Non-Ideal Practices: An Essay on Ethical Theory and Deliberation

Robbie Kubala

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018
ABSTRACT

Non-Ideal Practices: An Essay on Ethical Theory and Deliberation

Robbie Kubala

What role does ethical theory play in everyday deliberation? On the ideal view, agents are taken to have an overriding commitment to a theory that dictates the obligatory, permissible, and forbidden actions in every conceivable situation. I argue that the ideal view imposes undesirable psychological burdens, whereas a non-ideal view—on which agents act according to the norms of their local practices and appeal to theory only when those norms prove insufficient to resolve particular problems—does not. Inspired by J. S. Mill, I develop one non-ideal theory for practices of regret, toleration, punishment, and partiality.
# Table of Contents

Acknowledgments ........................................... ii

Introduction .................................................. 1

Chapter 1: Construction and Reconstruction in Ethical Theory ....... 9

Chapter 2: Justification for Non-Ideal Practices ...................... 61

Chapter 3: Partiality as a Non-Ideal Practice ....................... 114

Chapter 4: Objections and Replies ................................ 157

Bibliography .................................................. 180
Acknowledgments

Many people helped me along the way to a finished dissertation, and I am grateful to all of them, including those I have forgotten to mention due to availability bias.

My grandfather, Robert E. Wood, started it. From his impromptu childhood lectures on Plato to his 2002 summer course at the University of Dallas on “Philosophy and the Ethical Life” to our ongoing correspondence, I have been sustained for several decades now by his open ear and wise counsel. Though our intellectual paths have diverged, I am endlessly grateful for his inspiration and example. If I never consciously chose to pursue philosophy, but affirmed it only after the fact, it is because of my deep admiration of him.

I am fortunate to have a loving family who understood my decision to spend nearly a decade in graduate school and who continue to provide unstinting moral support. Thanks to Dan, Mary, and Aaron, and special thanks to my brother Daniel Kubala: our many conversations over the years, especially about consequentialism, have had an enormous influence on me. And while he isn’t technically a family member, Dan Roberts feels like one. He always makes me think.

At Boston College: Patrick Byrne, Arthur Madigan, Mark O’Connor, and Michael Resler encouraged and continue to encourage. What a dream to have had such warm and gracious mentors at a formative time in my life.

At St Andrews: Michael Wheeler, Adrian Haddock, and Ali Jawin.

At Cambridge: Hasok Chang, Henry Taylor, and especially Tim Crane.

I am daunted by the task of naming the friends and colleagues at Columbia who kept me going over the past seven years. A short list would have to include Mary Grace Albanese, Anuk
Arudpragasam, Adam Blazej, Jessica Collins, Julie Crawford, Lydia Goehr, Dhananjay Jagannathan, Evan Jewell, Pat Kitcher, Sandra Lindgren, Katie McIntyre, Jorge Morales, Matt Moss, Ignacio Ojea Quintana, Yoni Pasternak, Elliot Paul, Chris Peacocke, Shivani Radhakrishnan, Sébastien Rivat, Jenny Rhodes, and Achille Varzi. I also want to thank the extraordinary administrative staff, past and present, of both the Center for the Core Curriculum—Christine Butcher, Toni Gunthrope-Hardee, Jessica Ng, and David Truong—and the Department of Philosophy—Maia Bernstein, Asha Panduranga, Stacey Quartaro, and Clayton Rains. Stacey deserves special mention for her staggering efficiency, patient good humor, and invariable kindness. The prospect of naming friends and colleagues in the wider philosophical community is actually too daunting to consider, but I have to single out Andrew Huddleston for many enjoyable conversations that have helped me understand things in new ways.

Max Hayward, my erstwhile roommate, remains my best philosopher friend. Here’s to many more years of collaboration, conversation, and carousing.

I am lucky to have many friends outside the academy who have taught me much and whose presence alone is a gift. But here I’ll just name Tim Kelly, Amy Kyleen Lute, Anne Lyons, Tim Mooney, and Nicole Wong.

Thanks to my dissertation committee—Philip Kitcher, Michele Moody-Adams, and Macalester Bell—for probing comments and gracious encouragement. Thanks also to Justin Clarke-Doane and Elisabeth Ladenson for participating in the defense and contributing to a productive conversation.

Philip needs his own paragraph. I don’t believe in perfection, and I am dubious about the whole category, but if pressed to name the ideal supervisor I would immediately cave in, because
I cannot imagine a more generous, thoughtful, incisive, and kind philosopher. Philip is a model of how to be a charitable thinker without sacrificing rigor. Yet I have never been ashamed to send him extremely rough work, knowing that he will quickly be able to point the way forward. Philip, thank you for making my time as a Ph.D. student not only bearable but positively enjoyable.

Most of all, though, I thank Kyle Mahowald, who is the deepest source of support and delight in my life. For me, the special good distinctive of partiality as a practice is epistemic intimacy: knowing and being known by another person. I could not have reached that conclusion without knowing and being known by him. This dissertation is dedicated to him, and to Fogo.
Introduction

This is an essay about the relationship between ethical theory and deliberation. Deliberation—deciding on the basis of reasons—has been around ever since we became capable of more than instinctual response to environmental pressure. Ethical theory—a system of normative principles, some of which explain others—arrived on the scene much later, perhaps with Socrates or earlier religious traditions. Although there are plenty of cultural trends that emphasize spontaneity, habit, and unreflective action, everyone agrees that deliberation has some place in the good life, whereas many would do without ethical theory altogether. Critics of theory often observe that deliberating in the terms of an ethical theory—deciding on the basis of reasons of utility, autonomy, virtue, etc.—can work against achieving the very aims set by the theory, and defenders of theory typically respond by restricting theory’s deliberative role in various ways. But their exchanges presuppose the answer to a prior question, namely what role theory is supposed to play in everyday thought. Imagine that someone selects from the menu of theoretical options, settling for him- or herself the question of which theory is the correct one. What comes next? How should a commitment to that theory guide his or her ordinary life? Though I cannot pretend to a full answer, my arguments and discussions here are inspired by, and hopefully derive much of their interest from, that very large question.

My primary claim in this dissertation is that facts about our limited deliberative capacities should constrain both the role that ethical theory plays and the form that ethical theory takes. The first constraint is familiar from an older literature—ethical theory should offer a criterion of rightness without necessarily playing the role of decision procedure—although my account of the
details draws on more recent work. The second constraint is less familiar. Essentially, I hold that if theory is non-ideal in form, we can more readily integrate it into our everyday deliberative lives than if we take it to be ideal in form. I unpack these claims by way of giving an overview of the four chapters of the dissertation.

In Chapter 1, I develop a two-level model of ethical psychology that locates our ordinary motivating thoughts at the first level and ethical theorizing at the second, more reflective level. This model, which I take to be the best way to meet the desideratum that ethical theories should guide us without defeating the very aims they set, is common to philosophers as otherwise different as R. M. Hare and P. F. Strawson. But there is a crucial difference between them concerning theory’s role. For Hare, the second level is normatively fundamental: the only reason to have a first level is to implement an ideal criterion of rightness. This means that agents are required to reconstruct their everyday motivations and attitudes so as to best implement that theoretical criterion. Strawson denies this: the second level does not set an ideal standard but serves as a resource for explaining and thereby improving our practices. This means that agents are mostly permitted to act according to their local practices and to revise them only where problems arise. Adopting Hare’s version of the two-level model, I argue, is bad for us qua deliberators: it can cause an unhealthy bifurcation in our thinking, lead to alienation from our psychological makeup, and embroil us in rational inconsistency. On my diagnosis, these undesirable expected outcomes arise from taking theory to be ideal in form. This in turn motivates the idea, which I explore in Chapter 2, that theory should instead be non-ideal in form, serving not as an end-state standard of rightness from which we derive what to do in any
conceivable situation, but—in the spirit of the pragmatists—as a tool to diagnose problems with our current practices and frame hypotheses for how to ameliorate them.

Do facts about deliberation constrain the content of theory as well as its form? Here I am less sanguine. In a general sense of ‘constraint’, yes: if theory is to be capable of guiding us, then we must be able to deliberate in its terms at least sometimes. But nobody would deny that conditional (although some would deny the antecedent). In a more specific sense of ‘constraint’, no: different ethical theorists offer different fundamental value-orientations—eudaimonia for Aristotle, rational nature for Kant, happiness for Mill—and it is implausible to claim that the selection of a basic value is entirely settled by facts about deliberation. It is not the case, for instance, that in order to count as deliberating at all, we must universalize our maxims. Rather, it is an open question, for all we might say about deliberation, which basic value-orientation to affirm.

Thus although in Chapter 1 I try to remain neutral between competing ethical theories, in Chapter 2 I adopt a particular conception of theory’s content. Having argued that (i) theory should be non-ideal rather than ideal, I motivate the further claims that (ii) theory should be teleological rather than deontological, i.e., that considerations of goodness are the only fundamental determinants of rightness, and that (iii) theory should be practice-based rather than act-based, i.e., that theory’s criterion of rightness should be applied directly to practices and only indirectly to actions. I call the resulting theory Non-Ideal Practice Utilitarianism (NPU), show how it can derive a criterion of right action indirectly, and discuss how it fares better than standard maximizing act-utilitarianism with respect to practices of regret, toleration, and
punishment.\footnote{I recognize that some philosophers find any mention of ‘utilitarianism’ offensive, in virtue of its apparent reduction of the rich diversity of the thick goods we value to the thin notion of utility. But I prefer the term ‘utilitarianism’ to ‘consequentialism’, first, because it better connects my view to its historical predecessors (‘utilitarianism’ dates to the 18th century, while ‘consequentialism’ was first coined by Elizabeth Anscombe in an article of 1958), and second, because contemporary consequentialism, but not Millian utilitarianism, is wedded to the evaluation of consequences under strict deontic impartiality, i.e., under the idea that it makes no moral difference who is harmed or benefited (by an action). As I discuss further in Chapter 2, what matters to my theory is that rightness ultimately be derived from goods, which can be much more complex than pain and pleasure, and that there is a moral difference between self and other, i.e., that it makes a difference to my reasons whether, say, I am the one feeling vulnerable or you are.} Along the way, I offer readings of Mill and William James, though my primary purpose is not exegetical. Rather, I want to depict the psychology of an agent who might commit to these various theories and to evaluate the outcomes of his or her doing so. It turns out that, in virtue of its teleological component, adopting NPU requires agents to \textit{bracket}: not to treat some otherwise valid considerations as motivating reasons. But bracketing can seem psychologically puzzling, if not downright impossible—how is it possible to act on the basis of a reason not to act on the basis of a reason?—and so in the final section of Chapter 2 I respond to these further challenges.

Chapter 3 offers the lengthiest illustration of NPU, testing the theory against our practices of partiality. Here I show how NPU can vindicate our current practices to some degree while yielding new reasons to improve. Although I show how NPU offers a distinctive way of reconciling our partial commitments with a commitment to impartiality, thereby engaging with much recent philosophical discussion of partiality, my primary aim is to show how NPU guides deliberation. Chapter 4 considers various objections and replies, then brings the project full circle by returning to assess the developed NPU in light of the objections canvassed in Chapter 1.

Though my deepest influences are Mill and John Dewey, another of my guiding lights throughout is Bernard Williams, although I am inspired more by his general outlook than his...
particular arguments, which can be opaque. In particular, we are hampered by his own definition of theory in *Ethics and the Limits of Philosophy*: “An ethical theory is a theoretical account of what ethical thought and practice are, which account either implies a general test for the correctness of basic ethical beliefs and principles or else implies that there cannot be such a test” (1985: 72). Several difficulties arise, starting with the most grievous, namely the patent circularity of defining an ethical theory as a theoretical account. Relatedly, there is the oddity of claiming that the account implies either $x$ or $\neg x$, especially since the ‘negative’ ethical theories that Williams only briefly discusses, such as emotivism, all appear to be better classified as meta-ethical rather than first-order views. These concerns suggest that Williams’ definition is either too empty or too broad to be informative. Further, the definition itself seems to presuppose meta-ethical commitments in appealing to the realist notion of ‘correctness’ as opposed to other success terms. Likewise for the cognitivist focus on ‘beliefs’ as opposed to other attitudes. My own definition of ethical theory is much more minimal: its essential feature is that it offers a criterion of rightness—a ‘test’ of rightness, if you like—whether rightness attaches primarily to action or to some other evaluand, such as a motive, rule, or practice. Theories, whether ethical or scientific, aspire to explain more with less, in this case with a general principle from which more specific recommendations can be derived. Even when theory is understood in these minimal terms, however, Williams is resolutely anti-theory, whereas I want to secure a positive role for theory that does not fall prey to his many objections. Doing so requires going non-ideal.

It is a striking fact about recent philosophy that the debate over ideal and non-ideal theory has largely limited itself to political theory. There is admittedly a historical reason for this: it was John Rawls, in *A Theory of Justice*, who first made the distinction vivid, arguing that ideal theory...
—done under the twin assumptions of full compliance with the demands of justice and favorable natural and historical circumstances—provides “the only basis for the systematic grasp” of the actual pressing problems we face under non-ideal conditions (1971: 9). Yet Rawls himself is less ideal than some, since he claims that principles of justice should be tailored to realistic conditions of moderate scarcity and limited altruism, and more ideal than others. G. A. Cohen, who represents the farthest end of the ideal spectrum, holds that principles of justice are entirely independent of material conditions and other factual constraints. Williams, who himself represents the farthest end of the non-ideal spectrum, holds that ideal principles of justice are completely irrelevant to our real-world circumstances, since they cannot guide us in coping with oppressive existing power-structures and inequalities. Most philosophers today fall somewhere in between these two extremes: given that, contra Cohen, we want principles of justice to guide us, we have to ask precisely which facts are relevant to placing constraints on ideals. In moving the feasibility question squarely into ethics, it is my contention that facts about our deliberative lives are relevant to constraining our ideals.

Here my project intersects with recent work on reasons and guidance, particularly in my discussion of bracketing. Yet because my focus is the shape of ethical theory rather than the nature of reasons, I adopt a fairly minimal notion of deliberation: we deliberate in terms of a theory just when we represent the theory’s considerations in our minds and act on the basis of that deliberation. In order to be guided by an ethical theory, we must represent its terms to ourselves at some time or another, even if, as I hold, we should not do so at most times. But I do not take a stand on the relationship between reasons and (good) practical reasoning, and I avoid further complexities concerning the explanation of action, liberated by Anscombe’s dictum that
any intentional act can be placed under a myriad of descriptions and the pragmatist dictum that the success of inquiry is assessed in terms of the goals of that inquiry. Since my goal is to evaluate theories in terms of their effects on our deliberative lives, I have judged it better to remain neutral in many of these adjacent debates.

To some, it may be surprising that the non-ideal value orientation I adopt is utilitarian. Isn’t act-utilitarianism already non-ideal, at least in the sense that it realizes the good flexibly? That is, if act-utilitarianism requires taking whatever means are necessary to the overall good we can foresee now, then it doesn’t matter whether others are doing their fair share or whether we live in utopian conditions. That may be the case, but my objections target the fact that act-utilitarianism nonetheless sets an ideal criterion of rightness: what really, and finally, justifies our actions is that they promote the overall good, and all our motivation and deliberation must be instrumental to that aim. As I argue, not only does this offer a distorted picture of our deliberative psychology, but the very attempt to approximate that ideal can lead us astray: first, we are not entitled to assume that aiming at the ideal and failing is better than not shooting for the ideal at all, and second, deliberating in terms of ideals can yield false positives and false negatives in identifying problems with our current situation. Furthermore, existing discussions of non-ideal theory in ethics have had somewhat different understandings of the concept of the ‘non-ideal’ in mind. Non-ideal Kantianism, inspired by Rawls, has tended to focus on our ethical obligations under conditions of partial as opposed to full compliance: how far does our duty of beneficence extend in a world where most people are clearly falling short? And non-ideal virtue ethics has tended to focus on whether eudaimonia is attainable in our actual, unjust world, particularly for those subject to oppression based on race, gender, and other socially constructed categories.
Non-ideal utilitarianism, as I see it, takes the primary concept of the ‘non-ideal’ to be articulated in terms of transitional rather than end-state rightness: the goal of non-ideal utilitarianism, as for pragmatism more generally, is to make progress away from our imperfect state rather than progress toward a perfect state. While questions of compliance and realistic conditions are clearly pertinent to this investigation, they are secondary to this reconception of the notion of progress.

I conclude with one final note on the goals of the dissertation. As a work of philosophy, the main labor naturally consists in carving up conceptual space. But whereas political philosophy requires us to evaluate and reform real-world institutions, in ethics we can more easily attempt to implement the conclusions of theorizing in our own lives. I can only briefly report, autobiographically, that the transition away from ideal act-utilitarianism and toward accepting Non-Ideal Practice Utilitarianism has been a welcome one. NPU can reconcile the conservative thought that our current practices contain much of value with the progressive thought that there are innumerable ways in which they could be improved, whereas ideal act-utilitarianism tends to obsessively belabor the latter alone. In his essay on Mill, Isaiah Berlin writes: “He was committed to the answer that we can never tell (until we have tried) where greater truth or happiness (or any other form of experience) may lie. Finality is therefore in principle impossible: all solutions must be tentative and provisional” (1969 [1959]: 182). I think that this fact about ethics is equally true of ethical theorizing, and my hope is that the non-ideal theory offered in what follows captures something of that Millian spirit.
Chapter One: Construction and Reconstruction in Ethical Theory

… the function of theory is not to furnish a substitute for personal reflective choice but to be an instrument for rendering deliberation more effective and hence choice more intelligent.

— John Dewey and James Tufts, *Ethics*

The primary aim of ethical theorizing, as I understand it, is to identify a fundamental principle, or set of principles, that explains and justifies at least some particular ethical judgments. Ethical theorists differ in how revisionary they are with respect to common-sense morality, but most philosophers who go in for theory at all agree on this primary aim. A related disagreement—my focus in this dissertation—concerns how theory relates to everyday deliberation. Standardly, ethical theorists want to use their theories to derive what to do in any conceivable situation. They think that agents should have an overriding commitment to

In adopting this characterization, I follow Brad Hooker (2000: 4), although I do not rule out, as he does, moral pluralism, which has a plurality of principles at the fundamental level. I do rule out particularism, understood as the claim that there are no interesting fundamental moral principles. And although I use the terms ‘ethical’ and ‘moral’ interchangeably, both should be understood as referring to the broader domain of the good life, as opposed to the narrower domain of, say, what we owe to one another. I hasten to add that there are other valuable forms of ethical reflection that do not take the shape of a theory: stories, parables, poetry, images, and exemplars. Although in Chapter 2 I will express skepticism about reliance on exemplars for moral improvement, I believe that these other forms of reflection can also provide vital resources to help us live rightly, though they will not be my focus here. My aim in this dissertation is to reconceive the notion of a theory so that it can better serve as such a resource.

