presumably) and to enable recuperation and recovery (presumably the second stage). For his part, Klein (2015, 28) takes pain to be a homeostatic signal that functions to protect the body's physical integrity, particularly after injury.

Klein argues that this thesis about the function of pain supports an imperativist account of the nature of pain. It would take us too far afield to consider his detailed argument, but even if intentionalists ought to reject it, it seems clear that they also ought to recognize the strong grounds that Wall and Klein present in favor of thinking that there is a cohesiveness to the class of actions pain motivates us to perform: they aim to *protect* the (apparently) injured body part. Indeed they motivate us to protect it in very particular ways: by withdrawing it, covering it, not putting weight on it or manipulating it, etc.

Furthermore, this aspect of pain seems to be represented in normal pain experience. It would be truer to biological function and to experience for the intentionalist to hold that our unpleasant pain represents a bodily alteration (through the pain) as being a reason for a protective action than it

[78]Wall (1979). Part of the explanation of the persistence of pain long after healing has begun seems to be that a barrage of afferent C-fiber firing causes a long-term chemical change in the dorsal horn of the spinal cord, one of the effects of which is to increase the sensitivity of nociceptors. See Melzack & Wall (2008, 99-107).

would be to hold, as evaluativists do, that it represents the alteration as bad*. For there will likely be many ways that one could conceivably avoid badness*, very few of which are represented in every normal pain experience. For instance, David Bain (2013) argues that unpleasant pain experiences represent badness* "in the bodily sense". But there are many things that are bad for bodies and nearly as many ways to avoid or dispel that badness. If I am in pain, would it make sense to me to seek cover, or to vomit? Generally not. Bain and other evaluativists* could attempt to specify more precisely the disvalue* in question so that it becomes more obvious what types of actions are associated with avoiding it; they could perhaps say that pain signals badness* for the bodily tissue. But the same problem will likely arise again. Just to pick one kind of tissue, some skin conditions, or bad ways for the skin tissue to be in, merely require the application of a layer of ointment while others may require surgery, and there seems to be no class of actions that is immediately singled out to us as eligible when we learn that our skin is in a bad* way.

It seems that we will not capture pain's intuitive connection with action until such eligible actions, protective actions, are held to be represented as such in the pain experience. And it seems that the kind of content that can capture that is reasons-content: unpleasant pain represents an alteration in the body as a reason to protect that bodily part.[79]

4.6.3.3 Why affective states do not represent evaluative* properties

If we are being careful we should note that the last subsection does not give us an argument that evaluative* properties are *not* represented in normal pain experience. It merely aims to show that thinking of pain as representing a bodily part as instantiating a monadic evaluative* property does not suffice to capture its representational content, for it leaves out the relation of this bodily state to determinate types of aversive action. Nevertheless I do think there are decisive reasons to think that values* are not represented as such in affective experience.

Philosophers have typically motivated a connection between emotions and values* by pumping the intuition that the fittingness or correctness conditions of an emotional state correspond exactly to the presence of a certain value*. (For the most part it does not matter for the present argument whether these are fittingness conditions or correctness conditions, but for the sake of brevity and because the present account is committed to there being correctness conditions for affective states, that is the phrase I will use below. For one complication, see note 85 below.) One's fear of a

[79]In Chapter 2, I mentioned the possibility of unpleasant experiences without pain sensation. If that is indeed possible, then it seems the present theory would predict that such an experience would represent *there being reason* to protect a part of one's body without representing *a particular reason* to protect it. Indeed, I think it is a good idea to leave this possibility open for other affective states that are traditionally seen as non-intentional, such as moods. (See Fogal 2016 for an excellent linguistic and philosophical analysis of the distinction between "reason" in the mass-noun and count-noun senses.) Do note that accommodating this possibility would complicate the interpretation of AFR and its motivation, since we would then need to take action for a reason to include the case in which an agent acts for no *particular* reason but instead acts for the reason (in the mass-noun sense) she takes there to be for that action.

snarling dog is held to be correct just in case that dog is *dangerous*; one's amusement at a joke is correct just in case the joke is *funny*.[80] But in the general case there is a problem with the right-to-left direction of this biconditional, even if we acknowledge that correctness conditions for a mental state need not correspond to conditions in which one has reason to be in that mental state or in which one ought to form that mental state. (Correctness for full belief is truth, but it is not the case that I ought to believe any and all true propositions. Some are trivial and some I have no evidence for.) For the correctness conditions of an affective state depend not only on the presence of the value* but on the circumstances in which the state is realized.

Some dogs are just dangerous. But a dangerous dog does not become less of a dangerous dog the further I run from it. It will *pose* less of a danger to me, but I do not sap its dangerousness from it by putting distance between us. If I receive a call from a friend asking me what to do about a dangerous dog in front of him I could not very well console him by saying that there was nothing to fear because *I* was not near enough the dog for it to be dangerous. But it is *not* correct for me to fear a dangerous dog if I am at a safe distance from it. (Do note, however, that the further I run from the dog the less reason I have to keep running from it, which corresponds well to the lesser degree of fear it is correct to feel of the dog. So the thesis that affective states represent reasons to act holds up well to this line of thought.)

Philosophers will no doubt seem to sense some linguistic sleight of hand in this argument and wonder if a contextualist or relativist semantics for value-terms will save the evaluativist's position. But each of these options bring up well-known problems. A contextualist solution to the problem would require indexing 'dangerous' by person and place (and perhaps time): when I assert "This dog is dangerous", upon semantic analysis my assertion will turn out to mean, "This dog is dangerous relative to me and to where I am." But this view faces a problem with disagreement: if you and I are standing before a snarling dog and I say "That dog is dangerous" while you say "That dog is not dangerous", we seem to be semantically guaranteed to be disagreeing. But on the contextualist analysis that is not so, for we are different people standing in slightly different locations, and so we may both be speaking truly and not disagreeing. This view has the additional problem of multiplying values beyond recognition. On this view there are no dangerous dogs or funny jokes, *tout court.* Instead there is dangerousness relative to any given person and place, and funniness relative to any given person and time.

A view on which my utterance's truth-value is determined relative to a context of assessment (MacFarlane 2007) would seem to resolve the problem of disagreement, but at the price of another problem. If we are to hold on to the claim that it is correct to fear a dog just in case it is dangerous,

[80]See for instance de Sousa (1987); D'Arms & Jacobson (2000); Tappolet (2015); and many more. Do note that the equivalences are often stated with predicates like "fearsome", "amusing", or "terrifying" on the right side. The difference is significant, and we will return to it below.

then relativism about assertions of the dog's dangerousness would also be transmitted to assertions about the correctness conditions for fear of the dog. It appears there would be no context-of-assessment-independent answer to the question of what the correctness conditions of my state of fear of the dog are. This seems to be all but impossible to square with a naturalistic account of representation. To consider once more the much-belabored analogy with perception: what determines the correctness conditions of my visual experience as of there being a mesa in front of me are the function and/or causal history of my perceptual system, not this *and* the context from which I am assessed.[81]

A sensible-seeming response to this problem would be to move from talk of the dog's *dangerousness* to its *dangerousness for me*, from an absolute to a relational notion of the value in question. There does seem to be a sense in which the dog becomes less dangerous for or to me the further I get away from it, but here we must be careful. What the evaluativist* needs is not just the sense in which a dog may be dangerous to a toddler but not a grown man, since it is incorrect for a toddler to fear a dog that is dangerous to toddlers if it is very far away. She needs to posit a relatively fine-grained relational value, much like health: what food is healthy for me does not depend in the first instance on the broad species into which I fall but on my particular constitution.

However, it is unlikely that we will be able to find such fine-grained relational values to accommodate our intuitions about the correctness of each of our emotions.[82] Consider amusement. No doubt there are some jokes which had currency in the 17th century that were extremely well-constructed, aesthetically speaking. They were timely, inventive, and cutting. But with time we have also no doubt lost the cultural background necessary to appreciate the joke. The lambasted institutions have fallen away, and what was inventive about the jokes may now be clichéd. For those reasons it would not be correct for modern audience members to be amused upon hearing the joke. Nor would it be appropriate if they were filled in on the missing cultural background, for as a rule explanations kill jokes. Still, the passing of time has not made the jokes less *funny*. They are less funny in the sense that they now lack the power to *produce* amusement, but they retain their aesthetic value of funniness even if we lack the ability to appreciate them. And in this case the relationalizing move looks less plausible, for intuitively it is not that the jokes are aesthetically-good-with-respect-to-the-17th-century but not relative-to-us-now. There are no such relational aesthetic values. If our culture changed we might be able to *appreciate* and *rediscover* the value that was there all along; the joke would not simply change from being not aesthetically

[81] Of course, interpretivists about the mental (Dennett 1981; Davidson 1986) would not necessarily share this intuition, but this view has not been in recent favor among quasi-perceptualists about emotions. Indeed, in its motivation interpretationism seems to be at odds with such views insofar as it tends not to hold that the the mind possesses discrete representational states which causally interact.

[82] Note that there is also in ethics today a general resistance to posit relative or relational values outside of good-for. Schroeder (2007b) has been particularly influential in inculcating this resistance.

good to being aesthetically good.

Or consider the oft-noted interaction among emotion, value, and special relationships.[83] I could hardly deny that Steve's giving his mother a rose on Mother's Day would be just as valuable as my giving my mother a rose on Mother's Day, but it would be odd to insist that I must be just as pleased at Steve's giving his mother the rose as I am at mine own.

Moreover, there are problems with the left-to-right direction of the emotion-value equivalences too. There are, for instance, situations where it is correct to fear what is not in fact dangerous, as when one has strong evidence that one is in a dangerous situation. You are walking down a long and dark alleyway in the south side of Chicago, where recently many robberies have been reported. As it happens there are no robbers lurking about at the moment, so you are safe, but it is impossible for you to tell. It would be correct for you to be afraid of walking down the alley; it is a situation in which fearful alertness is appropriate.

All this evidence points in favor of a general lesson: values have a certain kind of perspective-independence that correctness conditions for emotions do not share. Our emotions, and affective states more broadly, are appropriately sensitive not just to the presence of value but to our relation to it in time, distance, evidence for its likely instantiation, etc.[84] And note that *none* of the considerations adduced above are inconsistent with the thesis that affective states' correctness conditions are given in terms of reasons for action, for an agent's reasons famously are sensitive to her circumstances.[85]

Before moving on I should make a few notes about the implications of this thesis. Certain concepts — *shameful, disgusting, fearsome, admirable* etc. — wear their connection to emotions (or attitudes) on their sleeves, and for these concepts an analysis in terms of the correctness or fittingness of the relevant emotion is hard to avoid: something is fearsome just in case it is correct (or fitting, merited . . .) to fear it. These terms are also evaluative in the broad sense since they are terms of praise and condemnation: to call something disgusting is to condemn it and to call something admirable is to praise it.[86] However, this does not imply that such terms are evaluations*

[83] See Ewing (1948, 159); Oddie (2005, 60-63).