I cannot pretend to lay exclusive claim to the term ‘theory’, whose definition must at some point be a matter of stipulation. But I do claim that a focus on principles, and in particular (as I discuss shortly) criteria, of rightness captures a core sense of the term. Consider the introduction to *Three Methods of Ethics*, co-authored by Marcia Baron, Philip Pettit, and Michael Slote—advocates of Kantian, consequentialist, and virtue ethical theories, respectively—“Ever since Socrates (if not before), [philosophers] have sought a general criterion or criteria for distinguishing between right and wrong” (1997: 1).

Why should we be interested in theorizing at all? While some philosophers think that “brute curiosity” should be sufficient to push us to systematize our moral judgments (Hooker 2000: 21), I think the ultimate upshot must be practical. First, coming to identify a fundamental principle might help us to settle unresolved moral questions (Scheffler 1992). But second, the very attempt to systematize can help us to be more consistent in our behavior, where consistency is itself ethically desirable. See also fn. 5 below.
following their favored theory. And they hold that because of this commitment, our non-
theoretical, everyday modes of deliberation are justified only to the extent that and because they
are sanctioned by the theory. I call this combination of claims the Reconstructive view, because it
aims to reconstruct all our thought and action in light of a theory. There’s a distinct view,
however, on which agents decide what to do based on the norms of their local practices and use
ethical theories as tools to improve those practices only when problems arise. On this
Constructive view, the kinds of questions we can reasonably ask from the perspective of ethical
theory are limited. In this chapter, I argue against Reconstruction. Because we have to live in our
ethical house throughout the renovation, we should not attempt to rebuild it from the ground up.
Rather, as I will discuss in later chapters, we should make improvements to the existing
structure.

Before turning to these two views, I want to set the scene. Even casual observers of
philosophical practice over the past fifty years know that there is a variety of competing ethical
theories: utilitarianism, deontology, virtue ethics, contractualism, and so forth. Each theory offers
a distinct fundamental principle (e.g., ‘an act is right just when, and because, …’) that is abstract
in content: it appeals to a consideration such as utility, well-being, the virtuous person,
autonomy, rationality, and so forth. These theoretical considerations are situation-independent:
they are supposed to apply across the whole range of ethical situations, however those situations
are themselves identified. But it is not the case that we always deliberate before we act, and it is
not the case that, when we do deliberate, we always appeal to the considerations offered by
ethical theories. Ordinary ethical deliberation, when it occurs at all, tends to center around
concrete features of immediate situations, such as the perceived desires, needs, and claims of others and oneself.

These familiar facts give rise to a problem facing ethical theory, namely the *deliberative role problem*: given that the fundamental considerations offered by ethical theory are abstract and situation-independent, what role are those considerations supposed to play in everyday ethical life? In short, when and how are we supposed to think in terms of an ethical theory? There are two extremes to be avoided here, or so I assume. On the one hand, the theory should play some first-personal deliberative role, rather than simply offering agents terms in which to approve or disapprove of various actions, motives, and lives from a third-personal perspective—theories should be more than *idle spectators*.

On the other hand, the deliberative role played by the theory should not be such as to alienate agents from those things that the theory holds to be good. Critics have charged that deliberating in terms of maximizing utility, complying with the Categorical Imperative, or emulating the virtuous person undermines the motivations necessary for close personal relationships, precludes spontaneous beneficence, and produces a narcissistic concern with one’s own ethical merits.

As Michael Stocker famously argued, ethical theories require that we “try to embody their reasons in our motives—as opposed to simply seeing whether our or others’ lives would be approved of by the theories” (1976: 466). Although Stocker’s is probably the majority position on the question of whether (as opposed to how) ethical theory should play a role in deliberation, Clayton Littlejohn has recently articulated a contrary perspective: “I’ve always been more sympathetic to the view that says that . . . it didn’t matter whether the considerations that figured in our theories were accessible to agents or could help agents see what they ought to do” (2016: 727). Although I do not have an argument for Stocker’s position, I believe that the project of ethical theorizing would lose much of its import if it never purported to offer guidance for agents to lead better lives.

Besides Stocker (1976), other notable discussions, in the case of consequentialism and Kantianism, can be found in Williams (1981a) and Baron (1984). Doris (1998), Hurka (2001), and Keller (2007) make similar objections to virtue ethics, but this issue is complicated by the question of whether virtue ethicists actually offer a criterion of right action. For defenses of virtue ethics on this point, see Annas (2008), Martínez (2011), and Pettigrove (2011). I should note that my brief list here does not exhaust the kind of objections that have been raised toward deliberating in terms of an ethical theory.

---

5 As Michael Stocker famously argued, ethical theories require that we “try to embody their reasons in our motives—as opposed to simply seeing whether our or others’ lives would be approved of by the theories” (1976: 466). Although Stocker’s is probably the majority position on the question of whether (as opposed to how) ethical theory should play a role in deliberation, Clayton Littlejohn has recently articulated a contrary perspective: “I’ve always been more sympathetic to the view that says that . . . it didn’t matter whether the considerations that figured in our theories were accessible to agents or could help agents see what they ought to do” (2016: 727). Although I do not have an argument for Stocker’s position, I believe that the project of ethical theorizing would lose much of its import if it never purported to offer guidance for agents to lead better lives.

6 Besides Stocker (1976), other notable discussions, in the case of consequentialism and Kantianism, can be found in Williams (1981a) and Baron (1984). Doris (1998), Hurka (2001), and Keller (2007) make similar objections to virtue ethics, but this issue is complicated by the question of whether virtue ethicists actually offer a criterion of right action. For defenses of virtue ethics on this point, see Annas (2008), Martínez (2011), and Pettigrove (2011). I should note that my brief list here does not exhaust the kind of objections that have been raised toward deliberating in terms of an ethical theory.
explicit terms of an ethical theory can actually work against achieving the aims set by the theory. Trying to follow the theory will defeat one’s ability to carry it out.

The best way to respond to the deliberative role problem is to adopt a two-level picture of ethical thinking and to restrict the role of ethical theorizing to the second level. This allows that agents need not always think in terms of an ethical theory—thus avoiding the self-defeat objection—and yet insists that they should sometimes think in terms of an ethical theory—thus avoiding the idle spectator objection. The first level would comprise the considerations that ordinarily motivate us to act (‘it’s wrong to lie’, ‘I made a promise’, ‘she needs help’), and the second level would consist of the more abstract and situation-independent considerations, offered by ethical theory, that explain why the first level considerations have justifying force, when they do (‘lying violates the Categorical Imperative’, ‘promising evinces the virtue of loyalty’, ‘we should alleviate suffering when we do not have to sacrifice anything of comparable ethical importance’). The most basic distinction can be drawn in terms of the reasons we act on versus the reasons that (ultimately) justify us in acting.7

In this chapter, I first develop the elements of the two-level picture shared by the Constructive and Reconstructive views (section 1). I then discuss the difference between them, which concerns the relation between the two levels. On the Reconstructive view, which I associate with R. M. Hare, the first level of ethical thinking has to be justified in terms of the second level: what we ought to do is determined from the perspective of ethical theory alone. On

7 For more on the distinction I have in mind, see Star (2015), who distinguishes derivative from fundamental reasons. There are other kinds of ‘two-level’ theory in the literature that should be kept distinct, such as Korsgaard’s (1986a) distinction between ideal and non-ideal reasons and Schliesser’s (2006) distinction, following Adam Smith, between thick local morality and thin universal morality, though my distinction has important affinities with the latter. See the first section of Chapter 3 for much more on this.
the Constructive view, which I associate with P. F. Strawson, the first level has a default justification and the second level is used only as a resource to guide improvement: what we ought to do is determined both by our local practices and by ethical theory (section 2). Developing arguments from Bernard Williams, I argue that the Reconstructive view is bad for agents in various respects: it causes an unhealthy bifurcation in our thinking, leads to alienation from our psychological makeup, and embroils us in rational inconsistency. The Constructive view avoids these undesirable results (section 3). And although I direct my arguments at utilitarian theories, I discuss how they extend to others as well, including recent theories from Parfit and Scanlon (section 4). Underlying my arguments is a more basic contention, namely that in addition to evaluating ethical theories on what we might call ‘epistemic’ grounds—how well they do in fulfilling the primary aim of explaining and justifying the ethical judgments that need justification—we can evaluate them on ‘practical’ grounds, in terms of the effects that adopting them would have on our lives. In later chapters, I will argue that this assessment of the practical role that theory plays in deliberation thereby constrains the form that an adequate ethical theory should take.

1. Two Levels of Ethical Thinking

Everyone distinguishes motivating reasons from justifying reasons (sometimes called normative reasons). The former are the reasons for which we act: the considerations that spur us to choose particular actions. The latter are reasons to act: the considerations that genuinely count in favor of actions. I might perform the same act-type, such as giving a dollar to a homeless person on the street, for any number of motivating reasons. I might have just found the dollar and
feel generous, I might believe that I ought to give charity when I can, I might feel moved by the sight of the person in need, or I might be feeling guilty about some recent transgression and seeking to assuage my conscience. And that act-type might be right for any number of justifying reasons, according to different ethical theories. Perhaps I have a duty of beneficence, or I ought to be moved by empathy, or the virtuous person would give a dollar in the same situation. The point is that our motivating and justifying reasons are often distinct.

There is an enormous debate about whether or not motivating and justifying reasons must coincide in order for an action to be ethically worthy (Kant famously argued that they must, that I must act from duty). But even in cases where they do coincide, and one’s motivating reasons really are justifying—I help because I believe I have a duty of beneficence—there is a distinction between levels of justification. The fact that I made a promise to help you presumably counts as a justifying reason on nearly any view of ethics—the fact of making a promise counts in favor of carrying it out, even if other considerations are ultimately weightier—but ‘keep your promises’ does not necessarily belong to the fundamental set of principles of an ethical theory. This is because there is almost certainly a more fundamental explanation—one that offers the reasons that are ultimately justifying—for why one ought to keep one’s promises. And that explanation, whatever it is—perhaps it appeals to the general utility of promise-keeping, or the notion of the rationality of a perfect duty, or the self-interest of having others keep their promises to you—is not likely to serve as a motivating reason for anyone to keep a particular promise. Moreover, even if that explanation did motivate someone to keep a promise, there are probably other motivating reasons that are more effective in meeting the aims of the ethical theory, such as the simpler thought, ‘I made a promise’.
Although I do not think that any general argument can be made for the claim that, for any ethical theory, there will always be a gap between motivating and justifying reasons, it seems to me that there is a variety of psychological reasons why such a gap will be common in everyday ethical life. If, for any action, your motivating reasons are the considerations that most proximally or immediately prompt you to act, and your justifying reasons are the considerations that you would sincerely offer, to yourself and to others, to make your action defensible, then there are any number of situations in which these will come apart. Consider the following non-exhaustive list:

**Unthinking Action:** I spontaneously jump into a lake to rescue my partner from drowning, without any thought at all about what makes my act right.

**Inappropriate Motive:** Dropping to my knee to propose marriage to my (now-safely rescued) partner, I think of how our marriage will contribute to aggregate happiness, when I should just be thinking of my partner.

**Telic Indirection:** Although I believe that the aim of golfing with my partner (now spouse) is to spend time together outdoors, I find that I best achieve that aim by playing to win, and hence am motivated by competition.

**Indeterminate Justifying Reason:** My spouse has convinced me that utilitarianism is the correct ethical theory. But utility is too indeterminate a justifying reason to motivate me in particular cases.

---

8 I say you would offer these reasons ‘sincerely’ in order to exclude contexts in which we are wrongly trying to persuade others, or ourselves, of our own praiseworthiness or lack of blameworthiness.

9 These examples come from Williams (1981a), Stocker (1976), Railton (1984), and Mill (1991), respectively.
It might seem that we could map the distinction between motivating and justifying reasons onto the two levels of ethical thinking, but, unfortunately, matters are not so simple, because justifying reasons can have varying levels of justification. Recall that, as I characterized it, the second level of ethical thinking consists of the more abstract and situation-independent considerations, offered by ethical theory, that explain why first level considerations have justifying force. So the second level of ethical thinking contains the reasons that ultimately justify us in acting (utility, autonomy, rationality, virtue, and so forth). Other reasons (keeping promises, telling the truth, lending a hand) are less fundamental but nonetheless justifying—they render our actions defensible to others—but because these are often the reasons for which we act, they belong to the first level of ethical thinking, the level of the considerations that ordinarily motivate us to act.

With this clarification in mind, we can see why the specific content of each level of ethical thinking will vary from person to person. Some people might have no second level, because they never even have occasion to reflect on what makes their actions justifiable. Young children, for instance, may have no conception of why they have to perform the actions their caretakers recommend. On the other hand, some adults might have very complicated second levels, because their reflections on what makes actions right take the form of a theory, or aspire to. Some theorists might also have more developed first levels, because they are motivated by the kinds of considerations that do not enter into others’ thoughts. And what is a justifying reason for one person might be a motivating reason for another, or the same person at a different time—

\[\text{\footnotesize{10}}\] Clearly the distinction between the two levels of ethical thinking is not entirely precise, because it is not always clear whether someone has a theory, or what degree of structure to one’s justifying thoughts would be necessary for those thoughts to count as comprising a theory. Insofar as my aim here is to ask what role a fully-fledged ethical theory should play in the psychology of an agent who adopts it, I do not need to draw a precise distinction.
I might refrain from telling a lie because it occurs to me that lying hurts your feelings, whereas you have no need for such a motivating thought because you are never even tempted to lie.

Given this understanding of the first level of ethical thinking as comprising one’s motivating reasons, and the second level as comprising the ultimately justifying reasons offered by ethical theory, it will be helpful to introduce two new terms for distinguishing these levels. Here I want to map the two levels onto the better-known distinction between decision procedures and criteria of right action.\footnote{To my knowledge, Bales (1971) was the first to make this distinction explicit, though it plays a role in utilitarian theorizing going back to Bentham and Mill. Baron (1984) and Keller (2007) employ the distinction in characterizing deontology and virtue ethics, respectively, although it is more controversial within those ethical traditions.} A decision procedure is, in the most general terms, a method for choosing what to do. Some worry about the notion of a procedure, typically because of the mechanical or calculative connotations of the term (e.g., Annas 2004). But the more neutral term ‘method’ admits of a range of interpretations, from acting on pure instinct—such that we could speak of an animal’s decision procedure—to acting on a fully-specifiable algorithmic function. On a broad understanding of the term, every agent has a decision procedure, one which takes as inputs the considerations that ordinarily motivate him or her.

A criterion of right action is the most fundamental principle of an ethical theory.\footnote{Although I speak in this chapter of criteria of right action, strictly speaking these could be criteria of rightness as applied to any evaluand. Some theorists want to identify criteria of right rules, or character, or motives, or institutions, etc.} Speaking loosely, a utilitarian criterion holds that an act is right just when it brings about the greatest aggregate happiness (or, for consequentialists, the best outcomes more generally); a Kantian criterion that an act is right just when it is sanctioned by the Categorical Imperative (or, for deontologists, in accord with our moral duties more generally); and a virtue-ethical criterion...
that an act is right just when it would characteristically be performed by the virtuous person under the circumstances. These criteria purport to fulfill the primary aim of ethical theorizing: they explain why we should be confident in our various ethical judgments, by offering considerations in terms of which we can assess the rightness of our actions, and justify those judgments from a more impartial point of view. (I discuss impartiality at greater length in Chapter 3.)

So far, a criterion of right action is merely a piece of formalism. What role does it play in the life of an agent? As the self-defeat objection suggests, criteria are not always suitable to serve as our motivating reasons: deliberating in terms of maximizing utility, complying with the Categorical Imperative, or emulating the virtuous person can undermine the motivations necessary for close personal relationships, preclude spontaneous beneficence, and evince a narcissistic concern with one’s own ethical merits. *Inappropriate Motive*, above, reflects this. But even if we think at the first level most of the time, the second level can nonetheless play a deliberative role in its conditional impact. As agents, we have various commitments, to our professions and projects, to other people, to ethical ideals and outlooks. Many of these commitments play a role at the first level of ethical thinking: we are motivated directly by our projects, relationships, and ideals. But we also have beliefs, whether explicitly or only tacitly held, about the conditions under which we would revise those commitments.

Consider a romantic relationship. Although common parlance occasionally refers to ‘unconditional love’, in fact most of our personal relationships are conditional in various ways. If your partner underwent a radical personality change, or joined a wildly offensive radical cult, or (to take a more homely example) simply stopped loving you, you might well stop loving them, at
least over time. These possibilities suggest that your love is conditional on your partner’s continuity of personality, political affiliation, and degree of affection for you, and that after reaching some (hopefully distant) threshold along each of these dimensions, you would eventually cease to love him or her. These conditions are justifiable, but others are not. If you were to stop loving your partner if he gained ten pounds or got a new haircut, he might quite reasonably find this distressing. These just aren’t the kinds of conditions that, by themselves, merit a withdrawal of affection. Even if you and your partner have a perfectly contented relationship now, you might well face problems if you were seriously prepared to leave him under those conditions.  

I propose to understand criteria of right action in the same way, as principled conditions under which we would revise our ethical outlooks. It is a familiar fact of ethical life that people more often agree about what to do than why to do it.14 You and I might both agree that we ought to vote in the midterm elections, though we have different views about what justifies voting. Similarly, we more often agree about why to do something—we agree on our justifying reasons—than about the conditions under which we would continue to act on those justifying reasons. We both believe that we should vote because we have a civic duty, but you believe that duty is in place only when one candidate is clearly preferable to the others, whereas I believe that the duty to vote exists even when every candidate is equally hateful. The same structure can be used to describe a disagreement between, say, Kantians and rule-consequentialists about lying. In most

13 There are tricky questions, which need not detain us here, about what the upshot would be if your partner never learned of your conditional criteria. They would almost certainly have some dispositional behavioral effects, so the problem would be averted only if those dispositions were never actualized.

14 As Appiah (2006: 71) aptly puts it, “We can live together without agreeing on what the values are that make it good to live together; we can agree about what to do in most cases, without agreeing about why it is right.”
particular cases, the two will agree that lying is wrong, but they will differ about the conditions under which lying in general is wrong. Understanding criteria of right action as commitments to change one’s behavior when certain conditions obtain is essential to my claim that we can assess ethical theories on practical grounds as well as on epistemic grounds.\(^\text{15}\)

To summarize: on any view of ethical theorizing, we need to distinguish two levels of ethical thinking—in order to avoid the self-defeat and idle spectator objections—and we need to understand the second level not merely as a piece of formalism but as a commitment on the part of the agent. The theory has to ‘get inside’ the agent, and once it does we can ask about the practical effects that such a commitment might have. In the next section, I distinguish Reconstructive and Constructive views of the relationship between the two levels before going on to argue that since the former has negative effects on an agent who adopts it, it should be rejected.

2. Reconstruction vs. Construction

On the Reconstructive view, associated with Hare, we ought to be committed to the second level of ethical thinking alone. In fact, the only value the first level has is to provide a decision procedure for implementing the deliverances of the second level. Put in terms of reasons: what we have reason to do is decided by the second level only. On the Constructive view, which I associate with Strawson, we ought to be committed to both levels; neither should automatically have deliberative priority. Our first level thinking has value independently of our second level thinking, although the second level justifications can aid us in improving the first

\(^{15}\) I would certainly not deny that many debates about the content and formulation of particular criteria of rightness can proceed without reference to agents’ psychologies.
level. Put in terms of reasons: what we have reason to do is decided by both levels of thinking and varies with our aims.\textsuperscript{16}

The main impetus behind the Reconstructive view is to take a fully worked-out second level (i.e., an ethical theory) and attempt to develop a first level of ordinary ethical thinking that will allow agents to implement the criterion of right action. This is why I call the view ‘reconstructive’: our motivations must be rebuilt, or at least reanalyzed, as implementations of an ethical theory. Hare arrives at his criterion by reasoning \textit{a priori}, but other Reconstructive views could arrive at the criterion in other ways (such as reflective equilibrium or divine revelation). On any such view, we need a first level for the reason that we typically lack the information needed to act on theoretical principles or criteria, as in \textit{Indeterminate Justifying Reason} above (Smith 1989; Feldman 2012). Our confidence is greatest in our criteria of right action—agents have greater credence in those principles than in their everyday ethical judgments—and our task is to modify our ordinary motivational capacities, to whatever extent possible, to bring them in line. The simplest view of the first level would be a pure ‘rules of thumb’ approach, on which our everyday thinking takes the form of readily revisable heuristics. As I show momentarily, Hare’s own position is more complicated than this, because he recognizes that some motivational capacities are not readily revisable. This makes his view more interesting and my argument more difficult.

\textsuperscript{16} This is not, of course, the only difference between Harean and Strawsonian conceptions of ethics, but it is the most salient difference as regards their views of ethical psychology. I suspect that their deeper differences stem from the fact that Strawson is an empiricist, who holds that ethical standards are fallible and revisable, and Hare is a rationalist, who holds that the ethical standard is knowable \textit{a priori} through reflection on the logical features of ethical language.
On the Constructive view I favor, the purpose of ethical theorizing is not to adopt a criterion and then design or approve of a psychology that can implement it. It is to take our actual practices, and the kinds of thinking they typically involve, as reason-giving, and only then to search for a plausible criterion—such as happiness, rationality, or virtue—that can explain and justify those practices. Our confidence is greatest in our ordinary practices, which enjoy a kind of default justification. The Reconstructive and Constructive theorists could, in principle, agree on the same criterion at the second level of ethical thinking. But there would still be a significant difference in the relationship between the criterion and ordinary deliberation: the Reconstructive view would see the criterion as the only source of justification for those practices, whereas the Constructive view would see it as a resource for improving them. In the remainder of this section, I explain the similarities between these two views—the elements that Strawson and Hare have in common—as a way of developing them in greater detail. Here I will be remaining neutral as to the content of the second level, i.e., neutral between theories.

In “Freedom and Resentment” (1962), Strawson famously introduced the notions of the reactive attitudes and the objective attitude, which he elaborated in his Woodbridge Lectures (1985). I think there is a fruitful analogy between these two kinds of attitudes and the first and

---

17 They would probably arrive at it in different ways, however. In particular, the Strawsonian, but not the Harean, could arrive at the standard through reflective equilibrium. Hare rules out intuitive judgments as having any probative force independently of theory, and therefore forbids the use of reflective equilibrium (1981: 75-6).

18 It is common for anti-theorists in ethics to wonder what authority an ethical theory could possibly have for us, such that we might give up our ordinary ethical judgments for its sake. But the Constructive view avoids this objection, because the epistemic route from the first to the second level is built into the very development of an ethical theory. See Star (2015: 27) for more on this objection, as well as Baier (1985), Noble (1979), and Louden (1992).
second levels of ethical thinking, respectively. The reactive attitudes—gratitude, forgiveness, anger, resentment, and love, to name a few—are the emotionally-toned reactions that we experience in response to the perceived good or ill will of agents. We respond to the perceived good or ill will of others, either as that will is directed at ourselves (participant reactive attitudes) or at other people (vicarious reactive attitudes), and also to our own good or ill will (self-reactive attitudes), where this last category includes attitudes of shame, guilt, compunction, and a sense of obligation. These attitudes characterize much, if not necessarily all, ordinary ethical motivation: we act out of forgiveness, love, compunction, and so on; such attitudes yield motivating reasons for which we act. The objective attitude, by contrast, is a more detached point of view in which we suspend our reactive attitudes and see ourselves and others as less than fully responsible and instead as the product of external causal forces. We can experience both kinds of attitude in response to the same situation, but they are nonetheless “profoundly opposed” to each other (1962: 9; 1985: 36).

In Moral Thinking (1981), Hare distinguishes ‘intuitive’ and ‘critical’ levels of moral thought. At the intuitive level, agents think in terms of prima facie principles, which are internalized universal prescriptions of the form ‘φ’ or ‘do not φ’ (e.g., ‘do not break promises’).

---

19 Here I am indebted to recent work by Miller (2014a; 2014b), who first developed Hare’s moral psychology with reference to Strawson. My argument differs from Miller’s in at least three ways, however: I claim that Reconstructive views are bad even for non-consequentialist theories; I think that Williams’ objections are stronger than Miller acknowledges; and my protagonist is Strawson, not Hare.

20 In some places Hare also distinguishes a third level, the metaethical, at which a moral theory’s formal component is established—for Hare the logical properties of universalizability and prescriptivity characteristic of moral judgments—but he does not take this to be part of everyday moral thinking (1989b: 237). Hare is frequently labeled an act-utilitarian, but this conceals a number of complications. For instance, while Hare’s universal prescriptivism is taken to issue in “judgements which are the same as a careful act-utilitarian would make” (1981: 43), it is also supposed to combine the merits of both act- and rule-utilitarianism, and is elsewhere said to be “practically equivalent to a rule-utilitarianism whose rules are allowed to be of any required degree of specificity” (1989a: 222; see also 1988: 226).
Because these principles are internalized, they are accompanied by strong dispositions to act in accordance with them, such that agents tend to feel guilt and compunction when they are violated (1981: 38-9). Indeed, Hare uses the terms ‘intuitive dispositions’ and ‘intuitive principles’ interchangeably: to *have* an intuitive principle just is to have a disposition to experience various feelings. At the critical level, agents step back to assess the efficacy of their prima facie principles, often but not always prompted by a situation in which the principles conflict, and then either affirm those principles or select new ones to attempt to internalize (1981: 50). While Hare recommends that we reason on the basis of utilitarian considerations, he acknowledges that a Kantian, for instance, could select new intuitive dispositions on the basis of deontological reasoning. Hare says little about the emotions and dispositions that agents experience at the critical level, but there is no reason to believe that critical thinking cannot be emotionally inflected in a range of ways.

I want to highlight four features common to an ethical psychology that both Hare and Strawson could, arguably, endorse: (1) the first level of ethical thinking is contentful, affectively laden, and appears to give us reasons for action; (2) the second level of ethical thinking allows us to step back to assess the situations we face in order to better understand what we have reason to do, sometimes but not necessarily in response to perceived conflicts; (3) we can think at both levels at the same time, where sometimes this will reinforce the reasons we take ourselves to have from the first level and at other times overturn them; and (4) we think at the first level the vast majority of the time, and because our first-level attitudes are so deeply internalized, they are generally resilient and only sometimes overridable. I develop these features in turn.
Both the reactive attitudes and our intuitive dispositions are understood as points of view in a full-blooded sense: they are not bare affects but contentful attitudes that provide a way of seeing the world and generate reasons for actions and attitudes. As Strawson puts it, from the standpoint of the reactive attitudes, “human behavior appears as the proper object of all those personal and moral reactions, judgments and attitudes to which, as social beings, we are naturally prone,” and when we experience the world through that standpoint, “human actions and human agents appear as the bearers of objective moral properties” (1985: 35). For instance, if you step on my toe with the intention of hurting me, or call me by a derogatory name with the intention of insulting me, I will be disposed, or “naturally prone,” to experience the attitude of resentment and thereby see you as responsible for wronging me and thus deserving of blame. Further, these moral properties (wrongness, blameworthiness) will appear to me to generate reasons for action, such as a reason to retaliate in kind, or to appeal to a third party, or to punish you in some other fashion (1962: 22).

Both the objective attitude and critical thinking also provide ways of seeing the world. The objective attitude can be characterized both negatively, in terms of the inhibition of the reactive attitudes, and positively, in terms of the appearance of human behavior as “yet another range of natural phenomena to be studied” (1985: 35). Instead of treating you as a responsible agent, I understand your behavior as determined by forces external to your intentions: perhaps you stepped on my toe only because you accidentally lost your balance, or because you suffer from a persisting mental illness. Understanding you in this way will tend to inhibit my

21 Although Strawson is not terribly explicit about ‘objective moral properties’, his paradigm example is blameworthiness (cf. Miller 2014b: 8). Interestingly, although his Woodbridge Lectures discuss praiseworthiness, “Freedom and Resentment” focuses entirely on blame; praise is nowhere to be found.
resentment. Yet the objective attitude is not one of affectless detachment. It “may be emotionally toned in many ways” (1962: 10), for instance by pity (for your illness) or fear (that you will strike again). And the properties revealed by the objective attitude generate their own reasons for action, in this case reasons for treatment and control. But the primary feature of the objective attitude, as I understand it, is explanation, regardless of whether that explanation goes on to alter my emotions (“finding in that very understanding a relief from the strains of involvement”) or the reasons I take myself to have (“determining our policy accordingly” (1962: 13)). When I take up the objective attitude or critical thinking, I have available to me new resources for understanding why you behave as you do and how I ought to treat you.

In order to employ the notion of the objective attitude, I want to clarify two points that are somewhat murky in Strawson. First, we must contextualize his claim that from the point of view of the objective attitude, “moral evaluation has no place” (1985: 40). Clearly, we could not employ the objective attitude to think about the moral value of our actions and practices if we could not engage in moral evaluation from that attitude. To avoid this conclusion, we have to understand that Strawson uses the term ‘moral’ to refer specifically to the kinds of evaluations and reactions characteristic of the reactive attitudes. So his claim that moral evaluation has no place from the objective attitude is purely terminological. It is evident that normative evaluation has a place from the objective attitude, since we can reason about what to do from that attitude.

22 There are two possible interpretations here. Either it is partially constitutive of the objective attitude that it inhibits resentment, or taking up the objective attitude has no constitutive emotional effects but rather gives me reason not to feel resentment. Whether or not the objective attitude generates reasons against certain emotions, it is clear that taking up the objective attitude generates reasons for action.

23 This point is well-emphasized by Sommers (2007), who defends the desirability of living with an exclusively objective attitude, which can heighten our sense of compassion and diminish our sense of indignation. Sommers downplays the instrumentalizing aspects of the objective attitude, however, which makes his argumentative task much easier.
In fact, Strawson occasionally seems to assume, though this is not essential to his argument, that the kind of practical reasoning that takes place at the objective attitude is consequentialist: if we could, *per impossibile*, choose to abandon our commitment to the reactive attitudes as a class, “then we could choose rationally only in the light of an assessment of the gains and losses to human life, its enrichment or impoverishment” (1962: 14). Second, whatever the exact nature of its reasons, taking up the objective attitude can *disable* the force of the reasons generated from the perspective of the reactive attitudes: “particular reactive attitudes, and reactive attitudes in general, may be, and, sometimes, we judge, should be, inhibited” (1962: 11). When you step on my toe, and I believe you to be a ‘normal’ agent, I take myself to have a reason to seek redress, even if I ultimately do nothing in response. But when I understand that you were pushed, or that you suffer from a psychological disorder, the force of that reason is disabled, and I take myself to have no reason to seek redress, even if I ultimately attempt to punish you. These two clarifications go beyond the letter of what Strawson says, but I believe them to be in keeping with the spirit of his discussion.

(3) We can combine both levels of thinking simultaneously. Not only can we rapidly toggle back and forth between them, as I might when struggling to balance my reactive frustration with my considered belief that you are not a full agent, but we can, to some extent, and for a brief period, experience both kinds of attitude at the same time. Strawson speaks, in the

---

24 Although Miller wants to develop this line of interpretation, he rightly points out that Strawson does not provide an argument that reasoning from the objective attitude must take this consequentialist form (2014b: 13). So we should not rule out a non-consequentialist understanding of the reasons we have from the objective attitude, particularly if this framework is intended to be neutral across competing theories of practical reason.

25 To interpret Strawson charitably, we should not understand “reactive attitudes in general” here to refer to the whole framework of reactive attitudes, but a particular set of reactive attitudes toward a single agent, or class of agents (e.g., the mentally ill).
case of the moral reactive attitudes, of being able to “more easily secure the speculative or political gains of objectivity of view by a kind of setting on one side, rather than a total suspension, of those attitudes” (1962: 18-9). In such a case, I would continue to have access to both kinds of reasons and would continue to feel their force. Strawson briefly discusses two examples of a “straddle” between reactive and objective attitudes: adults’ treatment of young children, and psychoanalysts’ treatment of patients. In the former case, adults have to deal with creatures who are not yet but someday will be capable of full agency, and who therefore must at least occasionally be treated like adults, at least for educational purposes. In the latter case, the psychoanalyst takes up the objective attitude toward her patient in order to work toward a state of affairs in which the objective attitude is no longer necessary because the patient can again be treated as a full agent. Going beyond Strawson, we might imagine that the patient should take up the objective attitude toward herself as a means to building or rebuilding that sense of agency. These examples illustrate the use of the objective attitude as a tool, a resource that we can draw upon to improve our everyday functioning at the level of the reactive attitudes: taking the objective attitude toward children helps us, as adults, to enter into more complex agential relations with them in the future, just as taking the objective attitude toward ourselves helps us to live as more fully responsible agents.

(4) We think in terms of the first level far more often than in the second. The reactive attitudes are our default setting, comprising the vast majority of our mental lives, such that we can step back from them only briefly. As Strawson puts it, “Being human, we cannot, in the normal case, [take the objective attitude] for long, or altogether” (1962: 10). This suspension can happen relatively automatically—in response to a perceived lack of agency in another person—
or it can be willed, “for reasons of policy or curiosity or emotional self-defense” (1985: 34; cf. 1962: 10). This last point is worth stressing: we can choose to take up the objective attitude, even without the stimulus of a conflict situation. Using Hare’s terminology, this implies that insofar as ethical theorizing takes place at the level of critical thinking, it is a project we can engage in at will, but also one that we only occasionally engage in, if at all. In general, the reactive attitudes are resilient: we cannot give them up entirely, and even if we take up the objective attitude for a time, we will always find ourselves returning to the reactive attitudes. Indeed, Hare’s recognition that our intuitive dispositions are deep marks one of the most plausible elements of his ethical psychology. Whereas other utilitarians conceive of our intuitive dispositions as mere ‘rules of thumb’, or substitute some other motivationally shallow mental state, Hare insists, and is right to insist, that intuitive dispositions “are associated, owing to our upbringing, with very firm and deep dispositions and feelings,” whereas the violation of a rule of thumb “excites no compunction” (1981: 38).

Yet even though Hare and Strawson are in agreement about these four features, there remains a crucial difference between them, which will play a key role in my arguments that the Reconstructive view is bad for agents. It is true that Hare is at pains to emphasize that intuitive and critical thinking are not rival procedures but rather “elements in a common structure, each with its part to play” (1981: 44). We cannot always be engaged in critical ethical thinking, yet it is sometimes necessary for us to go beyond intuitive thinking. To use Hare’s terms, we are less

---

26 I borrow the apt term ‘resilience’ from Miller (2014a: 49). This feature of Strawson’s ethical psychology is indebted to Hume, who famously wrote in his Treatise (1.4.7) that although he cannot rationally justify his beliefs in causation, induction, and so forth, nature cures him “of this philosophical melancholy” and he returns to his ordinary habits of backgammon and dining. For Strawson, however, reflection at the level of the objective attitude is not idle, but can be used to ameliorate our first level practices.
than “archangels,” who always think at the critical level, but must be more than “proles,” who think only at the intuitive level. But for Hare, our most fundamental commitment ought to be to the level of critical thinking, which is “epistemologically prior” to intuitive thinking: we have reasons at the intuitive level only because the critical level sanctions them. And that commitment ought to be overriding: “the right or best way for us to live or act either in general or on a particular occasion is what the archangel would pronounce” (1981: 46). We should be able to use our ethical theory to derive, in tandem with the facts, what we ought to do in any conceivable situation. Further, we should value intuitive thinking only as a mechanism for implementing the criterion of right action given by critical thinking. There are two factors about human agents that explain the need for an intuitive level and, from the critical point of view, justify the deployment of intuitive dispositions: the psychological reason that we are beings of limited cognitive ability and the pragmatic reason that “forming relatively simple reaction-patterns” helps us to cope with the world (1981: 36). The fact that intuitive thinking, in general, must be justified on the basis of critical thinking marks a significant divergence from the Constructive picture.  

In summary, let me try to be maximally clear about my strategy in this section. I began with an intriguing analogy or isomorphism between Hare’s moral psychology and Strawson’s distinction between the reactive and objective attitudes. Both claim that it is appropriate for us to have two different modes of practical thinking: a less reflective mode that governs our day-to-day interactions and a more reflective mode that prescinds from the attitudes and feelings we

27 There are many additional elements of Hare’s broader picture that one might reject or choose to remain agnostic about. For example, he insists that at the critical level there really are no moral dilemmas, that archangels thinking properly would all reach the same moral conclusions, that moral language is necessarily prescriptive and universal, and that because of these logical features moral thinking requires identifying with everyone’s preferences (see, respectively, 1981: 26; 46; 21; 91). Although these elements are far from incidental to Hare’s overall theory, they are orthogonal to my purposes here.
typically experience. And I tried to elaborate a picture of the kind of psychology that both could endorse, a picture that offers us resources for the broader, non-exegetical project of understanding the psychology of following an ethical theory. This requires some divergences from Strawson’s own view, as I have tried to explain, particularly since Strawson’s primary goal is to intervene in a metaphysical debate, not an ethical one. It also requires divergences from Hare’s view, and I will argue for further divergences in the next section.