[84] See Oddie (2005, Ch. 8) for a good articulation of this general point of view, albeit with desires substituted for emotions.

[85] One might object that there is at least one kind of case where my analysis of affective states' correctness conditions predicts the wrong result: cases where it seems fitting or merited to feel some emotion but the associated action is one that it impossible for me to do in my present circumstance, and so it seems it is one I have no reason to do. I recall Lincoln's assassination and am frustrated by it, say. The frustration seems to represent that I have reason to punish the assassin Booth, or perhaps even to have prevented it, but I cannot do either of these things. I think such emotions are best understood as representing reasons to act *in the fiction* that one is observing the imagined scene (cf. Chapter 3). But since I agree that the frustration is actually fitting to feel in these cases, I think all these cases show is that fittingness conditions and correctness conditions for affective states are not always the same.

[86] This is outside of special constructions such as "He's too admirable to be genuinely likable" and "Fried ice cream is so disgusting that it's good". Some might argue that the possibility of such constructions shows that these terms are not genuinely evaluative and instead only implicate evaluations in context; Blackburn (1998, 101-104) could be read

in the narrow sense of denoting values*. An analysis of terms like "shameful" in terms of a correct response may have the air of tautology, but an attempt at analyzing *value** in terms of a correct response would be taken as an attempt to analyze "the good* in terms of the right", which is quite a controversial thesis.[87] Indeed, I have been using "dangerous" as a canonical value*, but the dark alley may be fearsome — correct to fear — even when it is not dangerous.

Nor, it should be noted, does the present argument entirely preclude fitting emotion theories of value (let alone fitting attitude or fitting response theories), even if it turns out that the shameful, disgusting, etc. are values*. It could be, for instance, that for something to be shameful just is for it to be correct (or fitting) to feel ashamed of it *under certain ideal conditions*, e.g. when one experiences it at a proper distance and under full information.

4.7 Features of hard-line affectivism

In this section I will point out a few distinguishing features of HLA and explain how it can succeed where the Standard View failed.

Note that HLA is indeed a Guise of the Good view, for it holds that the sense in which humans act under the guise of the good is that their actions are based on a representation of the good in so acting, and in particular on the representation of a normative reason. Furthermore the view entails that that represented reason *is* the agent's motivating reason for acting, thereby specifying the sense in which the agent acts *because* of the good she sees in so acting, as GG requires.

4.7.1 A content-based view

Note as well that HLA is a version of what I earlier called a *content-based view*, since action's orientation to the good is given in virtue of its connection to phenomenal content. The content-based view is currently unpopular. Recall that the root of Velleman's rejection of the content-based view was his taking GG to be a thesis about desire, construed as a propositional attitude, and his contention (perhaps inherited from Davidson) that an attitude toward a proposition requires the possession of the concepts used to express it.[88] This led to the objection that if GG was an essential truth about desires then it could not be that small children and animals have desires, since they do not possess the concept of the good. But they do in fact desire, so the content-based version of GG was held to be false. This was one motivation for the shift to the attitude/aim theory.

in this way. But even thin evaluative terms such as 'bad' can be embedded in those constructions: "The movie was so bad that it was good." Since we are unlikely to say that 'bad' in this context is not a negative evaluation, the strategy seems unlikely to succeed.

[87] See Zimmerman (2015) for an overview of the issue.

[88] See p. 85 above.

I need not take a stand here on the connection between concepts and the content of desire. I need only note that because HLA holds that the appearances of the good relevant to GG are *affective* states, not desires, the force of this objection is attenuated, if not extinguished entirely. For the content relevant to HLA is *phenomenal* content, which few believe must be conceptual content.[89] Indeed, many hold with Tye (1995a) that the content of phenomenal experience *must* be non-conceptual. So it is entirely consistent with HLA that small children and animals act under the guise of the good.

HLA thus undercuts a major argument in favor the attitude/aim theory of GG. This was admittedly not the attitude/aim theory's main source of support, however, since that theory theory arises out of a long tradition of holding that the good is the formal aim of action,[90] much as truth is the formal aim of theoretical reasoning It is indeed sensible to think that believing a proposition is believing that proposition to be true, though it is is the proposition alone that is the content of one's belief and not *that the proposition is true*. But I also think the work done in §4.2.2 in clarifying this analogy between goodness and truth undermines the coherence of the attitude/aim theory as a *distinct option* from the content-based view.

In that section I argued that the analogy between truth and goodness boiled down to a pair of correctness conditions. From the idea that belief is governed by intrinsic correctness conditions, I derived the preliminary formulation:

A belief that X is correct iff it is true that X.

I then applied an indirect disquotational schema to this to derive the formulation expressed in 23:

A belief that X is correct iff X.

And then I noted that the corresponding correctness condition for an appearance of the good was 24:

An appearance of φing as good is correct iff φing is good.

But recall as well that it does not suffice for a GG view merely that when one acts one instantiates a state that has intrinsic correctness conditions, nor even that such a state be governed by its intrinsic correctness conditions. The agent's relation to the good must be more explicitly intentional. She must have mental access to the good; it must be presented to her. Therefore, it seems, the intrinsic correctness conditions in 24 must be presented to her (as obtaining). But *correctness conditions presented to a subject* just is the notion of representational content that has been in use in this dissertation since Chapter 2. So, the content of an appearance of φ as good just is that φing

[89] A notable exception along these lines is McDowell (1996).

[90] Aquinas has often been interpreted this way; see *ST* I-II q. 9, a. 1; q. 94, a. 2.

is good — and so goodness is part of the content after all. The attitude/aim theory is not, on its best version, distinct from the content-based view.[91]

To return to HLA: One common objection to the view that affective content is evaluative content is that it seems to require unintuitive verdicts about the contents of emotions. I am afraid of *the snarling dog* before me, or perhaps *that there is a snarling dog before me*; I am not afraid *that the snarling dog is fearsome*. Putting values into the content of emotion double-counts the connection between the emotions and evaluation, it is held, since emotions themselves are evaluative attitudes: fear of the dog is to be construed as Fearsome(the dog).[92]

The problem with this objection is that it ignores the fact that there can be multiple ways of describing a token mental state by its intentional properties. It is altogether possible that I fear *the dog*, and that my fear represents *the dog as fearsome*, and that it represents *that the dog is fearsome.* The fear can be characterized both as an attitude of the form Fearsome(the dog) and as one of the form IstheCase(the dog is fearsome). (In fact I have difficulty making sense of the former expression unless it is taken as equivalent to the latter). Gregory (2016), another defender of the Guise of the Good, thinks that one state can be described with two different contents. This is altogether similar to the claim I am making here, though for the reasons just given I believe it is important to point out the possibility that there are at least two notions of content in play here. One notion of content is representational content, cashed out in terms of correctness conditions, and another is the object at which a mental is directed.

In any case, HLA does not misdescribe the content of affective states, for it does not deny that when one fears a dog, one's fear is of the dog.

4.7.2 Affect and evaluation

In §4.3 I argued that desires were not essentially evaluations. One might wonder whether the same argument might be used to show that affective states are not essentially evaluative either. Furthermore, the intentionalism defended of affective states in Chapter 2 can easily be construed as entailing that states with phenomenal character are essentially representational states, so their falling prey to the same argument would undercut a main thrust of the present project.[93] I have already defended the claim that affective states represent normative facts — at length in Chapter 2, and I refined the view in §4.6.3 — but of course, I should also check that this argument does not

[91] Note that Dokic & Lemaire (2015, 276-281) offer a distinct but complementary argument against the view that emotions' connection to the good is through the nature of the emotional attitude, not the content of the emotion. They argue that if an emotion is not literally *about* a thing's being valuable or unvaluable then it cannot play a desired role in justifying evaluative beliefs about that thing.

[92] See Deonna & Teroni (2012) for an objection along these lines.

[93] NB, however, that this does not in turn entail that affective states essentially represent reasons. The hypothesis defended here is only that they actually represent reasons.

work against affective states.[94]

Take, then a modification of the case in which Determined and Ditherer are both motivated by an emotion: the exciting, thrilling prospect of petting a large and dangerous cat. Determined is reliably excited by the prospect up until the very end, while Ditherer's excitement wavers upon consideration of the fact that petting a hungry tiger may not be a good idea. (Of course it need not be supposed that they do not also desire to pet the cat, for the task here concerns only their emotions). Supposing that emotions can be evaluable as better or worse *qua* affective states, and that emotions are evaluations of one's reasons, are emotions better *qua* affective states to the extent that they better represent one's reasons? It seems to me that they are. One's emotion is correct to the extent that it correctly reflects a reason that one has, and it responds *well* as an emotion to the extent that it reflects the evidence available to it on that question — much as a belief is correct if it is true, and it is good as a belief to the extent that it responds to the evidence available to the believer. Ditherer's excitement is at least at times appropriately modulated by his (justified) thought that petting a tiger is not in the slightest respect a sensible thing to do, and is for that reason responding well as an emotion. So, the thought experiment in §4.3 cannot be used to draw a parallel conclusion that emotions are not essentially representational.

This conclusion can be drawn even though little has been said to settle precisely the motivational role of affect. I noted in Chapter 2 (§2.3.1) that affect has a direct role in motivation (by altering goals and focusing attention) as well as an indirect role (by feeding into conceptual evaluations of one's situation), and of course HLA assigns affective states a role in motivation as a basis of action. Nor have the relations among motivational strength, affective phenomenal intensity (the vividness of affect), and the strength of reasons that an affective state represents been settled. All these are crucial questions, but addressing them takes us beyond the present, preliminary inquiry.

4.8 Objections

Before concluding, I consider three potential objections to HLA.

The first and most obvious objection is that it certainly seems to be the case that we can act for a reason without having any *feelings* about our action at all. A candidate counterexample should be obviously done for a reason, avoiding actions done merely out of habit and actions one simply finds oneself doing for no reason at all. But even so, a diet of examples will spring to mind. Melinda Vadas (1984) gives us two excellent candidates:

[94]Of course, there were *two* arguments against the Standard View above, the second of which was the Wobbly Argument. But whereas the first argument aimed to show that *actual* desires were not evaluations *qua* desires, the Wobbly Argument merely aimed to show that it is possible for a creature to act for a reason but not under the guise of the good. Hence the Wobbly Argument poses no threat to HLA.

Mathematician Mara the mathematician is considering either of two strategies for completing a proof. She briefly considers both and decides that one would be shorter, and so chooses that one.

Drowning Cat You are engaged in a heated discussion with Professor X on a dock by a lake. Just are you are consumed by a desire to strangle X, "whose position on moral responsibility would befit that of the losing side at Nuremberg" (*op. cit.*, 273), you see that a cat has fallen into the water beside you and is floundering. Without withdrawing yourself from the conversation in the slightest, you bend down to toss the cat a nearby rope so that it may climb to shore. As you are about to lay into X in no uncertain terms, the owner of the cat rushes over to collect the cat and thank you. You, irritated by the interruption, assure the owner that you have no feelings for cats. You saved it simply because it was drowning.