Still, three clarifications are in order now. First, I am not claiming that all practical thinking is two-level, simply that ethical theorizing must be. Plenty of ordinary instrumental reasoning (‘I’m booking a flight so I can go see my mother’) does not require a second level to explain what is right or wrong with it, though it should also be said that nearly any decision could be ethically relevant under some circumstances. Second, and relatedly, I am not committed to the claim that all reactive attitudes or motivating reasons are ethical in character. For one thing, different ethical theories will have differing stances about this (Kantians are likely to distinguish ethical motivation from prudential motivation, while virtue ethicists may class prudence as a virtue, and consequentialists will likely hold that all motivation is ethically assessable), and I am trying to remain neutral on first-order topics here. For another thing, arguably both Hare’s rules of thumb and Strawson’s reactive attitudes span varieties of interpersonal conduct broader than the narrowly ‘moral’, though that point is spoiled somewhat by the fact that Strawson at least holds that “the moral sentiments” is “quite a good name for that network of human attitudes” (1962: 26). Third, and to reiterate a point made above, not all justifying reasons are second-level reasons: some are ordinary motivating reasons internal to the reactive attitudes and outside the perspective of ethical theory. One might worry that Strawson
actually *denies* that justifying reasons could be second-level: “Inside the general structure or web of human attitudes and feelings of which I have been speaking, there is endless room for modification, redirection, criticism, and justification. But questions of justification are internal to the structure or relate to modifications internal to it” (1962: 25). This might suggest that second-level theory has no role to play in justification at all. But the final clause undercuts this reading: justification can be ‘external’ to the reactive attitudes so long as it relates to *modifications*, and not wholesale abandonment or affirmation, of them. Strawson develops this possibility in the final paragraph of “Freedom and Resentment”: “It is far from wrong to emphasize the efficacy of all those practices which express or manifest our moral attitudes, in regulating behaviour in ways considered desirable; or to add that when certain of our beliefs about the efficacy of some of these practices turn out to be false, then we may have good reason for dropping or modifying those practices” (1962: 27). This is precisely the basis for the Strawsonian view I will develop in Chapters 2 and 3: employing a second-level standard like ‘efficacy’ (and here Strawson clearly has some kind of utilitarian theory in mind) can be a good reason to drop or modify our practices piecemeal. But Strawson goes on: “What *is* wrong is to forget that these practices, and their reception, the reactions to them, really *are* expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes. … Indeed the very understanding of the kind of efficacy these expressions of our attitudes have turns on our remembering this” (1962: 27). What I will suggest in the next section is that Hare, and Reconstructive views more generally, are guilty of precisely the amnesia that Strawson diagnoses, and that this turns out to be bad for agents in a variety of respects.
3. Against Reconstruction

Regardless of how well ethical theories do on epistemic grounds—how well they explain, or explain away, our particular judgments—we can assess them in terms of the effects that adopting them would have on our lives. No one has done more than Williams to emphasize that our theorizing must be continuous with and answerable to practice. I now want to develop three objections, inspired by his writings, and show that they apply to the Reconstructive, but not the Constructive, view of ethical theorizing.

According to Williams, an agent cannot experience, at the intuitive level, all the dispositions and feelings associated with following or flouting prima facie principles and think, at the critical level, that those dispositions and feelings have merely instrumental value for spurring her to act in ways approved by the ethical theory she endorses. As he puts it pithily, “the thoughts are not stable under reflection” (1988: 190; cf. 1985: 110). I think there are at least three possible kinds of ‘instability’ that should concern us, in increasing order of seriousness: an unhealthy bifurcation in our thinking, alienation from our psychological makeup, and rational inconsistency.

Before turning to these objections, two clarifications are in order. First, although Williams targets Hare’s utilitarianism in particular, these worries will apply to any Reconstructive view that claims that our first level attitudes have only instrumental value for implementing second level criteria. Second, while I agree with Williams that our first level thinking can be reason-giving independently of our second level justifications, I also hold that ethical theory has a role to play in improving that first level thinking. But my responses will rely on the idea that while the first level does have instrumental value, it does not have only
instrumental value. Translating from Hare’s language into Strawson’s: at the first level we think in terms of the contentful reactive attitudes, which offer us various motivating reasons. But at the second level, we step back to think in terms of the objective attitude, which assesses which reactive attitudes we ought to have. In doing so, we treat our reactive attitudes instrumentally: we consider which attitudes to adopt in terms of how well they help us to realize some other aim, and we evaluate them in light of their success or failure in meeting that aim. But we do not treat our reactive attitudes only instrumentally, since they are the default mode in which we experience the world.

3.1 Bifurcation

One objection to the Reconstructive view is that as a matter of psychological fact, an agent cannot combine seeing the world through the intuitive dispositions and from the point of view of critical thinking: as such, there can only be a problematic bifurcation between them. As Williams puts it in *Ethics and the Limits of Philosophy*, “The dispositions help to form the character of an agent who has them, and they will do the job the theory has given them only if the agent does not see his character purely instrumentally, but sees the world from the point of view of that character” (1985: 108). Williams’ idea is that once an agent has taken up a critical point of view on her intuitive dispositions, she simply will not be able to experience them in the same way again without an unhealthy bifurcation in her thinking. Once the sentiment of blame, for instance, is conceived not as a fitting response to wrongdoing as such but as a device that prevents future wrongdoing, then we can no longer see blame as appropriate when we have been wronged. Instead, we will have to severely compartmentalize our beliefs; if we do not, our
overriding commitment to critical thinking will lead to obsessive vigilance about whether blaming someone could best promote the good, regardless of whether they are guilty of wrongdoing. Even if we try to compartmentalize, striving not to calculate whether each particular act is optimific, and instead adopt a policy of blaming people just when that seems intuitively appropriate, then knowing that the policy is sometimes overridable will pressure us to consider overriding it in any particular case, even when there is no appearance of conflict. Note that this problem does not arise for a pure rules of thumb approach to the first level of ethical thinking, because rules of thumb can much more easily be overridden without the affective residue that accompanies the violation of deep dispositions.

Judging from his own responses, Hare interprets Williams’ objection as referring to an empirical matter of psychological fact only. First, Hare insists that he himself can readily combine intuitive and critical thinking. He supports this experiential assertion with the example of a battle commander, whose strategic thinking is intended to serve as an analogy to ethical thinking (1981: 52; 1988: 289-90). The commander has, at the critical level, endorsed a prima facie principle to the effect that one should engage in offensive action whenever possible. At the intuitive level, therefore, he evinces all the dispositions characteristic of having internalized the principle: he desires to engage in offensive action, and he feels reluctance and perhaps even guilt when he misses a chance to provoke the enemy. But he can experience all those dispositions at the intuitive level at the same time as he decides, at the critical level, that in some particular case he should pass up the chance to attack. That is, the decision not to attack does not in general sap the motivational force of the intuitive dispositions, even though they can sometimes be overridden. Second, Hare claims, further, that critical thinking actually reinforces dispositions at
the intuitive level. If I know why, at the critical level, it is wrong not to lie, that supports my tendency at the intuitive level to feel bad about lying (1988: 259).

One might well think that Hare cannot have it both ways: he cannot claim that critical thinking can in some cases override our intuitive dispositions and in general reinforce the motivation those dispositions provide. But in any case, it is evident that Hare has no principled response to this objection beyond mere counter-assertion, because he offers no developed ethical psychology to justify his claims. This is where Strawson’s picture proves useful, with its analogy between the reactive attitudes and our intuitive dispositions, and between the objective attitude and critical thinking. If our intuitive dispositions are resilient, then they are motivationally insulated from the force of critical thinking. In general, we will experience our blame reactions as appropriate or inappropriate in the normal way, but if there is a conflict (say, whether to blame someone for something she did a long time ago), then we have a reason, from within the perspective of intuitive thinking itself, either to revise our reactions in light of other intuitive dispositions or to switch to critical thinking to adjudicate the case. But the insulation of our intuitive dispositions means that we should feel no general pressure to override them. If we adopt the Constructive picture, then it is not the case that “some kind of willed forgetting is needed” (Williams 1985: 109): the forgetting will be involuntary, just as returning from the objective attitude to the involvement of the reactive attitudes is involuntary. So as a matter of psychological fact, combination of the attitudes is indeed possible without unhealthy bifurcation.

---

28 Indeed, I take this to be the conclusion of Miller’s work, and the feature that most motivates his Strawsonian picture.
Yet while the previous section admitted that it is open to Hare to agree with Strawson on this psychological point, there are other problems for the Reconstructive view, stemming from its requirement to justify our ordinary motivations on the basis of theory.

3.2 Alienation

A second objection is that even if we could in fact combine the two levels of thinking, the combination would lead to alienation from our psychological makeup. Alienation would come about because taking up a critical perspective on one’s intuitive dispositions means recognizing a deficiency in those dispositions that one previously did not appreciate. But since those dispositions can never be fully replaced, in virtue of their resilience, then having an overriding commitment to an ethical theory means that regret will always be appropriate. And whether or not regret has negative motivational consequences, it is an undesirable consequence of adopting an ethical theory that committing to it makes regretting one’s psychology appropriate. I will argue that while the Reconstructive theorist can, through a complicated series of dialectical moves, avoid alienation in the motivational sense, only the Constructive theorist can insist that regretting one’s intuitive dispositions, as a class, is not warranted.

Admittedly, the severity of this particular objection will depend on the nature of the ethical theory (i.e., the second level of ethical thinking) in question and the concomitant degree of imperfection it assigns to our intuitive dispositions. It may not be especially pressing for virtue ethicists, if their criterion of right action allows that the fully virtuous person has the ordinary sort of motivations, and it will be pressing for Kantians only to the degree that their criterion of right action appeals to a demanding Categorical Imperative procedure distinct from those
ordinary motivations. Because the objection is probably most pressing for consequentialists whose criterion of right action makes reference to an aggregate notion, such as pleasure or well-being, I concentrate on consequentialism in this section, hoping that the relevant argumentative changes can be made, *mutatis mutandis*.

Return to the blame example. If we come to believe that our intuitive blame responses are not producing the best consequences, but we find that we cannot give them up and that we continue to feel blame sentiments intuitively, then not only would the alienation we experience be likely to weaken our commitment to our dispositions, such a response would be, from the perspective of second level thinking, appropriate or rationally warranted. As with the first reading of Williams’ objection, this will not apply to pure rules of thumb approaches, which assume that the agent is ultimately committed to the criterion of right action found at the level of critical thinking, but it does apply to anyone who represents our intuitive dispositions as resilient and not readily revisable.

These ideas are compressed and require unpacking. We can begin by putting this version of Williams’ objection in the form of a conditional:

**Williams’ conditional:** if an agent truly took up an instrumental perspective on her intuitive dispositions, then that *ought* to weaken her commitment to those dispositions.

The force of that ‘ought’ is not merely predictive, but rational. It is supposed to follow from the nature of taking an instrumental perspective on something. The task for the defender of any two-level ethical psychology is to show how to deny the conditional. One answer would be simple:

---

29 See Rawls (2000: 167-70) for a detailed description of this procedure.
given that we experience regret only at the second level of ethical thinking, and not at the level of our admittedly resilient ordinary motivations, then Williams’ conditional will simply not have bad motivational outcomes. While I agree that this response is open to the Reconstructive theorist, it does nothing to overcome Williams’ conditional.

To do so, the Reconstructive theorist should distinguish one sense in which the conditional is true from another in which it is false. In the fitting sense of ‘ought’, the Reconstructive theorist must agree that our commitment to our intuitive dispositions ought to be weakened, but in the telic sense of ‘ought’, it is not the case that our commitment ought to be weakened. Or to put the same point in a different idiom, taking up an instrumental perspective gives us reasons of fittingness to regret our psychology, but does not give us telic reasons to regret our psychology. (I use ‘ought to’ and ‘have reason to’ interchangeably in this section.) I first show that the Reconstructive theorist is indeed committed to the view that we have reasons of fittingness to regret our psychology.\(^{30}\) I then show how she can argue that we have telic reasons not to regret our psychology. If those telic reasons are weightier than the reasons of fittingness, such that we have overall reason not to weaken our commitment to our intuitive dispositions, then the Reconstructive theorist can partially, but only partially, answer Williams’ objection.

\(^{30}\) The question of whether we have reasons of fittingness to regret our psychology is potentially ambiguous. It should be read as (a) ‘is it fitting to regret our psychology, given the instrumental deficiencies of that psychology?’ and not as (b) ‘is it fitting to regret our psychology, given the unfittingness of the psychology itself?’ I take no stand in this section on whether our attitudes themselves (e.g., blame, compunction) are unfitting. Most consequentialists would hold that there is nothing intrinsically bad about having unfitting attitudes and would therefore answer (b) in the negative; an exception is Hurka (2001), who holds that some unfitting attitudes—loving the bad, hating the good—are intrinsically bad. Thanks to an anonymous reviewer for alerting me to this potential ambiguity.
The claim that it would be fitting to regret one’s intuitive dispositions as a class is supposed to explain why the consequent of Williams’ conditional is true. That is, (i) taking up an instrumental perspective, at the level of critical thinking, on one’s intuitive dispositions (ii) gives one reason to regret those dispositions, which (iii) gives one reason to have a weaker commitment to them. The connection between (ii) and (iii) is stipulative: if one has reason to regret something, such as a past action, character trait, or state of affairs, then one should be _ceteris paribus_ less committed to it than if one did not have reason to regret it. The substantive philosophical challenge is to explain the connection between (i) and (ii), in particular the notion of regret at issue.

Begin with the idea of an instrumental perspective. In general, to treat something instrumentally is to value it only conditionally on its ability to facilitate or realize some other value. If I have a set of Ginzu knives, I value them instrumentally if I value them only for their ability to cut vegetables efficiently and effectively. A simple counterfactual test determines whether I value them _only_ instrumentally: were the knives no longer able to facilitate the chopping of vegetables, would I still value them? If the answer is no, then I value them only instrumentally. If the answer is yes, then I must value them in respect of some other properties. For instance, perhaps I value them aesthetically, admiring the pleasingness of their form, or perhaps they have extrinsic sentimental value, because they were a gift from a friend, or perhaps I value them instrumentally for some other end.

Suppose now that the Ginzu knives are rusty and dull, but they are the only tools I have to prepare vegetables. This gives me reason to regret that they are not better. I would replace the knives with sharper ones if I could, but instead I have to live with this set if I am to continue to
eat vegetables. The analogy to our intuitive dispositions is straightforward. If the goal is acting ethically as recommended by, say, Hare’s two-level utilitarianism, then the best tools for the job would be archangelic powers of the sort Hare describes—“superhuman powers of thought, superhuman knowledge and no human weaknesses,” not even “partiality to self” (1981: 44)—and anything less would be to that extent regrettable. The archangels would discard their rusty intuitive dispositions (if they had any remaining) and replace them with sharper cognitive skills.

In the same way, we mortals might regret our intuitive dispositions: a utilitarian committed to impartial benevolence might wish that her capacity to imaginatively identify and sympathize with others was greater, so that she would be motivated to aid more people in need. In all these cases, adopting an ethical theory gives one reasons to regret the tools that one has.

By claiming that the utilitarian has reason to regret her intuitive dispositions, I mean that some property of those dispositions is regrettable: regret would be fitting or appropriate to feel, because it accurately presents its object as having certain evaluative features. The emotions in general present the world in a certain evaluative light, and regret in particular presents the world in a certain negatively valenced way.\(^{31}\) As I gloss the evaluative presentation of regret, it presents its object as inadequate to facilitating at least one of the agent’s aims. Thus, regret presents the Ginzu knives as inadequate to chopping, and the agent’s intuitive dispositions as inadequate to her second level aims. If the knives were sharp enough, then regret would not be fitting, and the agent would have no reason to regret the knives. On this account, regret is not essentially tied to agency, since not everything that we regret need be an action of ours. Thus I distinguish regret

\(^{31}\) For more on this picture, see D’Arms & Jacobson (2000).
from remorse, which is essentially tied to agency: remorse presents a past action of the agent’s as blameworthy.\footnote{I recognize that this may be a revisionary usage of regret. Bittner (1992), for instance, follows the standard usage in asking whether it is reasonable to regret bad things that one does; in my terms, he is discussing remorse. My usage is not lacking in philosophical precedent, however: Wallace claims that regret “can be occasioned by reflection either on impersonal misfortunes or on voluntary past performances” (2013: 16), and he finds support in the secondary definition in the Oxford English Dictionary, which characterizes regret as a response to “some external circumstance or event.” One’s own character or psychological makeup is, in the relevant sense, external or impersonal, because it stands beyond the power of one’s agential powers to fully control (although see Smith 2005 for further discussion). Thanks to Michele Moody-Adams for urging me to clarify this.}

On this understanding of regret, it has a conditional structure: if one’s aims are a certain way, then regret is fitting to the degree that the means to those aims are suboptimal. If my aim is to be a champion marathon runner, then I have reason to regret that my physical capacities are not greater. Similarly for ethical aims vis-à-vis psychological capacities.\footnote{Hints of this view can be found in Hume’s account of justice (Treatise of Human Nature, 3.2.2). For Hume, the need for justice is contingent, since it would be unnecessary if we could “increase to a sufficient degree the benevolence of men, or the bounty of nature.” In my sense, then, it is to some degree regrettable that we need justice—which for Hume largely consists in respecting property rights—at all.} So on the fitting reading, Williams’ conditional comes out true for the Reconstructive theorist. On a different reading, however, the conditional comes out false, and it is not the case that one’s commitment to one’s intuitive dispositions ought to be weakened. This is the \textit{telic} reading, in which the consequences of what one does must be taken into account. A standard example in which the fitting and telic senses of ‘ought’ (or ‘reason’) come apart is that of an offensive joke. Suppose that some joke is funny, and offensive, and funny because it is offensive (its humor arises out of the stereotypes it employs, perhaps). In the fitting sense, I ought to laugh (I have reason to laugh), because laughter is an appropriate response to a humorous joke. But in the telic sense, it is not the case that I ought to laugh (I do not have reason to laugh), because things would go worse if I laughed, whether because I would be encouraging offensive humor or because I would...
be signaling to others that I am the sort of person who approves of offensive humor, or so forth.\(^{34}\)

Telic reasons are, we might say, grounded in their conduciveness toward our ends.\(^{35}\) Thus, whether a reason is telic depends on what our ends actually are.

Achieving our ethical ends requires making use of the tools that we have at our disposal. This claim should make it evident how to deny Williams’ conditional on the telic reading. What makes the consequent of the conditional objectionable at all, as I understand Williams, is that a weakened commitment to our intuitive dispositions is thought to produce a kind of paralysis, shaking our confidence in our intuitive dispositions as a class and thereby decreasing our motivation to act. Rather than being able to trust in the tools at our disposal, we would feel alienated from our very psychological makeup, perhaps suffering a psychic split with debilitating consequences. Yet if these are the expected outcomes of indulging in the emotional profile characteristic of regret, then the Reconstructive theorist has ample resources to claim that we should not give in to regret, even though we believe that certain features of our psychology are in fact regrettable. That is, the outcomes that Williams worries about simply need not arise for a mature agent capable of the self-mastery required (the serenity to accept the things she cannot change, perhaps).

One more premise is needed to deny Williams’ conditional, however, which is to claim that we have *overall* reason not to regret our intuitive dispositions. Issues of weighing reasons

---

\(^{34}\) One admitted complication of the example is that some might insist that laughter is not even a fitting response to an offensive joke. For defense of this claim, see Gaut (2007: Ch. 10); for the opposing view, which would support my example, see Jacobson (1997).

\(^{35}\) I borrow this locution from Muñoz (2018), who is inspired by Raz’s distinction between practical (i.e., telic) and adaptive (i.e., fitting) reasons (2011: 36-58).
are famously difficult, yet in this instance the Reconstructive theorist would have to insist that the balance of reasons speaks against regret, on the basis of the considerations just offered.

The Reconstructive theorist can deny Williams’ conditional, but only after an exhausting amount of dialectical work. Yet even if an agent could engage in some kind of cognitive therapy so as not to sap her motivations to act, surely there is still some normative failure in being made to feel alienated from one’s very psychological makeup on account of a commitment to an ethical theory. On the Constructive view, however, no such regret is warranted. If ethical theorizing is seen as a tool for improving our practices, and not the only source of their justification, then we do not have to insist upon the evaluation from which our intuitive dispositions, as a class, look imperfect. Of course it is still intelligible to regret individual instances of those dispositions, from within our practices—one might hope to become less susceptible to bouts of anger or resentment—but there will be no reason for regret simply because one adopts an ethical theory.

3.3 Inconsistency

A third objection is that even if agents can as a matter of fact combine thinking at both levels, and even if they can overcome the potential alienation occasioned by regret over their imperfect intuitive dispositions, their thinking will be rationally inconsistent. As Williams puts it in “The Structure of Hare’s Theory”:

you cannot think in these [intuitive] terms if at the same time you apply to the [intuitive] process the kind of thorough [critical] reflection that this theory itself advocates. That is not a merely psychological claim. It is a philosophical claim,
about what is involved in effective and adequate reflection on these particular states of mind (1988: 190).

Williams’ objection here is notoriously elusive. I have attempted to pin it down by labeling in brackets his allusions to intuitive and critical thinking. We can do even better by distinguishing two different claims Williams might intend by “you cannot think in these terms.” On the one hand, he might be suggesting that an agent would, from the perspective of critical thinking, have to consider the view from intuitive thinking to be illusory, and therefore could not rationally think in intuitive terms at all. On the other hand, he might be charging that if an agent were actually to apply the method of critical thinking to her intuitive dispositions, she would not find them justified, and therefore could not rationally think in these intuitive terms. As with the other versions of Williams’ objection, this would not be a problem for pure rules of thumb approaches, insofar as such heuristics would be put in place by an agent who is ready to admit that heuristic attitudes do not reveal the true ethical world and who relies on them only to the extent that they can be justified by critical thinking. But deep dispositions, as we have seen, reveal ‘objective moral properties’ that appear to generate reasons for action, and these reasons will at least sometimes be different in their content from the reasons we appreciate from the perspective of critical thinking.

3.3.1 The Inaccuracy Objection

The first reading of the rational consistency objection holds that even if we were to employ a criterion of right action, at the level of critical thinking, that approved all our intuitive

---

36 For expressions of frustration on this score, see Hare (1988: 290-1) and Zangwill (2013: 164 n. 17).
reactions, there would be a problem, because critical thinking insists that the reasons disclosed at the intuitive level are ultimately illusory. The objection was also developed by Thomas Nagel and has recently been pressed by Simon Keller and by Nick Zangwill; all three discuss the idea of illusion. Nagel admits that many of our intuitive dispositions have general utility. But, he continues, “It is very implausible to claim that intuitive repugnance at personal betrayal is just an artefact of upbringing on principles warranted by utility. Hare’s view requires us to regard the sense of immediacy that these claims have as a kind of illusion” (1982: 10). Keller says: “It would be both implausible and depressing to suggest that when we act well . . . we systematically misperceive our reasons” (2013: 27). And Zangwill claims that two-level consequentialist theories like Hare’s are “committed to a massive error theory about ordinary moral thought” (2011: 164), because they imply that most agents are wrong in their beliefs about what justifies their actions. Call this the ‘inaccuracy objection’.

As with the second version of Williams’ objection, the Reconstructive theorist can offer the motivational insulation response, which is that we do not, in fact, treat the deliverances of our reactive attitudes as illusory. Given that most of the time, we experience our reactive attitudes and intuitive dispositions in the ordinary way, acting on the reasons they offer without considering the reasons offered by the objective attitude and critical thinking, the worry about illusion simply does not apply. But Nagel, Keller, and Zangwill all suggest that we must, for the two-level theorist, be rationally committed to treating them as illusory. I argue that the Constructive view, but not the Reconstructive, has the resources to deny this as well.

\[37\] Indeed, Miller, although he acknowledges the possibility of this reading of Williams’ objection, finds it problematic only to the degree that believing one’s intuitive dispositions to be illusory leads to negative motivational consequences downstream. As with the second version of Williams’ objection, I think more needs to be said.
Zangwill offers the most developed formulation of this line of thought. He considers Cordelia, from Shakespeare’s *King Lear*, who loves her father “according to [her] bond.” The actions Cordelia has in the past performed for Lear—spending time with him, caring for him, treating him with respect—are reflectively self-endorsed by her and by what Zangwill calls ‘common-sense morality’, which we can identify with our ordinary reactive attitudes. And, let us suppose, those same actions would also be endorsed by second level ethical reflection. Even so, there will be a mismatch concerning the *grounds* of those actions. Cordelia, we suppose, judges that her actions are right in virtue of her bond with Lear. But the consequentialist judges that her actions are right in virtue of the fact that when people act on their loving bonds, things tend to go well. So there is a gap here.

Call the common-sense ground ‘B’ (for bond) and the consequentialist ground ‘C’ (for the general fact that things go well for everyone). Engaging in a bit of speculative belief-attribution, Zangwill supposes that Cordelia believes the conditional ‘if B, then M’, where M is some ethically relevant property, here the property of being under an obligation of some kind to Lear (and where the conditional is not believed to hold of necessity, hence it is not believed that B is a sufficient condition for M). But the two-level consequentialist—who accepts the different conditional ‘if C, then M’—allegedly cannot accept Cordelia’s conditional, for the reason that the grounds of Cordelia’s conditional are features intrinsic to her relationship, whereas the grounds of the consequentialist conditional are extrinsic to any particular relationship. They are more general considerations that obtain independently of any token relationship, considerations such as ‘piety toward fathers is good’. But such general facts seem, to intuitive thinking, to have nothing to do with Cordelia’s relationship to Lear; they do not typically show up in the contents
of reactive attitudes. That is, there appears to be a lack of fit between B (with its intrinsic features) and C (with its extrinsic features): “The Cordelia-Lear relationship is just a drop in the ocean of all the father-daughter relationships” (2011: 155). Hence, Zangwill concludes, anyone who comes to believe two-level consequentialism must give up belief in the common-sense conditional, ‘if B, then M’.  

On the Reconstructive view, an agent is indeed committed to believing that the content of the reasons given by first-level thinking is illusory, even if her ethical theory approves all her actions and is in no way revisionary. A Constructive Strawsonian view, by contrast, requires agents neither to see their intuitive thoughts as illusory nor to abandon all their ordinary beliefs about motivating reasons. The first level of ethical thinking does not derive its value only from implementing a second level standard, and it yields reasons for action independently of whether those reasons are given by the second level as well. This is the crucial difference between Hare’s conception of two-level thinking and Strawson’s: for Hare the intuitive level of ethical thinking must be justified from the point of view of critical thinking, whereas for Strawson the two perspectives are distinct and no such justification is required.

3.3.2 The Unjustifiability Objection

The second reading of the rational consistency objection holds that if an agent were actually to apply Harean critical thinking to her intuitive dispositions, she would not find them justified, and therefore could not rationally think in those intuitive terms. Call this the

38 Zangwill also considers a second argument against two-level consequentialism, namely that the ground of Cordelia’s conditional is indexical, whereas the ground of the consequentialist conditional is non-indexical, but he ultimately decides that this is weaker than the argument from intrinsicality.
‘unjustifiability objection’. This goes to the heart of the Reconstructive view, which requires that an agent’s everyday modes of deliberation must be justified on the basis of her theory. I first explain how the Constructive view differs before pressing the objection more forcefully.

The structure of Strawson’s argument is a matter of debate, but one way to read it employs an ‘ought implies can’ premise. If the reactive attitudes, as a class, are resilient—if “the general framework of attitudes itself is something we are given with the fact of human society” (1962: 25)—then they can never be given up entirely, even if we all came to believe that metaphysical determinism were true. But if we cannot give them up, then it is not the case that we ought to give them up: “As a whole, [the general framework] neither calls for, nor permits, an external ‘rational’ justification” (1962: 25; cf. 1985: 41). We cannot say either that the class of reactive attitudes, i.e., our practice of holding each other morally responsible, is justified or that it is unjustified: “it is useless to ask whether it would not be rational for us to do what it is not in our nature to (be able to) do” (1962: 20). It is not the case that we ought to give up our reactive attitudes, and it is not the case that we ought to keep them; the question of justification simply should not arise. The two points of view are simply different ways of thinking, which “tend at the limit to mutual exclusion” (1985: 36).

Yet while the framework of reactive attitudes is not subject to rational scrutiny, a major aim of Hare’s two-level picture of ethical thinking just is to subject the framework of intuitive dispositions, as a class, to rational scrutiny. In fact, Hare makes the stronger claim that it is only by subjecting intuitive thinking to scrutiny from the critical level that the intuitive dispositions can be justified: “The most fundamental objection to the one-level account of moral thinking called intuitionism is that it yields no way of answering” the question of which intuitions are
appropriate (1981: 39-40). Hare takes his primary opponent, the intuitionist, to be someone who is willing to rest content with various ethical intuitions, even when these conflict; he repeats a version of this charge in multiple places (e.g., 1981: 137; 1989a: 226; 1989b: 237-8).

Now one might object that I have established only that Hare is appealing to the need to justify particular intuitive principles, not the framework of intuitive thinking as a class, and Strawson similarly allows that particular reactive attitudes can be justified or unjustified. So there would be no disanalogy, according to this objection. But it is clear that Hare also believes that there are good reasons, from the perspective of critical thinking, to allow for the intuitive level. Recall that these are psychological and pragmatic reasons arising from our limited cognitive capacities and the need to rely on reaction-patterns in coping with the world. So there is a sense in which, according to critical thinking, intuitive thinking can be justified. But there is no sense in which someone adopting the objective perspective can see how the reactive attitudes as a class can be justified.

In fact, there is a further, even more significant disanalogy. Even if critical thinking can be said to justify intuitive thinking, in the sense of making intelligible what purpose it serves for the aims set by critical thinking, it is not the case that critical thinking fully endorses intuitive thinking. This is the place where the unjustifiability objection can be lodged. Hare clearly thinks that human beings would be ethically better if they were more like archangels, who do not need prima facie principles, because they confront each new moral situation with the ability to rapidly identify all its salient features, analyze the expected value of alternative actions, and frame a

---

39 Strawson says little about how to justify particular reactive attitudes—that is not his argumentative task—but presumably it will have to do with working out their appropriateness conditions, which are themselves internal to our practices.
universal and impersonal principle governing the rightness of their action. The most Hare can say for intuitive thinking is that it “certainly exists and is (humanly speaking) an essential part of the whole structure” (1981: 40). But it is clear that for Hare, it would be rational for us to become more purely archangelic, if that were possible.

Strawson makes no analogous claim about the objective attitude, however. In a curious footnote, he raises a question about the rationality of the objective attitude, bracketing the truth or falsity of determinism. (Recall that Strawson believes that accepting or denying determinism should have no effect on our belief in the rationality of the reactive attitudes as a class.) Strawson wonders whether, issues of determinism aside, “we should be nearer to being purely rational creatures in proportion as our relation to others was in fact dominated by the objective attitude” (1962: 28). Miller says of this quotation that Strawson “hastens to add that this does not entail that it would be rational for us to choose to dispense with the reactive attitudes entirely, because we have no choice in the matter” (2014a: 49). But this is not what Strawson says; that would be to repeat the same point. What Strawson does go on to say is that the answer to his question is yes: we might be more purely rational if we adopted the objective attitude more frequently, “only it would have to be added, once more, that if such a choice were possible, it would not necessarily be rational to choose to be more purely rational than we are” (1962: 28; emphasis added). It is difficult to be certain what Strawson means here, but I speculate that he believes that if we were to become more fully rational, there would be some “losses to human life” (1962: 14). The reactive attitudes—and here Strawson seems to take a step beyond his claim that there is nothing useful to be said about the justification of the reactive attitudes as a class—might be said to enrich human life. That is, the reactive attitudes are intrinsically valuable, so
trying to give them up would not constitute moral progress.\footnote{40}{Strawson’s later clarification of his footnote seems to bear this out: “When I said that the surrender of the reactive attitudes would bring us nearer to being ‘purely rational creatures’, I did not mean that some irrational elements would disappear from our lives; for I am not committed to the view that reactive attitudes are irrational. I meant only that some affective elements, elements of feeling or emotion, would disappear, leaving us more exclusively \emph{ratiocinative} creatures than before” (1980: 261). See Miller (2014b) for intriguing further discussion.} Whatever the correct interpretation of Strawson’s footnote may be, it is clear that Hare takes a different view: we should become as much like archangels as we can, given our current makeup, and beyond that it is regrettable that we have the makeup we do.\footnote{41}{Indeed, this view was precisely the basis for pressing Williams’ objection from alienation.}

To sum up, the most significant disanalogy between Hare’s and Strawson’s moral psychologies is that Strawson is a pluralist about practical reason, but Hare is not. Even if Strawson could be interpreted as holding that the reactive attitudes are irrational from the standpoint of the objective attitude, he would not want to say that the reactive attitudes are irrational \textit{tout court}. They would be irrational only relative to the objective attitude. But Hare maintains that our intuitive dispositions are irrational \textit{tout court}. Intuitive thinking, both as a class and as a series of individual dispositions, must be justified on the basis of critical thinking, and only to the extent that it is required by our psychological makeup. Furthermore, because the view from critical thinking shows the intuitive dispositions to be imperfect tools, critical thinking cannot fully affirm or endorse intuitive thinking. Intuitive thinking is a kind of necessary evil: we can justify relying on it because of contingent facts about who we are, but it would be even better if we could replace it.\footnote{42}{I am grateful to Dale Miller for helping me to formulate the claims made in this paragraph.}

When Williams objects that critical thinking cannot find our intuitive dispositions fully justified, there is therefore a sense in which Hare should agree. Hare explicitly recognizes that
intuitive thinking is not fully justified, from a critical point of view, and that it sometimes leads us morally astray. For example, he discusses a case in which, traveling to Czechoslovakia at a time when the country was under Communist rule and hostile to academic visitors, he was asked by the border officials the purpose of his visit (1981: 31). From the point of view of intuitive thinking, he would feel guilty if he told a lie, which, given his account of the intuitive dispositions, means that he thinks that he ought not to tell the lie. But he also judges, from the point of view of critical thinking, both that he ought to tell the lie, in order not to be deported from the country (i.e., that his intuitive dispositions ought to be overridden in this case), and that he ought to have been thinking that he ought to tell the lie (i.e., that it would be morally worse if he did not feel guilt at the prospect of lying). This means that Hare must think, critically, that in responding to those intuitive feelings in this case he would have been acting contrary to reason.

Other cases have the inverse structure, in which intuitive thinking is not critically justified and yet does not lead us morally astray. Hare considers various trolley-style rescue cases, in which, for instance, one has to choose between rescuing one’s family member and some larger number of strangers. To Hare, the utilitarian has the resources to claim that she would do the intuitively right thing, e.g., rescue her family member, because that is the course of action that flows from the best intuitive dispositions, and because these rescue cases are statistically rare and therefore not the subject of our customary intuitions. But nonetheless, the critically right thing to do would be to save the greatest number (1981: 138-9). In these cases, the right act—for Hare as an act-utilitarian, the very best act that could be performed—is one that no one with the best intuitive dispositions ever would perform.\footnote{For a similar analysis, see Railton (1984: 156-60). What is so fascinating about these cases is that they reveal that for the utilitarian, sometimes doing the right thing is both impossible and inappropriate.}
There are many objections that could be raised to Hare’s treatment of these cases, but I will remain neutral on the issues raised by his first-order reasons claims here. What I want to stress is Williams’ point that, given the possibility and potential frequency of these gaps between the critically right action and the course of action recommended by our intuitive dispositions, Hare has not justified his claim that critical thinking has selected these dispositions. In particular, he has not really undergone the process of critical thinking that he describes.

This challenge can be read in two ways: it’s a fantasy to imagine that we selected these dispositions, and in any case doubtful that we should have selected these dispositions. First, even if we could show that our intuitive dispositions had general utility, that would not establish that they were selected in any interesting sense, let alone selected by a procedure that aims to maximize general utility, or to implement any other second level criterion of right action. Second, it’s doubtful that critical thinking would select these dispositions, the ones that we as a matter of fact have. As Williams points out, if we start from the claim that critical thinking provides the correct account of right action, but that we cannot always think in critical terms, then there are “many styles of everyday moral thought that might in practice produce the best results” in applying that account of right action (1988: 191). But given that, by the consequentialist’s own lights, it ought to be an empirical question which kind of everyday thinking is the best stand-in for critical thinking, why should it be deep dispositions and not rules of thumb? It seems prima facie plausible that we would produce better consequences overall if we could entirely abandon our dispositions to blame and resentment, or even our entire concept of moral desert, and instead sanction others simply on the basis of whether punishing them would have greater expected utility. This is the source of Williams’ suspicion that Hare isn’t
really engaged in the process of reflection he thinks he is. In fact, Williams goes further and claims that Hare, like many utilitarians since Sidgwick, “makes pretty cavalier use of what are supposed to be evident matters of fact” in constructing his justifications for various existing practices (1995: 163). Hare himself admits that he wants to overcome “the objection to utilitarianism that it yields counter-intuitive prescriptions” (1988: 288), which provides some evidence that he is, despite appearances to the contrary, taking intuitive thinking as a given and constructing a justification for that, as opposed to undergoing the procedure of critical thinking in an open-minded way.

Given that an agent’s commitment to her theory is supposed to be overriding, then she faces a dilemma: either she hasn’t fully justified her ordinary dispositions, and by her own lights should not rationally think in those terms, or she has fully justified those dispositions, but by her own lights her ordinary thinking is illusory. In requiring agents to have an overriding commitment to the theories they adopt, the Reconstructive view has undesirable outcomes. These objections give us reason to abandon it. In the following chapters, I will develop one

---

44 As mentioned above, Miller mentions this objection but never explicitly discusses how his Strawsonian reading of Hare allows him to respond to it. As he nicely puts it, “Williams is alluding to the possibility that someone might give a kind of ‘lip service’ to the claims that only utility has intrinsic value and that her prima facie principles are only goads to encourage her to act in ways that are better from the utilitarian standpoint,” and thus “prove unwilling to approach the question of whether it might be better from the utilitarian standpoint for her to try to change [her prima facie principles] with the genuinely open mind that authentic critical thinking requires” (2014a: 45). I am about to suggest that taking Strawson seriously would require Hare to give up the idea that critical thinking, or any second level criterion of right action, should automatically have deliberative priority.

45 Hare’s justification for partiality, for instance, is remarkably flat-footed: “If mothers had the propensity to care equally for all the children in the world, it is unlikely that children would be as well provided for even as they are” (1981: 137). The facts that many societies raise children communally, or that equal care could be combined with targeted interventions on behalf of certain children, are ignored by Hare. Chapter 3 discusses partiality in much greater detail.
Constructive theory in greater detail. In closing, I want to briefly explain why the Constructive view in general does not face the objections discussed here.