I think the first response to these cases is the tried-and-true appeal to "calm passions" that are "more known by their effects than by the immediate feeling or sensation".[95] In such cases we are not disposed to *judge* that we feel or felt anything about what we did, but that is only because we are not focused reflexively upon our emotions. We are focused on the work at hand, or on Professor X's heinous opinions. But this is not at all incompatible with our feeling brief the sort of brief flash of affect that is hypothesized to play a role in cognition (see p. 117 above). Indeed, Mara might instantly *appreciate* how much more elegant the shorter strategy is and straightaway choose it without pausing to reflect on her own appreciation. Similarly, the thought that the cat is to be saved because it is drowning is surely an odd one unless it is backed up by *respect* for the cat as a sentient being.

Some might think this response does not itself appreciate the worry such cases can pose. Suppose then that Mara has no feeling at all about any of what she is doing; though she may feel strongly about certain styles of proof, the present work is routine for her. She is inured to it. Or suppose that in Drowning Cat, not only are you not disposed to care about cats in any case, you are so consumed with hatred for Professor X in the moment that it is as if you had no room left to care now about this cat.

I think the affectivist might well demur at these descriptions for the same reason as before, but there is also another response available to her. HLA requires that one's action be *based* on an affective state, and it is common to hold that the basis of one mental state can be at some temporal distance from it. Just as my basis for my belief today that Theorem Y is true is the derivation of it I performed yesterday, the affective basis of Mara's using the shorter strategy now is her previous appreciation of the elegance of using shorter proofs. This same feature can also explain how executions of intentions can be based on affective states: disappointed to discover that I have little food in the pantry one morning, I form the intention to go to the grocery store later that day. When after work I turn right to go to the grocery store instead of left as usual, I need not rehearse

[95] Hume, *A Treatise of Human Nature*, 2.3.3.8.

my morning's disappointment in the car in order for it to provide my reason for turning right when I execute my intention later in the day, for the intention itself was motivated by that disappointment.

A second objection to HLA is that it over-predicts action for a reason. It is usually thought that all cases of action for a reason are, relative to the range of possible expressions of one's agency, among the most full-blooded. Acting for a reason requires the agent to be fully participant in her action as it involves a recognition on her part not only of the fact *that* she is acting but an idea of the reason *why* she is. When I am watching a film I adjust my posture frequently to avoid discomfort and relieve muscle cramps. These may well be intentional actions, but it is held that they do not involve the same level of agency as when I change my posture so as to obscure the view of the noisy person sitting behind me. I am hardly aware of changing my posture in the first case, but I explicitly mean to change it in the second so as to give the guy behind me a taste of his own medicine. But according to HLA, it seems, both are equally cases of action for a reason since both are actions based on appropriate affective states — discomfort in one case and annoyance in the other.[96]

To the charge that I am lowering the bar for action for a reason, I plead guilty, but I protest that it is not a crime. It is a feature and not a bug of the account. If we like we can distinguish between acting for a reason and the special case of acting for a reason *that the agent has conceptually articulated*. And there may furthermore be moral reasons to care more about the special case, perhaps because in such cases the agent has more control over her action or because they are more expressive of her character. In contexts where these moral considerations are important it may be felicitous to deny that I change my posture for a reason in the first version of adjusting my posture in the movie theater. But even if one follows Anscombe (1963, 9) in thinking that action for a reason is action for which the question "Why?" has a positive response on behalf of the agent, one needn't follow her in thinking that the agent *actually* needs to be able to articulate her reason for acting. The core idea of acting for a reason is acting in light of (what one takes to be) an answer to the "Why?" question. And when I shift my posture to avoid discomfort, I do act in light of a reason. Just as Chapter 2 (as amended in §4.6.3.2) suggests, my reason for changing posture is the damage or bodily problem which my discomfort represents. Why should whether my reason is conceptually or non-conceptually represented make a difference as to whether my action is done for a reason at all?

A third objection to HLA raises a regress worry. It is clear that one can have reasons for beliefs, intentions, and other mental states. One can have reasons for affective states: my reason for being angry at the passerby is *that he hit me*, and my anger may well dissipate if I come to believe that he merely was knocked into me or hit me only unintentionally. But it seems that my account of acting for a reason in HLA cannot be extended to an account of holding a mental state for a reason. For

[96]For this objection I thank Sabine Döring.

one thing it is implausible that every mental state that is held for a reason is based in an affective state. Many of my everyday beliefs are based only in perception, for instance. And for another, even if the basing role of affect were ubiquitous in our cognitive economy, this would require either the ability of affective states to base themselves (which is implausible on its face and dubiously coherent if the basing relation is a causal relation) or a regress of affective states until one was reached that was not based in anything (which is again implausible).

So I will need at some point to appeal to a non-affective theory of how mental states can be had for reasons. But then the crucial question for me becomes why I could not apply this same theory to acting for a reason. I have a visual experience of the brake lights of the car in front of me lighting up, and I form the belief that it is slowing down. My belief, caused as it is in the right way by this experience, is based on that experience. Why, then, could I not simply form the belief that the car's slowing down is a reason for me to brake and then brake for that reason — all without affect?

To this I have two responses. First, I agree that in this situation if I *were* to (a) form the belief that the car ahead's braking is a reason for me to brake and (b) base my action on that belief (c) all without basing my action on an affective state, *I would still act for a reason*. So much is entailed by AFR and the extremely plausible thesis that to represent oneself as having a reason for an action is also to represent that action in a weak normative light. My claim in this chapter is that in humans, this does not happen. We are not designed this way. Indeed, my realizing that I *need* to brake is not an affectively neutral thought. If it were then although I might judge I *ought* to brake, I would not think *to brake*. The fact that such affectless conceptual judgments about reasons do not in fact generate action is what enables the judgment externalist module (E, p. 110) to explain accedie.

Second, there is good reason to think that an account of action for a reason should *not* extend straightforwardly to the general case of forming and holding mental states for a reason. When I act, I as a person am active. But not all cases in which I as a person am active involve actions of mine. Some cases of non-action activity are relatively uncontroversial, as when I follow a speaker's argument, realize upon looking at a familiar-looking storefront that what I took to be a stand-alone establishment is actually a chain, or am lost in thought about what to have for lunch. In these cases I am responding, realizing, or thinking things through, but unless every rationally-formed thought is a mental act, it seems in these cases I need have performed no actions. One might further more think that these non-action mental activities have a different kind of relationship to the mental states that rationally explain them and on which they are based than actions do.[97] According to an orthodox conception of action explanation familiar since Davidson, an action is explained in terms

[97] I do not mean to presume that only active mental states can be had for reasons. Aside from a handful of core cases such as the ones I mentioned, my intuitions about which states are active are in flux. However, one main current in philosophy, familiar since Kant at the latest, connects the active part of the mind to the faculty of reason.

of mental states (beliefs and pro-attitudes, usually) that are *about* that action and which causally guide it. The causing and explaining mental states are *posterior* to the action: they push it forward, so to speak.[98] But this kind of account is implausible for explaining how, say, my belief that the store is a chain is based on my perception. Even if the latter causes the former, my perception is not *about* my belief and does not causally guide it into existence in the same way; it is *lateral* to my belief. The distinction between affective and non-affective reasons-explanations may very well track the different ways in which action and non-action activity are explained.

4.9 Conclusion

In this chapter I argued that the standard understanding of the Guise of the Good was false in three significant respects. First, it is too strong: it is not a *necessary* truth about action for a reason that it occurs under the guise of the good, given the argument in Chapter 3 that GG views must avoid the weakness and inertness worries. That was what the Wobbly Argument showed. Second, desires are not essentially evaluations. For that reason GG as a thesis about the nature of desire is untenable. Third, I argued that the dominant understanding of the *guise* of the Guise of the Good, the attitude/aim theory, either misunderstands the relationship between mental content and the good (§4.2) or collapses into the content-based strategy (§4.7.1). I then proposed HLA, an affect-based version of GG that neatly avoids all of these problems. HLA is not committed to holding that GG is a necessary truth, since it holds it true only of actual humans (and animals, insofar as they act for a reason). It holds that the appearances of the good that GG is committed to are affective states, and because it depends on the representational conception of affect defended in Chapter 2, it is a version of the content-based strategy.

Most significantly, HLA, as a GG theory, is able to secure a version of internalism after all. It is not quite that agents in acting for a reason are absolutely *incapable* of opting out of the pursuit of the good. Rather, it is that they cannot so opt out *while remaining human*. And this, it seems to me, is attractive enough of a prize.[99]

[98] This is not to say that the conception is uncontroversial; far from it. Ginet (1990, Ch. 1) rejects this picture in many significant ways, for instance, and Broome (2013, §13.4) holds that all reasoning and even mental events such as calling a fact to mind are acts.

[99] An earlier version of this paper was presented at the Institut Jean-Nicod in Paris, and I thank the audience there for their feedback. Thanks also to Sarah Buss, Sergio Tenenbaum, Rohan Sud, and Damian Wassel for their helpful comments.

CHAPTER 5

Conclusion

The central goal of this dissertation was to articulate and defend a coherent view of the connections among several notions central to action theory and ethics, and in particular among affect, action for a reason, making sense of our actions, and the justification of action. In it I defended the view that affective states are representations of normative reasons for action, and I argued that such a representational theory of affect could underwrite a Guise of the Good theory.

There are a few themes that unite this dissertation. Chapters 2 and 4 do cover somewhat similar problems concerning the role of representations in the rational explanation of action. In Chapter 2 the problem was the access prong of the shooting the messenger problem: Ann has practically immediate access to the badness of her pain (or as I would rather put it now, to the fact that she has a reason to get rid of it), but in spite of the role evaluationism assigns to unpleasantness in representing the bad alteration in her knee, it seemed at first that her introspective access only gave her access to an experience with a bad-injury-in-knee character — and thus not to anything that would seem to justify her getting rid of her pain. The difficulty was that intuitively, Ann does have immediate access to an apparent justification for getting rid of her pain; it does make sense to her to get rid of it. Chapter 2 addressed this difficulty by appealing to the representational nature of secondary unpleasantness: in introspecting her pain, Ann feels fear or anxiety which represents her pain as bad.

Chapter 4 presented the *wrong kind of goodness* and *non-inheritance* problems in the context of the purported analogy between goodness and truth (§4.2.2). In the literature on GG one often sees the claim that just as the constitutive aim of belief is truth, the constitutive aim of desire (or better, of the candidate appearance of the good) is the good. But note that truth is a property of representational entities, and that beliefs inherit their truth from that of their propositional contents. The wrong kind of goodness problem points out that the kinds of good with which GG is concerned are not generally properties of representational entities, and the non-inheritance problem points out that desires do not inherit their goodness from the goodness of the actions, things, etc. which they are about. The solution to these problems proposed in Chapter 4 which preserved the analogy was, in a way, to understand both belief and appearances of the good as governed by world-directed

correctness conditions: beliefs aim to correctly represent the world, and appearances of the good aim to represent the goodness of actions. The two turn out not to be so very different after all, and that is what led to the result in §4.7.1 that the attitude/aim theory is not, on its best version, distinct from the content-based strategy. Thus the discussion of both of these problems in Chapters 2 and 4 demonstrate how difficult philosophical problems can be addressed and views clarified through careful attention to the properties of representations.