Return to the Strawsonian example from above. If you step on my toe, then from the perspective of intuitive thinking I will take myself to have reason to seek redress, i.e., to treat you as an agent and to demand some kind of apology or restitution. (A minor infraction, to be sure, but the structure of the case remains constant even as the gravity of complaint rises.) Yet these are not the only moral reasons I might have in such a situation. For Strawson, further reasons are available when I take up the objective attitude, whether in response to a conflict or not. Taking up the objective attitude offers me an explanation that makes no reference to intentions, decisions, or other agential concepts. Instead, the behavior is explained in terms of extra-agential forces, and that explanation in turn might give me reasons to alter my emotions (reasons not to feel resentment) or my actions (reasons to avoid further engagement). This invites the question of which reasons I should act on, and hence the prior question of which attitude I have most reason to take up. Sometimes, of course, the answer will be settled by appeal to knowable facts: learning that someone (perhaps even oneself) suffers from an autism spectrum disorder or a post-traumatic stress disorder ought to elicit the objective attitude and its concomitant reasons for treatment and control. But sometimes such explanations are not available, in which case we face a choice. And when we face a choice, whether we should treat someone as an agent or as determined depends on what our aims are. If my aim is to persevere in a close personal relationship with you, and I find that you are acting aberrantly, then I should choose to take up the objective attitude only when I believe that doing so would further that aim. Taking up the objective attitude at all times is not only impossible, given who we are, but would
distance us from enjoying close personal relationships with others.\textsuperscript{46} And the aim of getting along with others is one we are given from within the intuitive attitude, with the “general framework of human life.” So we should take Strawson at his word when he calls the objective attitude a “resource” (1962: 10) for explanation and, potentially, conflict resolution.\textsuperscript{47} If the objective attitude is to give us further reasons, then those have to be informed by our aims at the intuitive level. Of course, one of our aims can be to follow an ethical theory, and ethical theories can even help us to set our aims. But on the Constructive view, we have reasons not to adopt the ideal of reconstructing our motivations in light of theory. To anticipate the next chapters, we do not need to see ourselves as making progress toward that distant ideal so much as progress away from our current, non-ideal practices.

4. Conclusion

Clearly much more needs to be said about how second level justifications are supposed to operate. The next chapters will do just that, as I explore how the Constructive view of the relationship between theory and deliberation constrains possibilities for the form of justification in general. In particular, I adopt a non-ideal, teleological, practice-based theory of the right (Chapter 2), and then apply a Constructive view of such a theory to our practices of partiality in particular (Chapter 3). This chapter has focused much more on the first level of ethical thinking, arguing that our ordinary motivations should be understood not as rules of thumb for implementing a second level criterion of right action, but as resilient attitudes that can be

\textsuperscript{46} Shabo (2012) provides a defense of this claim against Sommers (2007).

\textsuperscript{47} Other writers who urge this pragmatist conception of ethical theories as tools include Hämäläinen (2009) and, with reference to Mill’s consequentialism in particular, Kitcher (2010).
modified only piecemeal by our second level justifications. Understanding the two levels of ethical thinking in this way allows us to avoid not just the self-defeat and idle spectator objections, but various kinds of instability.

How serious are the objections presented here? It depends in part on how widespread the Reconstructive view is. An objector might worry that my challenge is fairly trivial: even if the Reconstructive view is no strawman, Hare is something of an outlier in contemporary philosophy. After all, almost nobody believes that moral language is necessarily prescriptive and universal, and that because of these logical features moral thinking requires identifying with everyone’s preferences. But those metaethical views are orthogonal to the Reconstructive view, which is neutral across competing theories. My objections apply to any ethical psychology on which an agent is supposed to be most committed to a second level criterion of rightness that requires a distinct set of motivations to be implementable. For instance, a Kantian two-level psychology might hold that the maxims behind one’s actions should all be tested against the Categorical Imperative and yet allow that in ordinary decision-making we ought to be motivated by our reactive attitudes. Although Kant himself seems to have held that our ethical feelings are, to the extent they are deserving of that name, “emanations of reason that are necessarily congruent with the Categorical Imperative” (Miller 2014a: 58-9), many contemporary Kantians would deny this (e.g., Herman 1981; Baron 1984), and if they do, they will be vulnerable to the objections discussed here. The same goes for virtue ethicists who concede that there is a gap between how the virtuous person would do something, with her ideal set of ethical sentiments, and how the less than fully virtuous person would do it (e.g., Driver 2006). So there are Reconstructive views of these non-consequentialist theories.
Others need to be considered on a case-by-case basis. Derek Parfit’s recent Triple Theory offers a criterion—“An act is wrong just when such acts are disallowed by some principle that is optimific, uniquely universally willable, and not reasonably rejectable” (2011a: 413)—that purports to be a “single supreme principle” (2011b: 155) of morality, a principle to which we should have an overriding commitment. Parfit recognizes the need for a “motive theory” to tell us “which motives we ought to have, and what we ought to be disposed to do” (2011a: 407), though he does not himself develop one. This strongly suggests the Reconstructive claim that our ordinary modes of deliberation are justified only to the extent they are sanctioned by the theory. The status of T. M. Scanlon’s contractualism is less certain, since his formula that an act if wrong if it would be disallowed by a set of principles that no one could reasonably reject is intended as an analysis of what it is for an act to be wrong; what makes an act wrong “are the properties that would make any principle that allow it one that it would be reasonable to reject” (1998: 391 n. 21). Since the contractualist criterion, unlike the others I considered, does not explain what fundamentally makes acts right or wrong, it may or may not count as a theory in the sense described here; it depends on whether and how it is supposed to guide our everyday deliberation. In either case, though, it seems that the Reconstructive view is sufficiently widespread to be worth the arguments I have offered against it.

Williams is right to deflate the ambition to reconstruct our entire ethical life in the light of theory, pointing us to those features that cannot be seen as fully justifiable from outside. We find ourselves dragging around too many unselected dispositions. Yet he would be wrong to believe that we cannot employ a utilitarian criterion, or any second level ethical criterion, as a tool—an instrument, to use the word of Dewey and Tufts from my epigraph—constrained by our intuitive
dispositions to react to others in ordinary ways. On the Constructive view I have sketched in this essay, we do not have to indulge in impossible feats of psychological division, or be alienated from our own psychology, or succumb to rational instability, in order to make use of theory. We only have to recognize that our first level thinking can have value, and be a source of reasons, independently of our ethical theorizing.
Chapter Two: Justification for Non-Ideal Practices

Take the love of drunkenness; take bashfulness, the terror of high places, the tendency to seasickness, to faint at the sight of blood, the susceptibility to musical sounds; take the emotion of the comical, the passion for poetry, for mathematics, or for metaphysics, — no one of these things can be wholly explained by either association or utility. They go with other things that can be so explained, no doubt; and some of them are prophetic of future utilities, since there is nothing in us for which some use may not be found.

— William James, “The Moral Philosopher and the Moral Life”

The denial of a Reconstructive view of ethical thinking—on which our intuitive dispositions have value only as an implementation of an ideal criterion of rightness—is compatible with a wide range of positive views about which criterion to adopt in a Constructive manner. One could be a Constructive Kantian, a Constructive virtue ethicist, a Constructive egoist, and so forth. In the remaining chapters of the dissertation, I will be advocating just one such positive view: a Constructive utilitarianism. On this view, we are in general permitted to act on the reasons given by our ordinary first-level practices, so long as those practices contribute to the good, but at the second level, we employ a utilitarian standard to guide our efforts at reform. My primary aim in this chapter is to motivate this view. Constructive utilitarianism is a non-ideal, teleological, and practice-based theory, and each of those components requires defense. I first argue that theory should be non-ideal rather than ideal, not only because of ideal theory’s practical and psychological burdens—rehearsed in the previous chapter—but because ideal theory falls prey to the fallacy of approximation and epistemic inadequacy (section 1). I then
argue that theory should be teleological rather than deontological, since the fundamental purpose of ethics is to bring about the good (section 2), and that theory should be practice-based rather than action-based (section 3). Though I believe that anticipations of each of these three features can be found in Strawson, they are better developed by John Stuart Mill and William James. Like Strawson, Mill and James believe that ethical theorizing should operate in a highly indirect way on our everyday thinking. But whereas the Strawsonian Constructive view, as I have characterized it so far, is indirect only as a matter of practice—it allows that decision procedures are separable from criteria of rightness—a Constructive utilitarian view is indirect as a matter of theory as well. That is, it applies the utilitarian standard not directly to particular actions but indirectly to rule-governed patterns of action and attitude, or practices.

Having argued for this kind of a utilitarian theory, I then claim—by reference to our practices of regret, toleration, and punishment—that it is preferable to a more familiar maximizing act-utilitarian standard (section 4). For each practice, on my view, agents are required to *bracket* certain considerations, that is, to refrain from treating some valid, undefeated reasons as motivating reasons to act. This gives rise to a different challenge, which is not that deliberation is unstable but that it is psychologically impossible, and I respond to this challenge as well, returning to issues of how we ought to deliberate in everyday life (section 5), before briefly summarizing my conclusions and setting up the following chapter (section 6). The overall theory that emerges in this chapter can be called Non-Ideal Practice Utilitarianism, on which agents are equally permitted both to follow some of the actual practices of their societies, even where these are to some degree imperfect, and to work toward reform, guided by the utilitarian standard.
1. Non-Ideal Theory

By ‘non-ideal’, I mean a distinctive approach to theorizing, one that aims not at establishing ideal end-state principles but at diagnosing problems in our current practices and proposing hypotheses about how to ameliorate those problems. My first goal is to motivate the pursuit of non-ideal theorizing in ethical theory. In recent philosophy, most of the debate in this area has been about political theory, inspired by Rawls’ claims about ideal principles of justice. Ideal political theory offers principles of justice for a perfectly just society; ideal ethical theory offers principles of rightness for perfectly right actions.

In political theory, the basic motivation for going non-ideal is that we want theory to be able to guide action. That is, theoretical principles should be sensitive to facts about what we can feasibly do now. Ideal theorists do not necessarily accept this motivation. For instance, G. A. Cohen, in Rescuing Justice and Equality, puts the point starkly: “the question for political philosophy is not what we should do but what we should think, even when what we should think makes no practical difference” (2008: 268). Cohen simply rejects the idea that principles of justice should be capable of guiding us, that they should be tailored to the motivational and cognitive capacities of human beings as they are. Elizabeth Anderson, in The Imperative of Integration, offers a forceful disagreement: “I do not advance principles and ideals for a perfectly just society, but ones that we need to cope with the injustices in our current world, and to move us to something better” (2010: 3). The non-ideal theorist wants principles that are normative for
us, capable of giving us reasons to act, and not merely *evaluative*, telling us what is or would be best.\footnote{Laura Valentini (2012) distinguishes three different notions of non-ideal versus ideal political theory. One is the Rawlsian distinction between conditions of partial and full compliance with the demands of justice; this is not at all my focus here, though it is perhaps the most frequent sense of ‘non-ideal’ in the contemporary literature (cf. Stemplowska 2017). A second is the distinction between realistic and utopian conceptions of justice, which do and do not, respectively, place feasibility constraints on what constitutes justice. A third is the distinction between transitional and end-state justice, and whether we should content ourselves with making progress from some imperfect state rather than progress to a perfect state. My focus is the third, although feasibility constraints are clearly relevant to non-ideal theory in this transitional sense, because they constrain the possibilities for reasonable progress from our current state. Notice that Hare’s archangels elide Valentini’s three notions: archangels are imagined as fully compliant with utopian principles of end-state rightness.}

Reframing this motivation in terms of ethical rather than political theory, our criteria of rightness must be capable of *guiding* us; otherwise they will not be able to play a deliberative role, since we could not necessarily act on principles for beings with different capacities. An ideal ethical theorist such as Hare might seem to have a straightforward response to this, namely that the very purpose of distinguishing intuitive from critical thinking is to tailor our critical principles to our intuitive capacities. After all, since we cannot be archangels, we have to develop *prima facie* principles—intuitive dispositions—to implement the deliverances of critical thinking.

The whole point of my previous chapter was to object to just this philosophical move: the gambit of identifying an ideal criterion of rightness independently of our motivational and cognitive capacities and then insisting that the only value our admittedly imperfect capacities have is as approximations of that ideal. For one thing, I argued that we have reason to be skeptical of our ability to undergo the justificatory procedure that Hare describes, as opposed to simply reanalyzing our motives in a *post hoc* way. We did not select most of our motivational and cognitive dispositions, and, given the possibility and potential frequency of gaps between the
critically right action and the course of action recommended by our intuitive dispositions, it is
doubtful that we would have selected these dispositions if we were really thinking critically. But
even if we did think we could develop a feasible implementation of an independently-formulated
ideal criterion of rightness, I argued that there are good reasons not to do so, since Hare’s gambit
results in: an unstable bifurcation between the two modes of thinking; the potential for alienation
from one’s intuitive dispositions simply on the basis of adopting an ethical theory; and a
commitment to believing that one’s ethical thinking is rationally inconsistent, because from the
point of view of critical thinking, one is rationally required to act on reasons that are illusory
from that same point of view.

My arguments in the last chapter, then, have already done a lot of work to motivate the
pursuit of non-ideal ethical theory: they suggest that we do not want principles of rightness (or
justice) for archangels, but for humans. This is not to scrap thinking in terms of ideals altogether.
Rather, it is to conceive of ideals differently, not as standards of assessment, outside of practice,
for any society (in the case of political philosophy) or any person (in the case of ethics), but as
diagnostic tools for identifying current deficiencies and suggesting hypothetical solutions to be
tested in practice. Still, even if one found the previous chapter’s objections to be unpersuasive,
however lengthily rehearsed, there are at least two other objections from political theory that
speak against the ability of ideal principles to guide us successfully. I want to show that these
objections apply to ethical theory as well.

The first objection is that when we employ ideal theory to evaluate our current practices,
we risk falling into the trap of thinking that the solution is to adopt policies aimed at directly
closing the gap between the ideal and the current reality. This trap goes by the name of the
fallacy of approximation: the unjustified assumption that trying to approximate a desired result will produce the best outcomes. As Adam Morton memorably puts it, the fallacy consists in “thinking that an attempt to approximate an ideal will lead to approximately ideal results” (2012: 25). Suppose that your doctor has prescribed a drug cocktail of three pills, but you only have two of the pills. It would be a mistake to infer that taking only two of the pills is better than taking none, even though the former is a closer approximation of the prescription than the latter. This is because the value contribution of each pill might depend on the presence of the others and could even have harmful effects if taken alone. In this case, the relevant kind of interaction is causal, but the fallacy applies to non-causal interaction as well. Take an example of aesthetic value: if I am trying to paint like Monet, I might do worse if I use his color palette and pattern of brushstrokes, but not his technique for layering paint, than if I simply captured his layering technique alone. Rather than directly approximating Monet’s style, I have to build up my technique in a different sequence if I want to paint like him. A third example comes from Anderson’s political philosophy: if you think the ideal society is a color-blind meritocracy, it doesn’t necessarily follow that the best way to achieve that is to immediately end race-conscious policies. Creating a genuine meritocracy might first require initiatives to rectify inequalities of opportunity between whites and people of color.

I think there are similar cases within ethics as well. The virtue ethical tradition has long maintained that the virtuous person should be seen as an exemplar, a model to be followed for the imperfectly (i.e., only approximately) virtuous.49 And many in the Kantian tradition claim that I should treat my fully rational self as an ideal to imitate (e.g., Korsgaard 1986b). But this

49 The literature on this topic is vast and still-growing. One important recent contribution is Linda Zagzebski’s Exemplarist Moral Theory (2017).
exemplar view, whether virtue-ethical or Kantian, commits the fallacy of approximation, at least if unqualified. Consider Michael Smith’s example of the angry squash player. Suppose I have just suffered a humiliating defeat and now have a consuming desire to smash my opponent in the face with a racket. If I were fully rational, or a virtuous person, then I would exercise good sportsmanship and walk over to shake my opponent’s hand. But that is not what I—an imperfectly virtuous and less-than-fully-rational person—should do, since if I walk over, I might well act on my desire! According to Smith, “what I have reason to do in my uncalm and uncool state is to smile politely and leave the scene as quickly as possible” (1995: 111). Trying to approximate the ideal would have worse results than taking my non-ideal circumstances into account.

Now the ideal theorist may have a response at this point, which is to reject the exemplar model in favor of a template model. David Estlund, for instance, now holds that the role of ideals is not “as a goal to be promoted or approximated, but as a template against which to identify epistemically offending features” (forthcoming: 20). I will argue momentarily that my second objection targets even the template model, but let me first note that retreating from the exemplar model is itself a significant concession to the non-ideal theorist, since it eliminates one crucial role for the would-be exemplar: namely, as a model for action. That is, exemplars no longer play a role for mere mortals qua ideal agents: due to the fallacy of approximation, we are not entitled to assume that we should act as they do. One might object that an exemplar could still play a guiding role, not as a model, but as a giver of advice. But notice that, for the exact same reason—the fallacy of approximation—exemplars’ lofty epistemic status as advisors cannot derive from their status as ideal agents. Neither we mere mortals nor the exemplars are entitled to assume
that there is a straightforward inference from the ideal course of action to identifying the second-best approximation to that course of action. Rather, both we and the exemplars would need to arrive at that knowledge in some other way. I am not suggesting that exemplars might not have special abilities that could help them give good advice, just that whatever those abilities are, they would be distinct from their ability to choose the ideally right thing to do in any situation.

Hare’s own claims are insensitive to the distinction between treating archangels as exemplars and listening to their advice. Some of his language appears to suggest that the archangels are advice-givers: “the right or best way for us to live or act either in general or on a particular occasion is what the archangel would pronounce to be so if he addressed himself to the question” (1981: 46). But what the archangel pronounces is not tailored to our imperfections, but is simply the right act, full stop. Tellingly, intuitive thinking, for Hare, “has the function of yielding a working approximation … for those of us who cannot think like archangels on a particular occasion” (1981: 46). For all he has said, therefore, Hare runs afoul of the fallacy of approximation by assuming that the archangels are exemplars whose style of thinking we should try to approximate.

It might be open to a Harean to radically revise Hare’s position and argue that the archangels should be treated as templates (ignoring the claims quoted in the previous paragraph). The second objection to ideal theory, however, speaks against using ideals even as templates against which to identify problematic features. The worry is that ideal theory might prevent us from recognizing injustices in our current, non-ideal world. That is, not only are ideals potentially unable to guide action, they can be positively distorting insofar as they can blind us to actual problems. In Rawls’ perfectly just society, some class inequality will exist, and thus the
principles of justice must be justified to members of each class position. But no racial positions exist in Rawls’ ideal society, and *a fortiori* the principles of justice do not need to be justified to members of each racial position. Thus Rawls’ theory of justice is, as Anderson puts it, “epistemologically disabling: it makes us blind to the existence of race-based injustice” (2010: 5). Anderson considers the racial segregation of many neighborhoods in the contemporary United States, a pattern that often occurs because whites prefer to live in neighborhoods that contain few non-whites. This preference would be permissible even in an ideal Rawlsian society, given rights of free speech and association. Now admittedly, a Rawlsian could identify other problems with segregated neighborhoods, such as susceptibility to welfare harms, class inequalities, or the assault on individual dignity produced by hostile treatment from others. But he could not identify the expressive injuries done to someone simply in virtue of belonging to a racially stigmatized group. Such injuries are, as Anderson says, “invisible from the position of the putatively raceless individual” (2010: 6). I call this the problem of false positives, in which an objectionable state of affairs, such as racially segregated housing, is not diagnosable as such by ideal theory.

Hare’s archangels are, as characterized, similarly blind to certain problems. First, since Hare does not situate them in a determinate context, we do not know what kind of social and political situation they face. Is this a society of scarcity or plenty, of unrest or stability? Second, they lack many recognizably human dispositions, including “partiality to self” (1981: 44), and thus might be ill-considered templates for those of us who do evince such partiality. Third, their

---

50 For a blistering critique of ideal theory on the grounds that it can blind us to the working of oppression in our actual context, see Mills (2004).

51 Indeed, Mill’s sensitivity to the moral asymmetry between self and other is one of many reasons to prefer his flavor of utilitarianism to Hare’s. I expand on this discussion of partiality at length in Chapter 3.
perfections more generally make them insensitive to what the imperfect should do. The ideally virtuous person may well have started from a position of mixed virtue and have retained some familiarity with imperfection, but the archangels are envisioned as existing *ab aeterno.* Not only angry squash players, but jealous lovers, embittered siblings, envious bankers, ruthless careerists, listless depressives, and raging narcissists are all potentially unable to perform the single fully right act identified by the archangels. Even if Hare’s archangels could diagnose these characters’ problems *as such,* and not as failures of universalizability or what have you, they would again face the fallacy of approximation. In other words, the archangels may be able to tell such people that they would be better off losing those traits, but what are they to do *now?* How to get from here to there?

Not only can ideal theory produce false positives in blithely overlooking actual problems, however; it can produce false negatives as well. This is underappreciated even by non-ideal theorists. The flip-side of failing to diagnose problems in our current context that would not obtain in an ideal context is an overzealous criticism of features of our current context that are not in fact problematic now, even though they may or may not be present in an ideal context. For instance, affirmative action policies are arguably valuable within our current context, but from the perspective of the ideal of color-blind meritocracy, they might seem unjust. False negatives will be crucial to my argument against maximizing act-utilitarianism in section 4 below, which holds that currently unproblematic practices of regret, toleration, and punishment can be

---

52 St. Paul is, arguably, alive to this point when he writes that “we do not have a high priest [i.e., Jesus] who is unable to sympathize with our weaknesses, but we have one who was tempted in every way that we are, yet was without sin” (Hebrews 4:15). Insofar as he knows how the imperfect can overcome temptation, Jesus is therefore a better ethical ideal than a Harean archangel.
approved by the Constructive utilitarian but would be condemned by the Reconstructive ideal utilitarian.

What about within ethics? How could ideal ethical theories be epistemologically disabling? There is a variety of possibilities here, but I’ll offer just one example of a false negative, again from the virtue-ethical tradition. On eudaimonist theories, there is a close connection between living virtuously and living well: one should live so as to flourish, where flourishing is partially constituted by beneficial living conditions and partially constituted by an exercise of the virtues. But under non-ideal conditions of oppression, flourishing may be unattainable on both fronts, whether because of pervasive harmful external conditions or because of the kind of damage to the virtues produced by widespread relationships of domination and subordination. And if flourishing requires reasoning well about practical matters, or having one’s desires formed properly, and one’s reasoning and desires are distorted by racist, sexist, classist, or homophobic ideology, then one may be unable to flourish. So eudaimonism might wrongly criticize agents for being unable to flourish under conditions of oppression (cf. Tessman 2009).

To sum up, the first objection to ideal theory claims that it can’t help us get from here to there; the second claims that ideal theory often misdiagnoses where we actually ought to be going. At this point in the debate, an ideal theorist might object that non-ideal theory is too tied to the status quo and that it has fallen into a ‘sour grapes’ scenario in which we insist that we don’t want what we can’t have. Isn’t there a need for some theory in identifying objectionable practices and reasonable alternatives to them? To this the non-ideal theorist responds of course! Non-ideal theory still endorses ideals. We just have to reconceive their function, not as standards of assessment for any society or individual, but as diagnostic tools for identifying current
deficiencies and suggesting hypothetical solutions to be tested in practice. Non-ideal theory insists that we do not need knowledge of the best in order to have knowledge of the better.\textsuperscript{53} We do not need an ideal of perfect health—what would that even consist in?—in order to diagnose and treat obesity, allergies, viral infections, and other conditions. I do not want to pretend that these objections to ideal theorizing are conclusive, but simply to claim that they provide further motivation for exploring the possibilities of non-ideal theorizing, particularly in ethics.\textsuperscript{54}

2. Teleological Theory

In this section, I want to motivate the idea that (non-ideal) theory should be teleological rather than deontological. Philosophers in recent years have made increasing refinements to our understanding of these two terms, but I propose to start with a simple contrast. Teleological theorists hold, and deontological theorists deny, that considerations of goodness are the only fundamental determinants of rightness. According to teleologists, what ultimately matters in ethics is good outcomes or consequences: there is no intrinsic value to following moral rules, behaving as the virtuous person does, or acting rationally. These aspects of ethics are valuable only insofar as, and because, they promote or instantiate good outcomes.

The most powerful argument for adopting a teleological approach to ethics is a simple one: it is better placed to justify the content of our morality than deontology can. Here’s a short version of the argument: teleologists about ethics can answer a question that deontologists

\textsuperscript{53} See Dewey (1910), Sen (2009), and Anderson (2010). For forceful disagreement on this point, see Simmons (2010); for further discussion, see Valentini (2012).

\textsuperscript{54} Indeed, the most recent work on this topic in political philosophy suggests that most players in the debate now think of the distinction as lying along a continuum, with different degrees of ideality suited to different purposes (see Weber & Vallier 2017). This is a welcome development, since it emphasizes the various uses to which theory can be put.
cannot, or can only answer obscurely, namely why we should follow the moral rules. The answer
that they are intrinsically valuable is unsatisfying, and the answer that they are grounded in the
value of rational agency is obscure. But teleologists can explain that wrong acts harm, and that
right acts benefit, those who do them or those to whom they are done. And it’s their badness and
goodness that, in some way, make acts wrong or right. A deontologist who held, plausibly, that
we should not harm others would either have to deny that being harmed is bad, or deny that the
wrongness of harming is explained by its tendency to make things worse. This gives teleology an
explanatory advantage: it can help to justify the deontological restrictions we do have, without
pretending that they have any value as such.

While the most familiar incarnation of teleological ethics—maximizing act-utilitarianism
—can seem to give the wrong verdict about particular moral cases, recommending that we
punish an innocent, torture a child, or turn a trolley, even here, as Christian Coons puts it, “we
must concede that the utilitarian has one heck of a point” (2012: 208). For if such acts would
really make things best—best for everyone or best overall—then why shouldn’t we do them? Or
as Simon Keller argues, if morality really advised us to do something other than what’s best for
everyone, then “would we not be better off with something other than morality” (2009: 91)? I
return to this point in section 3, where I suggest that the best way to resist these particular
verdicts is by changing the evaluand to which rightness applies. In the remainder of this section,
I want to discuss two historical proponents of the teleological view more generally as a way of
elaborating its theoretical core.

Perhaps the most famous proponent of a teleological criterion of rightness was Mill. For
all I have said so far, one could be a teleologist and an egoist about ethics, and hold that the good
is one’s personal good. Mill denies this: he holds that the good is the good of all. In particular, he is an *axiological impartialist*: everyone’s good, if equal in quantity and quality, is equally valuable. This claim is the cornerstone of utilitarianism, which amounts to a teleological criterion of rightness plus axiological impartiality.

Mill’s key idea is that happiness “is the justification, and ought to be the controller, of all ends, but is not itself the sole end” (VIII: 952). The claim that there are multiple ends could be interpreted as a hedonist claim, namely that we should choose all other ends only in light of happiness, or as a pluralist claim, namely that other ends can to some degree conflict with happiness. Furthermore, there are various complications in understanding the notion of happiness itself, whether in terms of the presence of occurrent pleasure and the avoidance of occurrent pain, the satisfaction of desire across a longer timespan, a broader notion of well-being, or something else entirely. I will try to remain fairly neutral on these issues in the theory of the good in favor of focusing on the theory of the right, although Chapter 3 argues that the goods of partiality have a much richer structure than is typically picked out by the thin notion of ‘utility’.

Some may wonder why I do not use the term ‘consequentialism’ rather than ‘utilitarianism’ for my theory, given that I am happy to speak of various rich goods. One reason is simply terminological: to emphasize continuity with the Millian tradition by using Mill’s own preferred term (perhaps akin to the way that Mill used Bentham’s term while adopting a much richer conception of utility). Another reason is substantive, however: following Daniel Jacobson (2008), I take consequentialism to be committed to *deontic impartiality*, the claim that everyone’s happiness counts in the same way when evaluating actions as morally right or wrong. Utilitarianism without consequentialism amounts to the affirmation of a teleological criterion of rightness plus axiological impartiality.
rightness and axiological impartiality, and the denial of deontic impartiality, most notably the claim that it makes no moral difference who is harmed or benefited. For Mill, as I read him, the asymmetry between self and other is morally significant: though my pleasure is just as valuable as your pleasure, I do not have the same reason to promote your pleasure as I do my own.\footnote{I recognize that this may be a controversial interpretation of Mill, and in future work I hope to turn my attention to exegetical issues concerning Mill’s utilitarianism. Here I can only briefly mention Jacobson’s sentimentalist reading, on which plausible norms for the fittingness of, say, guilt and resentment are constrained by the nature of our sentiments, which do make a self/other distinction. Not only is this reading more consistent with my Strawsonian Constructivism, it also explains Mill’s rather extreme antipaternalism: “Mill clearly implies that when a person ‘does mischief’ to himself, his action is not yet amenable to moral disapproval [in the narrow sense of what we owe to each other], though both agent and act can be criticized in other terms—as selfish, intemperate, or foolish” (2008: 179). This is consistent with Mill’s commitment to axiological impartiality.} Mill would reject the trope of the impartial spectator, who is indifferent to the distribution of happiness: a false positive in which the ideal exemplar or advice-giver is blind to actual problems.

William James picks up on the Millian notion of happiness in the essay from which my epigraph is taken.\footnote{In fact, James dedicated his 1907 Pragmatism “To the Memory of John Stuart Mill from whom I first learned the pragmatic openness of mind and whom my fancy likes to picture as our leader were he alive to-day.”} For James, the aim of those who seek “an ethical philosophy”—what I would call an ethical theory—is to weave “the moral relations that obtain among things” into “the unity of a stable system” (1956 [1897]: 184-5). James is explicit that this is a philosopher’s ideal, and one that may well go unrealized insofar as “the world resists reduction to the form of unity” (1956 [1897]: 185). It may turn out, that is, that our moral concepts are irreducibly plural. James distinguishes three questions in ethics, of which the most directly relevant to ethical theorizing in my sense is the casuistic question, which “asks what is the measure of the various goods and ills which men recognize, so that the philosopher may settle the true order of human
obligations” (1956 [1897]: 185). To ask this question is to wonder whether “the heap of things demanded”—the many interests, preferences, and needs of individual people—could prove less chaotic than they initially seem, whether “they furnished their own relative test and measure” (1956 [1897]: 200). The casuistic question asks, in short, for a criterion of rightness.

James surveys a number of potential responses, including recognizably virtue-ethical, Kantian, divine voluntarist, social Darwinist, and intuitionist criteria, and declares, “The best, on the whole, of these marks and measures of goodness seems to be the capacity to bring happiness” (1956 [1897]: 201). The non-ideal tenor of this sentence is clear: of the many possible criteria of rightness, happiness stands out as better than the rest. There is no implication that this is the ideal criterion independently of all particular times and places; James’ animadversions against a priority in ethics make that clear. Happiness is an ideal in the non-ideal sense: a hypothesis or solution to be aimed at in practice. But this idea has a Millian caveat: although happiness is the best criterion we have now, it cannot be “so used as to yield to the philosopher anything like a scientifically accurate and genuinely useful casuistic scale” (1956 [1897]: 201). James may have a variety of claims in mind here, but one is surely the denial of any kind of algorithmic decision procedure in ethical life. Another is the denial that happiness is the “one universal underlying kind of motive” (1956 [1897]: 201); as for Mill, happiness is not the ‘sole end’ in our deliberations. Still, for both Mill and James, it is currently the best ‘justification’ of our other ends.

57 The other two questions, the “psychological”—which inquires into the genealogical origins of our moral ideas—and what James idiosyncratically calls the “metaphysical”—the semantic question of what our moral words mean—are less directly relevant to first-order ethical theorizing.

58 Though James speaks of a test of ‘goodness’, I think the reference to settling the order of human obligations—a deontic notion—speaks in favor of interpreting him as offering a criterion of rightness.
However the theory of the good shakes out, what matters for non-ideal teleological theory is that happiness is an aim that can be affirmed at both levels of ethical thinking. Not only should it be endorsed at the second level of ethical theorizing, it is pervasive in our ordinary first-level practices. Rather than interpolate a criterion of rightness from outside our practices entirely, we isolate one element of our practices, make it explicit, and use it to revise our practices going forward. As T. M. Scanlon, a committed contractualist, puts it,

it seems to me that a large part of the appeal of utilitarianism lies in the fact that it identifies, in the idea of ‘the greatest happiness’, a substantive value which seems at the same time to be clearly connected to the content of morality and, when looked at from outside morality, to be something which is of obvious importance and value, capable of explaining the great importance that morality claims for itself (1998: 151).

Following Mill and James, therefore, I take axiological impartiality as an assumption going forward. Mill’s own proof has seemed flawed to many, and a better argument may be simply to insist that, as John Skorupski puts it, we are, or at least should be if brought up well, “naturally disposed to hold that everyone’s good is of legitimate concern to all of us” (2006: 22). Similarly, with James, we might hold that the demands of others are, in themselves, worthy of our ethical consideration; egoism and skepticism are not live philosophical possibilities. Thus far, then, my non-ideal utilitarianism offers a claim about the good—axiological impartiality—and a claim about the right—that happiness is criterial of it. The next step in filling out the theory is to answer two questions: rightness of what, and what suffices for rightness?

59 This natural disposition might be understood as a sentiment rather than an affectless cognition, which would be more in keeping with Jacobson’s and Skorupski’s sentimentalist interpretations of Mill.
3. Practice-Based Theory

The most familiar variety of utilitarianism applies the criterion of rightness directly to actions and claims that each action should, in order to count as right, promote the good. But a better variety of utilitarianism stresses the importance of evaluating rightness more systematically, in terms of rules rather than one-off actions. Richard Miller (2009) offers four reasons to prefer an indirect, rule-based criterion. First, determining right and wrong in terms of rules makes individual rationalization less likely. If agents are given greater flexibility in applying the criterion of rightness to each of their actions, they may be more susceptible to the temptation to tilt the scales of deliberation in their own favor. Second, determining right and wrong in terms of rules rather than one-off actions better promotes social coordination. If agents can reasonably expect that others are following the publicly promulgated rules rather than acting according to their individual judgment, then they can better secure the goods that can be achieved only through cooperation. Third, and relatedly, the mere fact of adhering to a public set of rules, whatever the content of those rules, facilitates the good of belonging to a moral community, of living on good terms with others. Fourth, adhering to a public set of rules secures freedom from blame, since rules are generally less vague in their application than the act-utilitarian criterion. The demand to maximize the good on each occasion is often only hazily fulfilled, leaving individuals uncertain whether they are doing enough good.60 Miller notes finally, and crucially, that all these reasons speak in favor of applying the criterion of rightness to

---

60 Larissa MacFarquhar richly describes the pathologies of one philosophy graduate student who became obsessed with living as maximizing act-utilitarianism demands before allowing herself to deny the doctrine: “It took her a long time to get to this point—several years of guilt and self-laceration—but at the end of it, she no longer believed that she was obliged to dedicate every waking moment to saving the world, or to pry ever more waking moments from her hours of sleep” (2015: 293).
actual rules rather than ideal rules, since these four advantages are all better secured by adherence to actual rules. If we are going to argue for a rule-based criterion over an action-based criterion on the basis of these instrumental advantages, then for the sake of consistency we had better prefer actual rules over ideal rules.

Rather than focus exclusively on actual rules, however, I prefer to apply the criterion of rightness to actual practices. Practices are of course governed by rules, but practices also encompass more than rules: they embody habits of thought, feeling, and action beyond what can be encoded in rules alone. As such, they comprise more of ethical life. I develop the notion of a practice further in Chapter 3; for now, I hope simply to have motivated the shift away from an action-based criterion to a broader one.

The question of whether or not Mill subscribed to an act-utilitarian or a rule-utilitarian view is endlessly vexed, and I will not pursue that exegetical project here. James is more readily recruited to the anti-act-based view of rightness. Immediately after claiming that the best criterion we have is the capacity to bring happiness, he qualifies this in two important ways: “But in order not to break down fatally, this test must be taken to cover innumerable acts and impulses that never aim at happiness” (1956 [1897]: 201). First, then, James refers not only to acts, but to ‘impulses’, a wider notion that gestures at the idea of habits of thought and feeling as well as action. Second, neither acts nor impulses always aim at, or are motivated by, happiness. In order for happiness to continue to offer a useful standard of assessment, therefore, we must ‘zoom out’, as I would put it, from a narrower evaluand to a wider one. The constitutive ethical aim of our

---

61 The currently dominant view holds that Mill was an act-utilitarian: see Crisp (1997). Rule-utilitarian interpretations began with Urmson (1953) and continue today with Miller (2010) and others. Eggleston (2017) provides a useful overview of the debate.
practices can be happiness without its being the case that each particular action falling under the practice aims at happiness. James makes a related point in the epigraph: even if there are certain evaluands—including various emotions and psychological tendencies—that cannot be explained, or justified, by utility, they go with broader categories of evaluand that can be so justified.

The second question to ask is what suffices for rightness. Again, the most familiar variety of utilitarianism is maximizing, which means that rightness of an evaluand (whether an act, a rule, a motive, or something else) is tied to its producing the greatest amount of good possible. But maximizing criteria face notorious objections, especially when evaluated from the perspective of our actual, non-ideal practices. First, maximizing criteria eliminate the possibility of supererogation: going above and beyond what is required. If what is required is to maximize, then as a matter of definition there can be nothing above and beyond that level. Second, and relatedly, maximizing criteria imply that even small divergences from the standard are wrong. But this fails to fit our actual practices, on which wrongness is tied to blameworthiness. As Mill famously puts it in *Utilitarianism*, “We do not call anything wrong, unless we mean to imply that a person ought to be punished in some way or other for doing it” (X: 246). Suppose that giving away 50% of your income to charity is the right thing to do, from a maximizing point of view. That means that giving only 49%—which is still wildly revisionary of our practices—is wrong, and thus blameworthy. But we just do not seem capable of blaming someone who gives away that much of their income.⁶² Although we can inquire about whether a better standard for blameworthiness would be an improvement on our current practices, it is clear that by the lights of our current practices, giving 49% of one’s income away is not only not wrong, but would be

---

⁶² I modify this example from one given in a talk I heard by Christopher Macleod (2016), called “Mill, Morality, and Malleability.”
highly praiseworthy. Third, it is not always clear how to maximize. As mentioned above, there is
greater uncertainty in applying the act-utilitarian standard than in following the moral rules of
one’s own society. And to repeat a point from section 1, we often know how to make something
better without knowing how to make it the best. Rather than having to generate an ideal, we only
need to work out whether a particular proposed revision will be better than the alternatives when
promulgated in our immediate context. Thus, something less than the very best should suffice for
rightness.