GG holds that there is a deep connection between acting for a reason and the good, and Chapters 3 and 4 investigated the nature and modal strength of this connection. Chapter 3 argued that if GG is to avoid weakness it should hold that to see an action as good is to see it as meeting a standard of practical reason, while Chapter 4 argued that GG was not a necessary truth, much less an essential truth about the nature of desire. And all three main chapters investigated how we human agents make sense of our actions, with Chapter 3 drawing out a view of the sort of intelligibility of action which is commonly held among GG theorists to be a necessary feature of action, and Chapters 2 and 4 emphasizing the role of affect as a representation of reasons for action in making sense of one's own actions. It is worth noting in this regard that Chapters 2 and 4 thus share the challenge of explaining both how it *could* be that affective states represent reasons —that is, what representational account of phenomenal states would enable affective states to represent normative considerations — and why they in fact *do* represent normative reasons.[1] But adequately addressing this problem would take a dissertation by itself.

A few other open questions remain, especially concerning the relation between the views defended in the previous two chapters. So, in the remainder of this concluding chapter I would like to make note of a few outstanding issues and how they might be resolved.

Chapter 3 developed the intelligibility motivation for GG, which in outline goes as follows. It is held to be a pre-theoretic datum that an agent φs for a reason only if φing is intelligible$_A$ to her; this is the intelligibility constraint. GG then offers an account of intelligibility$_A$: an action is intelligible$_A$ to an agent just in case, and because, it appears good to that agent. From this and the intelligibility constraint, GG follows: an agent φs for a reason only if φing appears good to her. (See §3.3.1.)

The only problem with this argument is that the Wobbly Argument of Chapter 4 showed that GG is not a *necessary* truth. This does not show the conclusion of the intelligibility motivation to be actually false, but it does restrict the strength of the premises used to derive it: at least one must be not necessarily true. Given the difficulty of finding an *a priori* defense of the Guise of the Good's account of intelligibility$_A$ against the possibility that near-goods could make actions intelligible$_A$ (§3.3.2, p. 59), it seems the evidence points in favor of thinking that the account of intelligibility$_A$ offered in Chapter 3 is not true as an account of the very nature of intelligibility$_A$. It

[1] This might be called "Schroeder's Challenge", after the discussion in Schroeder (2008, 127-129).

is not necessarily true that agents make their actions intelligible$_A$ by seeing them as good. Instead the account should be considered to offer an explanation of the way in which humans make their actions intelligible$_A$. But given the anthropocentric view of the connection between human action and the good broached in Chapter 4, this result was perhaps to be expected.[2]

It is often pointed out that affective states can render certain actions intelligible to the agent who instantiates them.[3] But perhaps the most interesting question left on the table is the role that affect may play, not just in making our actions intelligible in some sense or other, but in the specific sense of intelligibility$_A$ developed in Chapter 3. Could it be that humans find their actions intelligible$_A$ by affectively representing there as being reason for those actions? HLA would naturally suggest this, and this would also easily explain the intelligibility$_A$ of expressive actions. Moreover, affective representation may also prove crucial to explaining INI, for it is plausible that there is a deep connection between empathy, understood as the ability to understand what feelings others are having, and sympathy, the capacity to feel for others in light of how they feel.

Here is one way the connection could go. If affect is at the bottom of our own ability to act for reasons in the way that HLA requires, then affect is what enables us to see a *point* to our acting — indeed one might say that according to HLA the point of one's action is more felt than seen. Plausibly, then, my understanding the point that another agent, Myrtle, sees in her own action — that is, my making it intelligible$_A$ — requires me to possess much the same affective machinery that Myrtle must use in rending her own action intelligible$_A$ to herself. To see the point she sees in acting, perhaps I must be able to feel about it the way she feels about it — and that may mean that what concerns Myrtle about her action must be of concern to me too. Of course, this is not to say that I must be actually disposed to care about the very same thing Myrtle cares about. She may feel so moved by a Thomas Kinkade painting that she feels she must buy it, and yet I may find the painting so bright that it practically pains me to look at it. But to understand why Myrtle buys the painting I must understand how the painting makes her feel the way I would feel in a peaceful country hamlet — and I would feel cozy and at peace. So Myrtle and I do in a broad sense share an affective sensibility with respect to a certain value, even if we are not disposed to find that value in the same things. But if I share that affective sensibility, then it seems I can exercise it in making Myrtle's action intelligible$_A$ to me, just as INI predicts.

I regard the foregoing as an interesting and plausible possibility, but it will need to remain

[2]One question raised by this possibility is how much the thesis that *values* are *public* might depend on particularly human capacities and how much might depend on more abstract, structural features of agency. For instance, it may be that the argument in §§3.3.2-3.3.3, for the thesis that intelligible$_A$-rendering appearances of any actions must have the force to rationally support either performing those actions or tolerating them depending upon one's standpoint, does not depend on GG's interpretation of intelligibility$_A$. (Perhaps even the Wobbly must favor tolerating others' actions *if* it finds those actions intelligible$_A$ and conducive to the point it thereby appreciates in the action.) And given HLA and the anthropocentric view of value, it may be that affect is a large part of what ties human action to value.

[3]Aside from the literature on evaluationism about pain, see for instance Tappolet (2003); Döring (2007).

undefended for the time being. And given that the argument for the publicity of values in Chapter 3 depended on the intelligibility constraint, which was taken as a starting-point of that chapter, the thesis that the goods under which we act are publicly shared will also need to remain without a complete defense. For that reason, what we have thus far is only the sketch of a program for avoiding the inertness worry for GG. Nevertheless, the GG theory set out in Chapter 4 does indeed avoid the weakness worry. That is because it holds that in acting for a reason, human agents represent their action as meeting a standard of practical reason, and in particular as there being a reason to perform that action. In that chapter we also saw what we can learn about action from HLA as a version of the Guise of the Good: for humans, at least, it is not possible to act for a reason without doing so for what one takes to be a *good* reason, that is, a normative reason. While it does not quite provide a vindication of practical reason, HLA would at least show that humans, such as they are, cannot reject practical reason wholesale.

APPENDIX A

Conceptualizations of Intelligibility

It should be noted that the characterization of the intelligibility requirement given in §3.3.1 simplifies conceptualizations of the intelligibility requirement to an important extent. One can find at least five *prima facie* distinct conceptualizations in the literature:

Formal Aim The constitutive question of practical reason, "What to do?", just is the question of what is *good* to do, and actions are unintelligible if the agent cannot offer an answer to the question which takes the required form (i.e., which shows the action to be thought good in some respect). As Vogler (2002, 31) puts it, "... the questions 'Why are you doing that?' and 'What's the point of doing that?' and 'What's the good of doing that?' are basically the same question." See also Boyle & Lavin (2010, 191).

Blindness Behavior would be "blind" if it were not done under the guise of the good, and blind behavior is not action for a reason. Unintelligible behavior is blind in the relevant respect.

Active/passive Action is necessarily on the active side of the active/passive distinction, and action is only active if it is done under the guise of the good. The agent is passive with respect to actions that are unintelligible to her.

Internal/external Action necessarily is or emanates "from within" the agent. This motivation is skeptical of behavior caused by "alien" desires, especially desires that we do not bring within ourselves (as it were) by endorsing them. Unintelligible actions are without the agent.

Ownership Actions do not *belong* to us unless we see them as good, and our actions for a reason necessarily belong to us. Seeing the action as good may be a precondition of our being able to endorse it, and we do not act for a reason unless we make the action ours by endorsing it. Unintelligible actions do not belong to the agent.

These perspectives are not mutually exclusive and theorists relate them in different ways. Frankfurt (1977, 59), for instance, analyzes the active/passive distinction in terms of the internal/external one. Often the last three are mentioned practically in the same breath (e.g. Schapiro 2014, 132-134).

The intelligibility motivation has, in recent years, been most strongly connected with these last three conceptualizations. However, my own intuitions about them are not robust enough to incline

me to put much theoretical weight on them. Furthermore, the Formal Aim motivation seems to me to beg the question against non-GG interpretations of IC. This leaves Blindness. Vague though it is, Blindness reflects the intuitive and popular notion that action for a reason requires having an idea of what one is doing and why.

APPENDIX B

On Davidson and Intelligibility

One might wonder whether there is a practical analogue to a Davidsonian view of truth and meaning that could enable us to derive INI from *a priori* premises. Davidson famously argues that it is a condition on the sentences of anyone's language meaning anything at all that they be translatable into our own language, and further argues from this possibility and from the thesis that meaning is holistic to the claim that the beliefs of those it is possible for us to interpret must overlap ours to a significant extent. (See for instance Davidson 1973b,a.) The first step is especially suggestive of a theoretical analogue of INI, and one might think Davidson's theory would give us insight into an intersubjective condition on the intelligibility$_A$ of action. Indeed, one might be all the more inclined to think this given that Davidson does indeed think that it is a condition of having intentional attitudes at all that one meet global requirements of coherence and plausibility that include being interpretable as a "lover of the good" (Davidson 1970b, 221-222).

But there are many reasons to be cautious about this strategy. First, it is not clear that such global requirements of coherence and plausibility would enable us to derive INI as we have interpreted intelligibility$_A$. As we have noted (note 71, p. 62), intelligibility$_A$ requires more than mere coherence, and for all that Davidson says seeing someone else as a lover of the good may require only the sort of understanding described in condition 14. It may not require that one find their action intelligible$_A$ (condition 15, p. 68). Second, Davidson's contention that in order for an entity to be an agent we must be able to find their attitudes on the whole plausible given their circumstances *depends on* his presumption that he would be able to extend the argument above from the nature of radical interpretation to mental states more generally. But we began by wondering precisely whether this is true, so we cannot look to Davidson for help here.

Third, the argument from radical interpretation adumbrated above depends upon claims about the nature of truth that do not clearly have a practical analogue, and are moreover highly controversial. For instance, Davidson seems to assume that our understanding of the truth predicate requires that we can only apply it to a sentence if there exists a translation of that sentence into our own language, or that the only metalanguage in which we can competently apply the truth-predicate is our own (Davidson 1973a, 194-195). It is not clear why this should be so. Let '*P*' name a sentence

in a language which is completely unfamiliar to me. Still, it seems that I can understand perfectly well that ⌜'*P*' is true if and only if *P*⌝ is true. For these reasons, I doubt that we should look to Davidson's theory as a model for supporting INI.

BIBLIOGRAPHY

Albritton, Rogers. 1985. "Freedom of Will and Freedom of Action." *Proceedings and Addresses of the American Philosophical Association*, vol. 59 (2): 239–251.

Alvarez, Maria. 2010. *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford UP, New York.

Anscombe, G. E. M. 1958. "Modern Moral Philosophy." *Philosophy*, vol. 33 (124): 1–19.

—. 1963. *Intention*. Blackwell, Oxford, UK, second edn.

Arpaly, Nomy. 2000. "On Acting Rationally against One's Best Judgment." *Ethics*, vol. 110 (3): 488–513.

Arpaly, Nomy & Timothy Schroeder. 2014. *In Praise of Desire*. Oxford Moral Theory. Oxford UP, New York, NY.

—. 2015. "A Causal Theory of Acting for Reasons." *American Philosophical Quarterly*, vol. 52 (2): 103–114.

Aydede, Murat. 2003. "Is Introspection Inferential?" In *Privileged Access*, Brie Gertler, editor, 55–64. Ashgate, Aldershot.

—. 2005. "Introduction: A Critical and Quasi-Historical Essay on Theories of Pain." In *Pain: New Essays on Its Nature and the Methodology of Its Study*, 1–58. MIT Press, Cambridge, MA.

Aydede, Murat & Matthew Fulkerson. 2014. "Affect: representationalists' headache." *Philosophical Studies*, vol. 170 (2): 175–198.

—. 2015. "Reasons and Theories of Sensory Affect." In *Forthcoming in: The Nature of Pain*, David Bain, Michael Brady & Jennifer Corns, editors, 1–37. Oxford UP, Oxford.

Aydede, Murat & Güven Güzeldere. 2002. "Some Foundational Problems in the Scientific Study of Pain." *Philosophy of Science*, vol. 69 (S3): S265–S283.

Bain, David. 2003. "Intentionalism and Pain." *The Philosophical Quarterly*, vol. 53 (213): 502–523.

—. 2013. "What makes pains unpleasant?" *Philosophical Studies*, vol. 166 (1): 69–89.

—. 2014. "Pains that Don't Hurt." *Australasian Journal of Philosophy*, vol. 92 (2): 305–320.

Bargh, John A. & Ezequiel Morsella. 2008. "The Unconscious Mind." *Perspectives on psychological science : a journal of the Association for Psychological Science*, vol. 3 (1): 73–79.

Behrens, Timothy E. J., Laurence T. Hunt, Mark W. Woolrich & Matthew F. S. Rushworth. 2008. "Associative learning of social value." *Nature*, vol. 456 (7219): 245–249.

Behrens, Timothy E. J., Mark W. Woolrich, Mark E. Walton & Matthew F. S. Rushworth. 2007. "Learning the value of information in an uncertain world." *Nature Neuroscience*, vol. 10 (9): 1214–1221.

Benn, Stanley I. 1988. *A Theory of Freedom*. Cambridge UP, Cambridge.

Bermúdez, José Luis. 1998. *The Paradox of Self-Consciousness*. MIT Press, Cambridge, MA.

Berna, Chantal, Siri Leknes, Emily A. Holmes, Robert R. Edwards, Guy M. Goodwin & Irene Tracey. 2010. "Induction of Depressed Mood Disrupts Emotion Regulation Neurocircuitry and Enhances Pain Unpleasantness." *Biological Psychiatry*, vol. 67 (11): 1083–1090.

Berthier, Marchcelo, Sergio Starkstein & Ramon Leiguarda. 1988. "Asymbolia for pain: A sensory-limbic disconnection syndrome." *Annals of Neurology*, vol. 24 (1): 41–49.

Blackburn, Simon W. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Oxford Clarenon Press, Oxford.

Block, Ned. 1996. "Mental Paint and Mental Latex." *Philosophical Issues*, vol. 7: 19–49.

—. 2003. "Consciousness." In *The Encyclopedia of Cognitive Science*. Wiley, New York, NY. Reprinted in "Consciousness, Function, and Representation", Collected Papers vol. 1. 2007. MIT Press, Cambridge, MA. pp. 111-127. Page references to this edition.

Bond, E.J. 1983. *Reason and Value*. Cambridge UP, Cambridge, UK.

Boyle, Matthew & Douglas Lavin. 2010. "Goodness and Desire." In *Desire, Practical Reason, and the Good*, Sergio Tenenbaum, editor, 161–201. Oxford UP, New York.

Bradley, Richard & Christian List. 2009. "Desire-as-Belief Revisited." *Analysis*, vol. 69 (1): 31–37.

Brady, Michael S. 2013. *Emotional Insight: The Epistemic Role of Emotional Experience*. Oxford UP, New York.

Bratman, Michael. 1987. *Intentions, Plans, and Practical Reason*. Harvard UP, Cambridge, MA.

Bratman, Michael E. 2009. "Intention, belief, and instrumental rationality." In *Reasons for Action*, David Sobel & Steven Wall, editors, 13–36. Cambridge UP, Cambridge.

Brentano, Franz. 1889. *The Origin of Our Knowledge of Right and Wrong*. Humanities Press, New York. Trans. Roderick M. Chisholm and Elisabeth H. Schneewind, 1969.

Brewer, Talbot. 2009. *The Retrieval of Ethics*. Oxford UP, New York.

Broome, John. 2013. *Rationality Through Reasoning*. The Blackwell/Brown Lectures in Philosophy. Wiley Blackwell, Malden, MA.

Buchak, Lara. 2014. *Risk and Rationality*. Oxford UP, Oxford.

Burgh, W. G. de. 1931. "Right and Good: Action "Sub Ratione Boni"." *Journal of Philosophical Studies*, vol. 6 (21): 72–84.

Buss, Sarah. 1999. "What practical reasoning must be if we act for our own reasons." *Australasian Journal of Philosophy*, vol. 77 (4): 399–421.

Byrne, Alex. 2001. "Intentionalism Defended." *Philosophical Review*, vol. 110 (2): 199–240.

Carver, Charles S. & Michael F. Scheier. 2013. "Goals and Emotion." In *Handbook of Cognition and Emotion*, M.D. Robinson, E.R. Watkins & E. Harmon-Jones, editors, 176–194. Guilford Press, New York.

Chalmers, David J. 2004. "The Representational Character of Experience." In *The Future of Philosophy*, Brian Leiter, editor, 153–181. Oxford UP, Oxford, UK.

Clark, Austen. 2005. "Painfulness is Not A Quale." In *Pain: New Essays on Its Nature and the Methodology of Its Study*, Murat Aydede, editor, 177–198. MIT Press, Cambridge, MA.

Clark, Philip. 2001. "Velleman's Autonomism." *Ethics*, vol. 111 (3): 580–593.

—. 2010. "Aspects, Guises, Species, and Knowing Something to Be Good." In *Desire, Practical Reason, and the Good*, Sergio Tenenbaum, editor, 234–244. Oxford UP, Oxford.

Clore, Gerald L. & Jeffrey R. Huntsinger. 2007. "How emotions inform judgment and regulate thought." *Trends in Cognitive Sciences*, vol. 11 (9): 393–399.

Cohen, Jonathan & Matthew Fulkerson. 2014. "Affect, Rationalization, and Motivation." *Review of Philosophy and Psychology*, vol. 5 (1): 103–118.

Cullity, Garrett. 2015. "Neutral and Relative Value." In *The Oxford Handbook of Value Theory*, Iwao Hirose & Jonas Olson, editors, 96–116. Oxford UP, New York.

Cutter, Brian & Michael Tye. 2011. "Tracking Representationalism and the Painfulness of Pain." *Philosophical Issues*, vol. 21 (1): 90–109.

—. 2014. "Pains and Reasons: Why It Is Rational to Kill the Messenger." *The Philosophical Quarterly*, vol. 64 (256): 423–433.

Dancy, Jonathan. 1993a. "Agent Relativity - The Very Idea." In *Value, Welfare, and Morality*, R.G. Frey & Christopher W. Morris, editors, 233–251. Cambridge UP, Cambridge.

—. 1993b. *Moral Reasons*. Oxford UP, Oxford.

—. 2000. *Practical Reality*. Oxford UP, New York.

D'Arms, Justin & Daniel Jacobson. 2000. "Sentiment and Value." *Ethics*, vol. 110: 722–748.

Darwall, Stephen L. 1983. *Impartial Reason*. Cornell UP, Ithaca.

—. 1992. "Internalism and Agency." *Philosophical Perspectives*, vol. 6: 155–174.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." *The Journal of Philosophy*, vol. 60 (23): 685–700. Reprinted in Davidson, Donald. 2001. Essays on Actions and Events, 2nd edition. Oxford UP, Oxford: pp. 3-20. References to this edition.

—. 1970a. "How is Weakness of the Will Possible?" In *Moral Concepts*, Joel Feinberg, editor, Oxford Readings in Philosophy. Oxford UP, Oxford. Reprinted in Davidson, Donald. 2001. Essays on Action and Events, 2nd edition. Oxford UP, Oxford: pp. 21-42. References to this edition.

—. 1970b. "Mental Events." In *Experience and Theory*, Lawrence Foster & J.W. Swanson, editors. University of Massachusetts Press, Amherst, MA. Reprinted in Davdison, Donald. 2001. Essays on Actions and Events, 2nd edition. Oxfrod UP, Oxford: pp. 207-228. References to this edition.

—. 1973a. "On the Very Idea of a Conceptual Scheme." *Proceedings and Addresses of the American Philosophical Association*, vol. 47: 5–20. Reprinted in Davidson, Donald. 1984. Inquiries Into Truth and Interpretation. Oxford: Oxford UP, pp. 183-198. References to this edition.

—. 1973b. "Radical Interpretation." *Dialectica*, vol. 27 (1): 314–328. Reprinted in Davidson, Donald. 1984. Inquiries into Truth and Interpretation. Oxford UP, Oxford: pp. 125-140. References to this edition.

—. 1974. "Psychology as Philosophy." In *Philosophy of Psychology*, S.C. Brown, editor. Macmillan and Barnes and Noble, New York. Reprinted in Davidson, Donald. 2002. Essays on Actions and Events: 2nd edition. Oxford: Oxford UP, pp. 229-244. References to this edition.

—. 1978. "Intending." In *The Philosophy of History and Action*, Yirmiaku Yovel, editor. D. Reidel and the Magnes Press, Jerusalem, second edn. Reprinted in Davidson, Donald. 2001. Essays on Actions and Events, 2nd edition. Oxford UP, Oxford: pp. 83-102. References to this edition.

—. 1985. "Incoherence and Irrationality." *Dialectica*, vol. 39 (4): 345–354. Reprinted in Davidson, Donald. 2004. Problems of Rationality. Oxford UP, Oxford: pp. 189-198. References to this edition.

—. 1986. "A Coherence Theory of Truth and Knowledge." In *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson*, Ernest LePore, editor, 307–319. Basil Blackwell.

Dennett, Daniel C. 1981. "True Believers : The Intentional Strategy and Why It Works." In *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*, A. F. Heath, editor, 150–167. Clarendon. Reprinted in Dennett, Daniel C. 1987. The Intentional Stance. Cambridge: MIT Press, pp. 13-42. References to this edition.

Deonna, Julien A. & Fabrice Teroni. 2012. *The Emotions: A Philosophical Introduction*. Routledge, London.

—. 2014. "In What Sense are Emotions Evaluations?" In *Emotion and Value*, Sabine Roeser & Cain Todd, editors, 15–31. Oxford UP.

Dokic, Jérôme & Stéphane Lemaire. 2013. "Are emotions perceptions of value?" *Canadian Journal of Philosophy*, vol. 43 (2): 227–247.

—. 2015. "Are Emotions Evaluative Modes?" *Dialectica*, vol. 69 (3): 271–292.

Döring, Sabine A. 2003. "Explaining Action by Emotion." *Philosophical Quarterly*, vol. 53 (211): 214–230.

—. 2007. "Seeing What to Do: Affective Perception and Rational Motivation." *Dialectica*, vol. 61 (3): 363–394.

—. 2009. "The Logic of Emotional Experience: Noninferentiality and the Problem of Conflict Without Contradiction." *Emotion Review*, vol. 1 (3): 240–247.

Dreier, James. 1993. "Structures of Normative Theories." *The Monist*, vol. 76 (1): 22–40.

Dretske, Fred. 1995. *Naturalizing the Mind*. MIT Press, Cambridge, MA.

Dutton, Donald G. & Arthur P. Aron. 1974. "Some evidence for heightened sexual attraction under conditions of high anxiety." *Journal of Personality and Social Psychology*, vol. 30 (4): 510–517.

Ekman, Paul. 1977. "Biological and cultural contributions to body and facial movement." In *The Anthropology of the Body*, J. Blacking, editor, no. 15 in A.S.A. Monographs, 39–84. Academic Press, London.

Ellsworth, Phoebe C. 2013. "Appraisal Theory: Old and New Questions." *Emotion Review*, vol. 5 (2): 125–131.

Evans, Gareth. 1982. *The Varieties of Reference*. Oxford UP, Oxford.

Ewing, Alfred Cyril. 1948. *The Definition of Good*. Routledge and Kegan Paul, London.

Fields, Howard L. 1999. "Pain: An unpleasant topic." *Pain*, vol. 82, Supplement 1: S61–S69.

Finlay, Stephen. 2007. "Responding to Normativity." In *Oxford Studies in Metaethics*, Russ Shafer-Landau, editor, vol. 2, 220–239. Oxford UP, New York.

Finlay, Stephen & Mark Schroeder. 2012. "Reasons for Action: Internal vs. External." In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, editor. CSLI Stanford, winter 2012 edn. URL `http://plato.stanford.edu/archives/win2012/entries/reasons-internal-external/`.

Fogal, Daniel. 2016. "Reasons, Reason, and Context." In *Weighing Reasons*, Errol Lord & Barry Maguire, editors, 74–103. Oxford UP, Oxford.

Foot, Philippa. 2001. *Natural Goodness*. Oxford UP, Oxford.

Frances, A. & L. Gale. 1975. "The Proprioceptive Body Image in Self-Object Differentiation—A Case of Congenital Indifference to Pain and Head-Banging." *Psychoanal Q.*, vol. 44: 107–126.

Frankfurt, Harry G. 1977. "Identification and Externality." In *The Identities of Persons*, Amélie Oksenberg Rorty, editor. University of California Press. Reprinted in Frankfurt, Harry G. 1988. The Importance of What We Care About. Cambridge, Cambridge UP: 58-68. References to this edition.

Freeman, Walter & James W. Watts. 1950. *Psychosurgery in the Treatmeant of Mental Disorders and Intractable Pain*. Charles C Thomas, Springfield, IL, second edn.

Frege, Gottlob. 1918. "The Thought: A Logical Inquiry." *Mind*, vol. 65 (1): 289–311.

Frijda, Niko H. 1986. *The Emotions*. Studies in Emotion and Social Interaction. Cambridge UP, Cambridge, UK.

Gazzaniga, Michael S. 2000. "Cerebral specialization and interhemispheric communication." *Brain*, vol. 123 (7): 1293–1326.

Gert, Joshua. 2007. "Normative Strength and the Balance of Reasons." *The Philosophical Review*, vol. 116 (4): 533–562.

Gewirth, Alan. 1978. *Reason and Morality*. University of Chicago Press, Chicago.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings*. Harvard UP, Cambridge, MA.

—. 2003. *Thinking How to Live*. Harvard UP, Cambridge, MA.

Ginet, Carl. 1990. *On Action*. Cambridge UP, Cambridge.

Goldie, Peter. 2000. *The Emotions: A Philosophical Exploration*. Oxford UP, New York.

Goldman, Alvin I. 1970. *A Theory of Human Action*. Prentice-Hall, Englewood Cliffs, NJ.

Gracely, Richard H. 1992. "Affective dimensions of pain: How many and how measured?" *APS Journal*, vol. 1 (4): 243–247.

Grahek, Nikola. 2007. *Feeling Pain and Being in Pain*. MIT Press, Cambrige, MA, second edn.

Gregory, Alex. 2013. "The Guise of Reasons." *American Philosophical Quarterly*, vol. 50 (1): 63–72.

—. 2016. "Why Do Desires Rationalise Actions?" Unpublished manuscript.

Grice, Paul. 1974. "Method in Philosophical Psychology (From the Banal to the Bizarre)." *Proceedings and Addresses of the American Philosophical Association*, vol. 48: 23–53.

Griffiths, A.P. 1997. *What Emotions Really Are: The Problem of Psychological Categories*. University of Chicago Press, Chicago.

Hájek, Alan & Philip Pettit. 2004. "Desire Beyond Belief." *Australasian Journal of Philosophy*, vol. 82 (1): 77–92.

Hall, Richard J. 2008. "If It Itches, Scratch!" *Australasian Journal of Philosophy*, vol. 86 (4): 525–535.

Hampton, Jean. 1998. *The Authority of Reason*. Cambridge UP, Cambridge.

Hardin, C. L. 1988. *Color for Philosophers*. Hackett, Indianapolis.

Hardy, James D., Harold G. Wolff & Helen Goodell. 1952. *Pain Sensations and Reactions*. Hafner Publishing Company, New York, NY.

Hare, Richard M. 1963. *Freedom and Reason*. Oxford at the Clarendon Press, Oxford, UK.

Harman, Gilbert. 1990. "The Intrinsic Quality of Experience." *Philosophical Perspectives*, vol. 4: 31–52.

Hawkins, Jennifer. 2008. "Desiring the bad Under the Guise of the Good." *The Philosophical Quarterly*, vol. 58 (231): 244–264.

Heathwood, Chris. 2007. "The Reduction of Sensory Pleasure to Desire." *Philosophical Studies*, vol. 133 (1): 23–44.

Heck, Richard. 2000. "Nonconceptual Content and the "Space of Reasons"." *Philosophical Review*, vol. 109 (4): 483–523.

Helm, Bennett W. 2001. *Emotional Reason: Deliberation, Motivation, and the Nature of Value*. Cambridge UP, Cambridge.

—. 2002. "Felt Evaluations: A Theory of Pleasure and Pain." *American Philosophical Quarterly*, vol. 39 (1): 13–30.

—. 2009. "Emotions as Evaluative Feelings." *Emotion Review*, vol. 1 (3): 248–255.

Hohfeld, Wesley Newcomb. 1913. "Some Fundamental Legal Conceptions as Applied in Judicial Reasoning." *The Yale Law Journal*, vol. 23 (1): 16–59.

Horgan, Terence & John Tienson. 2002. "The Intentionality of Phenomenology and the Phenomenology of Intentionality." In *Philosophy of Mind: Classical and Contemporary Readings*, David J. Chalmers, editor, 520–533. Oxford UP.

Hubin, Donald C. 2001. "The Groundless Normativity of Instrumental Rationality." *Journal of Philosophy*, vol. 98 (9): 445–468.

Hulse, Donovan, Cynthia Read & Timothy Schroeder. 2004. "The Impossibility of Conscious Desire." *American Philosophical Quarterly*, vol. 41 (1): 73–80.

Hursthouse, Rosalind. 1991. "Arational Actions." *The Journal of Philosophy*, vol. 88 (2): 57–68.

Jacobson, Hilla. 2013. "Killing the Messenger: Representationalism and the Painfulness of Pain." *The Philosophical Quarterly*, vol. 63 (252): 509–519.

Jeffrey, Richard C. 1983. *The Logic of Decision*. University of Chicago Press, Chicago, second edn.

Johnston, Mark. 2001. "The Authority of Affect." *Philosophy and Phenomenological Research*, vol. 63 (1): 181–214.

Jones, Karen. 2003. "Emotion, Weakness of Will, and the Normative Conception of Agency." In *Philosophy and the Emotions*, Anthony Hatzimoysis, editor, no. 52 in Royal Institute of Philosophy Supplement, 181–200. Cambridge UP, Cambridge.

Katsafanas, Paul. 2016. "Constitutivism About Practical Reasons." In *Oxford Handbook of Reasons and Normativity*, Daniel Star, editor. Oxford. Forthcoming.

Klein, Colin. 2007. "An Imperative Theory of Pain." *The Journal of Philosophy*, vol. 104 (10): 517–532.

—. 2015. *What the Body Commands*. MIT Press, Cambridge.

Kolodny, Niko. 2005. "Why Be Rational?" *Mind*, vol. 114 (455): 509–563.

Korcz, Keith Allen. 2015. "The Epistemic Basing Relation." In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, editor. CSLI Stanford, fall 2015 edn. URL `http://plato.stanford.edu/archives/fall2015/entries/basing-epistemic/`.

Korsgaard, Christine M. 1993. "The reasons we can share: An attack on the distinction between agent-relative and agent-neutral values." *Social Philosophy and Policy*, vol. 10 (1): 24–51. Reprinted in Korsgaard, Christine. 1996. Creating the Kingdom of Ends. Cambridge, Harvard UP: pp. 275-310. References to this edition.

—. 1996. *The Sources of Normativity*. Cambridge UP, Cambridge, UK.

—. 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford UP, New York, NY.

—. 2013. "The Relational Nature of the Good." In *Oxford Studies in Metaethics*, Russ Shafer-Landau, editor, vol. 8, 1–26. Oxford UP, New York.

Kriegel, Uriah. 2013. "The Phenomenal Intentionality Research Program." In *Phenomenal Intentionality*, Uriah Kriegel, editor, 1–26. Oxford UP, New York, NY.

Lavin, Douglas. 2004. "Practical Reason and the Possibility of Error." *Ethics*, vol. 114 (3): 424–457.

Lawrence, Gavin. 1995. "The Rationality of Morality." In *Virtues and Reasons: Philippa Foot and Moral Theory*, Rosalind Hursthouse, Gavin Lawrence & Warren Quinn, editors, 89–148. Oxford UP, Oxford.

LeDoux, Joseph. 1998. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Touchstone, New York.

—. 2012. "Rethinking the Emotional Brain." *Neuron*, vol. 73 (4): 653–676.

Lenman, James. 2005. "The Saucer of Mud, The Kudzu Vine and the Uxorious Cheetah: Against Neo-Aristotelian Naturalism in Metaethics." *European Journal of Analytic Philosophy*, vol. 1 (2): 37–50.

Lewis, David. 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy*, vol. 50 (December): 249–58.

—. 1988. "Desire as Belief." *Mind*, vol. 97 (418): 323–32.

—. 1996. "Desire as Belief II." *Mind*, vol. 105 (418): 303–13.

Little, Margaret Olivia. 2000. "Moral Generalities Revisited." In *Moral Particularism*, Brad Hooker & Margaret Olivia Little, editors, 276–311. Oxford UP, Oxford.

Lycan, William G. 2003. "Dretske's Ways of Introspecting." In *Privileged Access*, Brie Gertler, editor, 15–29. Ashgate, Aldershot.

MacFarlane, John. 2007. "Relativism and Disagreement." *Philosophical Studies*, vol. 132 (1): 17–31.

MacIntyre, Alasdair. 1986. "The Intelligibility of Action." In *Rationality, Relativism, and the Human Sciences*, Michael Krausz, Richard M. Burian & Joseph Margolis, editors, 63–80. M. Nijhoff, Dordrecht.

Martínez, Manolo. 2011. "Imperative Content and the Painfulness of Pain." *Phenomenology and the Cognitive Sciences*, vol. 10 (1): 67–90.

—. 2015. "Pains as reasons." *Philosophical Studies*, vol. 117 (9): 2261–2274.

Massin, Olivier. 2016. "Desires, Values and Norms." In *The Nature of Desire*, Federico Lauria & Julien Deonna, editors. Oxford UP, Oxford. Forthcoming.

Matsumoto, David. 2009. "Affect." In *The Cambridge Dictionary of Psychology*, David Matsumoto, editor, 19–20. Cambridge UP, Cambridge.

McDowell, John. 1979. "Virtue and Reason." *The Monist*, vol. 62 (3): 331–350.

—. 1985. "Values and Secondary Qualities." In *Morality and Objectivity*, Ted Honderich, editor, 110–129. Routledge and Kegan Paul, London.

—. 1996. *Mind and World*. Harvard UP, Cambridge.

Melzack, Ronald & Patrick D. Wall. 2008. *The Challenge of Pain*. Penguin, London, revised edn.

Millgram, Elijah. 2012. "Practical Reason and the Structure of Actions." In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, editor. CSLI Stanford, summer 2012 edn. URL `http://plato.stanford.edu/archives/sum2012/entries/practical-reason-action/`.

Millikan, Ruth G. 1984. *Language, Thought and Other Biological Categories*. MIT Press.

Millikan, Ruth Garrett. 1995. "Pushmi-Pullyu Representations." *Philosophical Perspectives*, vol. 9: 185–200.

Montague, Michelle. 2007. "Against Propositionalism." *Noûs*, vol. 41 (3): 503–518.

—. 2014. "Evaluative Phenomenology." In *Emotion and Value*, 32–51. Oxford UP, New York.

Nagel, Thomas. 1970. *The Possibility of Altruism*. Princeton UP, Princeton, NJ.

—. 1986. *The View from Nowhere*. Oxford UP, New York.

Navratilova, Edita & Frank Porreca. 2014. "Reward and motivation in pain and pain relief." *Nature Neuroscience*, vol. 17 (10): 1304–1312.

Neumann, Roland, Jens Förster & Fritz Strack. 2003. "Motor Compatibility: The Bidirectional Link Between Behavior and Evaluation." In *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*, Jochen Musch & Karl Christoph Klauer, editors, 371–392. Lawrence Erlbaum, Mahwah.

Nummenmaa, Lauri, Enrico Glerean, Riitta Hari & Jari K. Hietanen. 2014. "Bodily maps of emotions." *Proceedings of the National Academy of Sciences*, vol. 111 (2): 646–651.

Nussbaum, Martha C. 2015. "Transitional Anger." *Journal of the American Philosophical Association*, vol. 1 (1): 41–56.

Oddie, Graham. 2005. *Value, Reality, and Desire*. Oxford UP, Oxford, UK.

Orsi, Francesco. 2015. "The Guise of the Good." *Philosophy Compass*, vol. 10 (10): 714–724.

O'Sullivan, Brendan & Robert Schroer. 2012. "Painful Reasons: Representationalism as a Theory of Pain." *The Philosophical Quarterly*, vol. 62 (249): 737–758.

Panksepp, Jaak. 1998. *Affective Neuroscience*. Series in Affective Science. Oxford UP, New York, NY.

Papineau, David. 1987. *Reality and Representation*. B. Blackwell, Oxford.

Parfit, Derek. 1984a. *Reasons and Persons*. Oxford UP, Oxford.

—. 1984b. "What Makes Someone's Life Go Best." In *Reasons and Persons*, 493–502. Oxford UP, Oxford.

Peacocke, Christopher. 1983. *Sense and Content: Experience, Thought, and Their Relations*. Oxford UP, New York.

—. 1992. “Scenarios, concepts and perception.” In *The Contents of Experience*, Tim Crane, editor, 105–135. Cambridge UP, Cambridge.

Ploner, M, H. J Freund & A Schnitzler. 1999. “Pain affect without pain sensation in a patient with a postcentral lesion.” *Pain*, vol. 81 (1–2): 211–214.

Portmore, Douglas W. 2011. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford UP, New York, NY.

Price, D. D., S. W. Harkins & C. Baker. 1987. “Sensory-affective relationships among different types of clinical and experimental pain.” *Pain*, vol. 28 (3): 297–307.

Price, Donald D. 2000. “Psychological and Neural Mechanisms of the Affective Dimension of Pain.” *Science*, vol. 288 (5472): 1769–1772.

Price, Donald D & Murat Aydede. 2005. “The Experimental Use of Introspection in the Scientific Study of Pain and Its Integration with Third-Person Methodologies: The Experiential-Phenomenological Approach.” In *Pain: New Essays on Its Nature and the Methodology of Its Study*, Murat Aydede, editor, 243–273. MIT Press, Cambridge.

Price, Donald D. & James J. Barrell. 2012. *Inner Experience and Neuroscience: Merging Both Perspectives*. MIT Press, Cambridge, MA.

Prinz, Jesse. 2004a. “Embodied Emotions.” In *Thinking About Feeling: Contemporary Philosophers on Emotions*, Robert C. Solomon, editor, 44–58. Oxford UP, Oxford.

—. 2004b. *Gut Reactions: A Perceptual Theory of Emotions*. Oxford UP, Oxford.

Quinn, Warren S. 1990. “The Puzzle of the Self-Torturer.” *Philosophical Studies*, vol. 59 (1): 79–90.

—. 1993. “Putting rationality in its place.” In *Value, Welfare, and Morality*, Christopher W. Morris & R.G. Frey, editors, 26–49. Cambridge UP, Cambridge.

Railton, Peter. 1997. “On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action.” In *Ethics and Practical Reason*, Garrett Cullity & Berys Gaut, editors, 53–80. Oxford UP, New York, NY.

—. 1998. “Red, Bitter, Good.” In *Response Dependence*, Roberto Casati & Christine Tappolet, editors, no. 3 in European Review of Philosophy, 67–84. CSLI Publications, Stanford.

—. 2012. “That Obscure Object, Desire.” *Proceedings and Addresses of the APA*, vol. 86 (2): 22–46.

Rawls, John. 1999. *A Theory of Justice*. Harvard UP, Cambridge, revised edn.

Raz, Joseph. 1997a. “Incommensurablity and Agency.” In *Incommensurability, Incomparability, and Practical Reason*, Ruth Chang, editor, 110–128. Harvard UP, Cambridge.

—. 1997b. "When We Are Ourselves: The Active and the Passive." *Proceedings of the Aristotelian Society, Supplementary Volumes*, vol. 71: 211–246. Reprinted in Raz, Joseph. Engaging Reason. New York: Oxford UP, pp. 5-22. References to this edition.

—. 1999. "Agency, Reason, and the Good." In *Engaging Reason*, 22–45. Oxford UP, New York, NY.

—. 2010. "On the Guise of the Good." In *Desire, Practical Reason, and the Good*, Sergio Tenenbaum, editor, 111–137. Oxford UP, New York, NY.

Ridge, Michael. 2011. "Reasons for Action: Agent-Neutral vs. Agent-Relative." In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, editor. CSLI Stanford, winter 2011 edn. URL `http://plato.stanford.edu/archives/win2011/entries/reasons-agent/`.

Ridge, Michael & Sean McKeever. 2006. *Principled Ethics: Generalism as a Regulative Ideal*. Oxford UP, Oxford.

Roberts, Robert C. 2003. *Emotions: An Essay in Aid of Moral Psychology*. Cambridge UP, Cambridge, UK.

Rolls, Edmund T. 2014. *Emotion and Decision-Making Explained*. Oxford UP, Oxford.

Rolls, Edmund T. & Fabian Grabenhorst. 2008. "The orbitofrontal cortex and beyond: from affect to decision-making." *Progress in Neurobiology*, vol. 86 (3): 216–244.

Rosenthal, David. 1991. "The independence of consciousness and sensory quality." In *Consciousness: Philosophical Issues*, Enrique Villaneuva, editor, vol. 1, 15–36. Ridgeview, Atascodero.

Roy, Mathieu. 2015. "Cerebral and Spinal Modulation of Pain by Emotions and Attention." In *Pain, Emotion and Cognition: A Complex Nexus*, Gisèle Pickering & Stephen Gibson, editors, 35–52. Springer, Cham.

Saemi, Amir. 2015. "Aiming at the good." *Canadian Journal of Philosophy*, 1–23.

Salmela, Mikko. 2011. "Can Emotion Be Modelled on Perception?" *Dialectica*, vol. 65 (1): 1–29.

Scanlon, Thomas M. 1998. *What We Owe To Each Other*. Harvard UP, Cambridge, MA.

Scarantino, Andrea. 2014. "The Motivational Theory of Emotions." In *Moral Psychology and Human Agency*, Justin D'Arms & Daniel Jacobson, editors, 156–185. Oxford UP, Oxford.

Schafer, Karl. 2013. "Perception and the Rational Force of Desire." *The Journal of Philosophy*, vol. 110 (5): 258–281.

Schapiro, Tamar. 2011. "Foregrounding Desire: A Defense of Kant's Incorporation Thesis." *Journal of Ethics*, vol. 15 (3): 147–167.

—. 2014. "What Are Theories of Desire Theories Of?" *Analytic Philosophy*, vol. 55 (2): 131–150.

Schilder, Paul & Erwin Stengel. 1928. "Schmerzasymbolie." *Zeitschrift für die gesamte*, vol. 113: 143–158.

Schroeder, Mark. 2007a. *Slaves of the Passions*. Oxford UP, New York, NY.

—. 2007b. "Teleology, Agent-Relative Value, and 'Good'." *Ethics*, vol. 117 (2): 265–295.

—. 2008. "How Does the Good Appear To Us?" *Social Theory and Practice*, vol. 34 (1): 119–130.

Schroeder, Timothy. 2004. *Three Faces of Desire*. Oxford UP, New York.

Schroeter, Laura, François Schroeter & Karen Jones. 2015. "Do Emotions Represent Values?" *Dialectica*, vol. 69 (3): 357–380.

Schueler, George F. 1996. *Desire: Its Role in Practical Reason and the Explanation of Action*. MIT Press, Cambridge.

Schwarz, Norbert & Gerald L. Clore. 1983. "Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states." *Journal of Personality and Social Psychology*, vol. 45 (3): 513–523.

—. 2003. "Mood as Information: 20 Years Later." *Psychological Inquiry*, vol. 14 (3/4): 296–303.

Schwarz, Norbert, J. Xu & H. Cho. 2005. "Diverging Inferences from Identical Inputs: The Role of Naive Theories." Würzburg, Germany. Paper presented at the conference of the European Association Experimental Social Psychology.

Schwitzgebel, Eric. 2008. "The Unreliability of Naive Introspection." *Philosophical Review*, vol. 117 (2): 245–273.

Seager, William. 2002. "Emotional introspection." *Consciousness and Cognition*, vol. 11 (4): 666–687.

Searle, John. 1983. *Intentionality*. Cambridge UP, Cambridge.

Sen, Amartya. 1982. "Rights and Agency." *Philosophy & Public Affairs*, vol. 11 (1): 3–39.

Setiya, Kieran. 2007. *Reasons without Rationalism*. Princeton UP, Princeton, NJ.

—. 2010. "Sympathy for the Devil." In *Desire, Practical Reason, and the Good*, Sergio Tenenbaum, editor, 82–110. Oxford UP, New York, NY.

Shah, Nishi & J. David Velleman. 2005. "Doxastic Deliberation." *Philosophical Review*, vol. 114 (4): 497–534.

Siegel, Susanna. 2010. *The Contents of Visual Experience*. Oxford UP, New York.

—. 2014. "Affordances and the Contents of Perception." In *Does Perception Have Content?*, Berit Brogaard, editor, 39–76. Oxford.

Simon, Herbert A. 1967. "Motivational and Emotional Controls on Cognition." *Psychological Review*, vol. 74 (1): 29–39.

Sinhababu, Neil. 2009. "The Humean Theory of Motivation Reformulated and Defended." *Philosophical Review*, vol. 118 (4): 465–500.

—. 2015. "Advantages of Propositionalism." *Pacific Philosophical Quarterly*, vol. 96 (1): 165–180.

Smith, Michael. 1987. "The Humean Theory of Motivation." *Mind*, vol. 96 (381): 36–61.

—. 1994. *The Moral Problem*. Blackwell, Malden, MA.

—. 2009. "Two Kinds of Consequentialism." *Philosophical Issues*, vol. 19: 257–272.

—. 2011. "Deontological Moral Obligations and Non-Welfarist Agent-Relative Values." *Ratio*, vol. 24: 351–363.

—. 2013. "A Constitutivist Theory of Reasons: Its Promise and Parts." *Law, Ethics and Philosophy*, vol. 1: 9–30.

—. 2015. "The Magic of Constitivism." *American Philosophical Quarterly*, vol. 52 (2): 187–200.

Smithies, Declan. 2012. "The Normative Role of Knowledge." *Noûs*, vol. 46 (2): 265–288.

Soames, Scott. 2011. "What Vagueness and Inconsistency Tell Us About Interpretation." In *Philosophical Foundations of Language in the Law*, Andrei Marmor & Scott Soames, editors, 31–57. Oxford UP, Oxford.

de Sousa, Ronald. 1987. *The Rationality of Emotion*. MIT Press, Cambridge, MA.

Stampe, Dennis W. 1987. "The Authority of Desire." *The Philosophical Review*, vol. 96 (3): 335–381.

Stocker, Michael. 2004. "Raz on the Intelligibility of Bad Acts." In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, R. Jay Wallace, Philip Pettit, Samuel Scheffler & Michael Smith, editors, 303–332. Oxford UP, New York.

—. 2008. "On the Intelligibility of Bad Acts." In *Moral Psychology Today: Essays on Value, Rational Choice, and the Will*, David K. Chan, editor, no. 110 in Philosophical Studies Series, 123–140. Springer, Dordrecht.

Strawson, Galen. 1994. *Mental Reality*. MIT Press, Cambridge, MA.

Street, Sharon. 2012. "Coming to Terms with Contingency: Humean Constructivism about Practical Reason." In *Constructivism in Practical Philosophy*, James Lenman & Yonatan Shemmer, editors, 40–59. Oxford UP, Oxford, UK.

Sussman, David. 2009. "For Badness' Sake." *Journal of Philosophy*, vol. 106 (11): 613–628.

Swenson, Adam. 2009. "Pain's Evils." *Utilitas*, vol. 21 (2): 197–216.

Tappolet, Christine. 2000. *Émotions et valeurs*. Presses Universitares de France, Paris.

—. 2003. "Emotions and the Intelligibility of Akratic Action." In *Weakness of the Will and Practical Irrationality*, Sarah Stroud & Christine Tappolet, editors, 97–120. Oxford UP, New York, NY.

—. 2010. "Emotion, Motivation, and Action: The Case of Fear." In *The Oxford Handbook of Philosophy of Emotion*, Peter Goldie, editor, 326–345. Oxford UP, Oxford.

—. 2015. "Value and Emotions." In *The Oxford Handbook of Value Theory*, Iwao Hirose & Jonas Olson, editors, 80–95. Oxford UP, New York.

Tenenbaum, Sergio. 2007. *Appearances of the Good*. Cambridge UP, Cambridge, UK.

—. 2008. "Appearing Good: A Reply to Schroeder." *Social Theory and Practice*, vol. 34 (1): 131–138.

—. 2013. "The Guise of the Good." In *The International Encyclopedia of Ethics*, Hugh LaFollette, editor. Wiley, Hoboken. URL `http://philpapers.org/archive/TENGOT.pdf`.

Tenenbaum, Sergio & Diana Raffman. 2012. "Vague Projects and the Puzzle of the Self-Torturer." *Ethics*, vol. 123 (1): 86–112.

Thomson, Judith Jarvis. 2008. *Normativity*. Open Court, LaSalle, IL.

Titchener, Edward Bradford. 1896. *An Outline of Psychology*. Macmillan, New York, NY.

Tye, Michael. 1995a. "A Representational Theory of Pains and Their Phenomenal Character." *Philosophical Perspectives*, vol. 9: 223–239.

—. 1995b. *Ten Problems of Consciousness*. MIT Press, Cambridge, MA.

—. 2003. "Blurry Images, Double Vision, and Other Oddities: New Problems for Representationalism?" In *Consciousness: New Philosophical Perspectives*, Quentin Smith & Alexandar Jokic, editors, 7–32. Oxford UP, New York.

—. 2005. "In Defense of Representationalism: A Reply to Commentaries." In *Pain: New Essays on Its Nature and the Methodology of Its Study*, 163–177. MIT Press, Cambridge, MA.

—. 2009. *Consciousness Revisited*. MIT Press, Cambridge, MA.

—. 2014. "Transparency, qualia realism and representationalism." *Philosophical Studies*, vol. 170 (1): 39–57.

Vadas, Melinda. 1984. "Affective and Nonaffective Desire." *Philosophy and Phenomenological Research*, vol. 45 (December): 273–280.

Velleman, J. David. 1992a. "The Guise of the Good." *Noûs*, vol. 26 (1): 3–26. Reprinted in Velleman, J. David. 2000. The Possibility of Practical Reason. Oxford UP, New York, NY. References to this edition.

—. 1992b. "What Happens When Someone Acts?" *Mind*, vol. 101 (403): 461–481. Reprinted in Velleman, J. David. 2000. The Possibility of Practical Reason. New York: Oxford University Press, 123-143. References to this edition.

—. 1996. "The Possibility of Practical Reason." *Ethics*, vol. 106 (4): 694–726. Reprinted in Velleman, J. David. (2000). The Possibility of Practical Reason. New York: Oxford UP, pp. 170-199.

Vogler, Candace. 2002. *Reasonably Vicious*. Harvard UP, Cambridge.

—. 2014. "Good and Bad in Human Action." *Proceedings of the American Catholic Philosophical Association*, vol. 87: 57–68.

Wade, James B., Linda M. Dougherty, C. Ray Archer & Donald D. Price. 1996. "Assessing the stages of pain processing: a multivariate analytical approach." *Pain*, vol. 68 (1): 157–167.

Wade, James B., Daniel L. Riddle, Donald D. Price & Levent Dumenci. 2011. "Role of pain catastrophizing during pain processing in a cohort of patients with chronic and severe arthritic knee pain." *Pain*, vol. 152 (2): 314–319.

Wall, Patrick D. 1979. "On the relation of injury to pain." *PAIN*, vol. 6 (3): 253–264.

Wallace, Jay. 2001. "Normativity, Commitment, and Instrumental Reason." *Philosophers' Imprint*, vol. 1 (4): 1–26.

Wallace, R. Jay. 2009. "The Publicity of Reasons." *Philosophical Perspectives*, vol. 23 (1): 471–497.

Watson, Gary. 2003. "The Work of the Will." In *Weakness of the Will and Practical Irrationality*, 172–200. Oxford UP, New York.

Wedgwood, Ralph. 2003. "Choosing Rationally and Choosing Correctly." In *Weakness of the Will and Practical Irrationality*, Sarah Stroud & Christine Tappolet, editors, 201–229. Oxford UP, New York.

—. 2006. "The Normative Force of Reasoning." *Noûs*, vol. 40 (4): 660–686.

Weinstein, Edwin A., Robert L. Kahn & Walter H. Slote. 1955. "Withdrawal, inattention, and pain asymbolia." *A.M.A. Archives of Neurology & Psychiatry*, vol. 74 (3): 235–248.

Wiggins, David. 1998. "A Sensible Subjectivism?" In *Needs, Values, Truth*, 185–214. Oxford UP, Oxford, third edn.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. Harvard UP, Cambridge.

Winkielman, Piotr, Kent C. Berridge & Julia L. Wilbarger. 2005. "Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value." *Personality & Social Psychology Bulletin*, vol. 31 (1): 121–135.

Zajonc, R. B. 1980. "Feeling and thinking: Preferences need no inferences." *American Psychologist*, vol. 35 (2): 151–175.

—. 2000. "Feeling and Thinking: Closing the Debate Over the Independence of Affect." In *Feeling and Thinking: The Role of Affect in Social Cognition*, Joseph P. Forgas, editor, 31–58. Cambridge University Press.

Zhu, Jing & Paul Thagard. 2002. "Emotion and Action." *Philosophical Psychology*, vol. 15 (1): 19–36.

Zimmerman, Michael J. 2015. "Value and Normativity." In *The Oxford Handbook of Value Theory*, Iwao Hirose & Jonas Olson, editors, 13–28. Oxford UP, New York.