The theory that is emerging can still be used to generate a criterion of right action, but it
will be derived from compliance with practices and hence only indirectly utilitarian:

**Criterion of right action**: An act is morally right if it is permitted by a justified
actual practice in one’s society, where a justified practice is one that better
promotes the happiness of those affected by it than the viable alternatives do.

This criterion embodies all three theoretical desiderata: it is non-ideal, in claiming that a justified
practice need not be perfect or end-state but only better than the viable alternatives; it is
teleological, in appealing to happiness to rule out some practices and rule in others; and it is
practice-based, in evaluating actions not directly in terms of happiness-promotion but indirectly
in terms of compliance with happiness-promoting practices. Notice that non-ideal theory is still
theory: it offers an explanatory principle that helps us to tell when actions are right.

My aim in this section has not been to provide a full argument for the theory, but to paint
some of its basic attractions. The next section will move away from abstract theorizing into
discussion of three concrete practices for which, I claim, Non-Ideal Practice Utilitarianism is
superior to maximizing act-utilitarianism. On the latter approach, these practices—or, more
precisely, actions that fall under them—are unjustified, but on the former we can see how they can be justified.

4. Three Case Studies

For each of these cases, I have two aims. The first is to describe and respond to the skeptical challenge from maximizing act-utilitarianism, namely that the phenomenon in question is in itself unjustified by a forward-looking utilitarian standard. Zooming out to a broader practice, of which the phenomenon is a constitutive part—attitudes of regret as partially constitutive of our practices of valuing, the flourishing of harmful opinions as partially constitutive of our practices of toleration, and the punishment of the guilty as partially constitutive of our practices of deterrence and rehabilitation—allows us to see how that phenomenon can be justified indirectly. But becoming aware of this strategy requires individual agents to bracket, that is, not to treat certain otherwise valid considerations as motivating reasons. Bracketing itself raises a puzzle: what justifies me in putting aside, in my deliberation, perfectly valid reasons for action? My second aim is to begin to address this puzzle by answering, for each case, three questions: Which considerations should be bracketed? How difficult is it, psychologically, to bracket? What justifies this bracketing? In section 5, I will address the broader challenge that bracketing is psychologically impossible.

4.1 Wallace on Regret

One essential feature of utilitarian standards, no matter which evaluand they are applied to, is that they are forward-looking: they claim that outcomes or consequences provide better
terms of ethical assessment than backward-looking terms such as moral desert or reparability. As mentioned, the standards could be maximizing, or they could be something weaker: for whatever is being evaluated—actions, rules, motives, characters, attitudes, and so forth—we might ask only whether it ultimately contributes to *more* expected good effects than expected bad ones, and not whether it contributes to the *most* expected good effects. Act-utilitarians, no matter what their stance on maximization, hold that this is a compelling criterion for right action, given that action can affect only present and future states of affairs. But some attitudes, such as regret, are backward-looking, directed at past states of affairs that can no longer be changed. This raises a skeptical challenge. Can regret be justified by the forward-looking criterion? One response claims that since regret is not itself an action, it falls outside our theoretical purview. But this is too quick: since regret can be the primary motive for, or even embodied in, actions—such as expressions of grief and mourning—it is clearly relevant to a forward-looking evaluation.

Let me make the challenge more precise. Regret, as I characterized it in Chapter 1, is a negatively valenced emotion directed at a state of affairs that can no longer be changed. Some of those states of affairs are brought about through our own agency, in which case we feel remorse, a species of regret, but the more general emotion can also be occasioned by circumstances external to our agency. Regret has two features that give rise to the skeptical challenge: it is unpleasant to experience, and it appears to have no constructive point, given that, as Jay Wallace puts it, “it is no longer possible to do anything about the events that gave rise to”

---

63 On this definition, therefore, we can regret future states of affairs, at least if the probability of their realization has attained some sufficiently high threshold. I can appropriately regret the coming catastrophes caused by climate change, for instance.
regret (2013: 30). Thus it cannot be justified directly by the forward-looking criterion; it must
be justified indirectly, if at all.

A response naturally suggests itself: regret ultimately brings about more good effects than
bad ones. Others judge me negatively if I do not seem subject to retrospective sadness over
negative states of affairs: if I appear not to experience any grief at the funeral of a loved one, for
instance, then others will find this disturbing and shun me. So it would be better to experience
regret. But this response will not work. First, it cannot account for regret experienced in private.
Second, it yields only a reason to appear to experience regret, not actually to be regretful. If
experiencing regret were the only way to display regret, then the response might be more
successful, but the possibility of faking regret and other negative backward-looking emotions
means that the response fails. Perhaps the response can be reformulated: in some cases, actually
experiencing regret brings about an overall balance of good effects not simply to others but to
oneself, because it motivates me to do better in the future than I would in the absence of regret.
But this cannot account for regret over impersonal states of affairs, such as the destruction of
valuable objects by natural forces that are beyond anyone’s control. Besides, there are other
emotions that can motivate me to do better that do not entail regret. Indeed, I can admit that I did
wrong, and resolve to improve in the future, without feeling remorse. So the skeptical challenge
—to justify regret in terms of the forward-looking criterion—stands.

The challenge can in fact be strengthened by two further assumptions. First, it is plausible
to suppose that, empirically, regret and its ilk can never fully be given up. Though we can go

---

64 Wallace does not explicitly adopt a forward-looking criterion, but it seems to me that these two features
would not appear problematic from the point of view of a backward-looking criterion on which regret is
simply the appropriate emotion to feel, regardless of its effects.
some way toward weaning ourselves from regret—by improving ethically so that we do fewer regrettable things, or by insulating ourselves from distressing news about the state of the world—it is, like the reactive attitudes, resilient. Second, although it is easy to see how regret is fitting—as a very rough gloss, regret fits its object if that object is in fact negatively valenced and works counter to our aims—that does not mean that regret is apt, or all-things-considered appropriate, where the ‘things’ considered include the prudence, reasonableness, and moral value of a token emotion. Although some emotions are perhaps intrinsically incapable of being fitting, such as superbia, the feeling of inordinate pride at one’s character and accomplishments, other emotions are capable of being fitting without ever being apt, such as schadenfreude, the feeling of pleasure at the misfortune of another. In other words, we need a reason for thinking that regret can be apt as well as fitting. These two additional claims raise the stakes for responding to the skeptical challenge.

The best response is to zoom out to a broader evaluand. Rather than trying to justify token instances of regret directly, we can justify regret in general indirectly, as a constitutive part of something which can be justified directly. Wallace proposes that regret is constitutively connected to our capacities for valuing, a complex set of attitudes and dispositions that entails emotional vulnerability to the things we value. Thus, as he puts it, the “retrospective emotions will have a point so long as valuing itself does” (2013: 25). And not only can we not give up valuing (as a matter of brute probability, this is even less likely than giving up regret, since regret is a constitutive part of valuing), it is clear that our capacity for valuing is directly justifiable.

---

65 Wallace relies on the account of valuing given by Scheffler (2010a). I have argued for a modification to this account in Kubala (2017). Since, on my view, we can value something without judging it valuable, our valuing practices are even more insulated from ethical assessment than many have claimed.
Even setting aside the nature of the particular objects that we value, the stance of valuing is *itself* valuable, in giving rise to forms of experience that make our lives worth living. If we were not valuing creatures, who cared about the fortunes of at least some of the people, projects, relationships, and ideals around us, our lives would be almost immeasurably impoverished. Valuing is therefore justified by the forward-looking criterion directly, and regret and its ilk are justified indirectly as constitutive parts of something which is justified directly.  

I find Wallace’s argument convincing, in its broad outlines, but the zooming-out strategy raises two general questions. First, when are we justified in switching the evaluand? In this case, we began with a desire to find our attitudes justified, and it turned out that seeing regret as constitutively connected to valuing helps us to satisfy that desire. So the justification for zooming out often depends on what our interests are in asking the question. But it is important to emphasize that in doing so, we do not actually change the lower-order negative evaluation. We can still find individual instances of regret problematic in various ways, although not because they fail the forward-looking test directly.

This leads to the second question: what if we would be better off without regret? Shouldn’t we at least consider the possibility of changing? As a matter of practice, the answer is probably no. As C. S. Peirce puts it, “Let us not pretend to doubt in philosophy what we do not doubt in our hearts” (1935: 157). But in theory, we can examine this question by means of what Nick Bostrom and Toby Ord have called the Reversal Test:

---

66 Although Wallace wants to claim that “the value of [valuing] … does not lie solely in its production of beneficial effects” (2013: 31), it seems to me that the forward-looking criterion, and utilitarianism more generally, can acknowledge the value that valuing has in the present; our attitudes of attachment may not always produce beneficial effects, but they themselves instantiate or constitute good effects. To put the point another way, although Wallace thinks that valuing has intrinsic value, the justificatory strategy is successful even if its value is (constitutively) instrumental.
When a proposal to change a certain parameter is thought to have bad overall consequences, consider a change to the same parameter in the opposite direction. If this is also thought to have bad overall consequences, then the onus is on those who reach these conclusions to explain why our position cannot be improved through changes to this parameter. If they are unable to do so, then we have reason to suspect that they suffer from status quo bias (2006: 664-5).

Bostrom and Ord’s example concerns cognitive enhancement. Since most people would agree that a method of safely lowering intelligence would not have good overall consequences, they should be prepared to give a reason why they would resist a method for increasing intelligence. Otherwise, they would be committed to claiming that the net value for society of our current level of intelligence is at a local optimum. Now, there are many reasons we might justifiably resist particular modes of cognitive enhancement, such as transition costs, uncertainty about expected outcomes, deliberation costs, or anticipation that the benefits of enhancement will be unequally distributed. But the burden of proof is on those who resist improvement to offer at least one such valid reason.

Does valuing pass the Reversal Test? In one sense, it is difficult to assess this question, because it is difficult to know what it would mean to enhance our capacities to value. Would it involve heightening our affective capabilities, or the number of items we could fully value? Perhaps we are at a local optimum for valuing. After all, we clearly would not want to diminish our capacities to value. But in different circumstances we might find it reasonable to value fewer things, or to value them to a less intense degree. Consider the apocalyptic scenarios discussed by Samuel Scheffler (2013). In one scenario, we know that an asteroid will destroy the world in
several weeks. In another, inspired by the P. D. James novel *The Children of Men* (1992), mass infertility will lead to the extinction of the human species in some rapidly approaching near future. In these scenarios, perhaps generalized anhedonia would be an apt response, and valuing would lose its rationale in virtue of the loss of a future for humanity. Pondering such scenarios enables us to acknowledge that in some circumstances, which do not currently obtain, we might indeed be better off without our valuing capacities.

Furthermore, thinking in terms of Non-Ideal Practice Utilitarianism can provide pointers for how we might improve our practices of valuing in the here and now. Although the theory does not in itself spit out recommendations automatically—it is “an instrument for rendering deliberation more effective,” in Dewey’s words—deliberating in its terms can suggest action-guiding directions for reform. In particular, although we have seen that there may be little reason to change our practices now, we can identify potential signs of trouble, such as an overly sentimental attitude of nostalgia for a purely imagined past, an indulgent amount of grief as a mode of insulation from confronting a future without the absent beloved, or a paralyzing inertia in the face of coming environmental dangers. These and other examples could trigger an attempt to diminish our regret-reactions on forward-looking grounds, while continuing to respect the beneficial role that some amount of regret plays in our practices.

To sum up, return to the three questions with which I began this section. Which considerations should we bracket? Here we should not treat the facts giving rise to the skeptical challenge as reasons; we should not act on, say, reasons to avoid the emotional vulnerability that makes us susceptible to regret. How difficult is it, psychologically, to bracket? In this case, it is rather easy, since the facts that give rise to the skeptical challenge are, although genuinely
reason-giving, not typically salient in deliberative contexts. What justifies bracketing? Here, it is our desire to make sense of our habits of valuing in a way that provides regret with a constructive point, a desire that we meet by appealing to the value of our actual practices of valuing.

4.2 Lewis on Toleration

David Lewis (1989) considers our practices of toleration. In addition to legal institutions for the protection of free speech, we have private customs of publishing and publicizing the opinions of those we disagree with, as well as the habit of associating with people of different, sometimes extremely different, persuasions on vexed religious, moral, scientific, social, and political questions. And even if we choose not to spend time with such people, we often do nothing to prevent the dissemination of their opinions. Some of those opinions are, by our lights, not simply false but harmful: they would have deleterious effects if they gained even greater traction than they already have. Given the possibility of such potentially awful expected outcomes, how can we justify our acts of toleration by any forward-looking criterion, even a non-maximizing criterion?

The main burden of Lewis’ argument is to show that a forward-looking defense of toleration—one that would show its expected benefits to outweigh its expected costs—cannot succeed so long as it is constrained by a rule of neutralism. Neutralism would require a defense of toleration to proceed without appealing to any premises that are in dispute, only to considerations that are accepted by all parties to a disagreement. An atheist cannot defend toleration of her beliefs, to a group composed partly of religious adherents, by insisting that since there is no God, it cannot be harmful to stand on the street corner crying out words to that effect;
the orthodox would reject the premise that such an action is not harmful. A truly neutralist defense would instead require both sides to agree to a common list of benefits and costs of toleration versus suppression. There are several neutralist points that speak against suppression, such as the risk of error, the possibility that truth will be mixed in with falsity, the loss of rational grounds and meaning that accompany the hardening of a belief into dogma (this point is stressed by Mill), the chance to build character in debate, the importance of individuality, the insult of paternalism, and the obvious harms, to the heretics, of a secret police or prison, which can be recognized by all sides (1989: 155-7). But neutralism, according to Lewis, cannot persuade an Inquisitor, even a utilitarian Inquisitor who discounts appropriately for his own uncertainty, who insists that the neutral list is outweighed by the consequences of heresy. What chance does such a list have against the harms of eternal damnation?

We don’t even have to imagine such an extreme character, in fact, since many of us have beliefs about our own ‘heretics’ that would tip the balance over the neutralist list and decide in favor of suppression. The potential harms of climate change denial, or nuclear hawkishness, which stretch to the decimation or even annihilation of the human species, surely approach, or even outpace, the catastrophic consequences of hellfire for a smaller chunk of the population (and if there are unsaved souls in the coming apocalypses, so much the worse). And even more localized harmful beliefs, such as racism in the police force, can have devastating effects not only on oppressed individuals but in their communities at large. Why are we not justified in taking greater action to expunge these beliefs, beyond our toleration-sanctioned attempts at reasoned persuasion and sensitivity training?
If we wish to find a justification for acts of toleration, therefore, we must look elsewhere than to the neutralist defense. Lewis proposes a treaty of toleration. Suppose, for simplicity’s sake, that there are only two parties to a disagreement, the orthodox and the heretics. For each side, it is reasonable to suppose that at least sometimes, by their own lights, the fear of defeat in a war of suppression will outweigh the hope of victory. If that is so, then each side would prefer toleration to defeat more than it prefers victory to toleration. This generates a preference for conditional toleration—toleration so long as the other side is tolerant—and in any case, each side might still hope to do reasonably well persuading the other under the Ireneic practices of toleration that result. The equilibrium that results can be characterized as a treaty. This would be a “contract for utilitarians” (1989: 166), justified on forward-looking grounds, in terms of each side’s preferences, without appealing, at the most fundamental level of justification, to backward-looking inviolable rights that must be upheld at any cost.

The origin of the treaty in a given society does not matter to its justification: the treaty could be formally codified (as with some peace agreements) or tacitly reached by the drift of history. It could be attained by non-utilitarian reasons, or it could be present all along by ancient custom. Still, the contract itself offers only grudging grounds for toleration, whereas many of us tolerate cheerfully today. And not only do we celebrate our practices of toleration when they are mutually endorsed, we also tolerate the weak, the intolerant, and the extremely dangerous, who cannot be characterized in terms of the original treaty. How did we get from that game-theoretic treaty to our full-throated practices today? Lewis’ answer is that we have developed “a climate of thought” over and above “a constraint of conduct” (1989: 170), one that extends beyond a two-
sided case to a many-sided and shifting alliance. That climate of thought arises out of a commitment to the basic, almost exceptionless, and well-established treaty. As he puts it,

If, in the end, you will always decide that the balance of cost and benefit comes out in favor of complying with the treaty, why should you ever stop to think about the harm done by tolerating a dangerous error? Eventually you will be tolerant by habit, proudly, cheerfully, and without thought of the costs. You will proceed as if the neutralist tally were the whole story about the costs and benefits of suppression. You will bracket whatever you may think about the harm done by others’ opinions. You might still think, in some compartment of your mind, that certain opinions are false and harmful. If the treaty of toleration has become second nature, you will be hard put to explain why these opinions are not dangerous enough to be worth suppressing. But you will never think of the danger as a reason to suppress (1989: 170).

We can unpack this passage in light of the three questions that began this section. What do we bracket? That very fact, that others’ opinions can be dangerous. That fact is a reason: it genuinely counts in favor of suppression. But we must not treat it as a reason for action, not even as a reason that is outweighed. It should not be part of our personal motivational tally. We can still hold, by our own lights, that it would be good if harmful opinions could be suppressed (as Lewis puts it, “the first choice” would be “to suppress and yet be tolerated” (1989: 165)). But that axiological fact is not directly reason-providing for us, living under the treaty, so long as we accept the further axiological fact that conditions are better with the treaty than without.
How difficult is it for us to bracket? In Lewis’ treaty, as with the objective attitude, the bracketing is more often than not done for us, as it were: it has become ‘second nature’ and so we typically do not stop to ponder the harms of toleration, although “in some compartment” we may continue to recognize that “certain opinions are false and harmful.” In other words, the challenge is not that we must avoid thinking about the reasons given by such facts, but rather that we must avoid acting on such reasons.

Why then should we bracket? As Lewis puts it, “bracketing might be not just a consequence of the treaty but part of its very content” (1989: 170). That is to say, the treaty itself may prescribe bracketing, because the practices of toleration are themselves strengthened when we bracket. Lewis is not committed to any particular historical claim, however. He is telling a story that describes how it might have been possible that toleration arose in the past. And, importantly, this story is a device that enables us to see how to construct a second-level justification for a practice that remains otherwise unintelligible on a forward-looking criterion. The ‘neutralist tally’ is not ‘the whole story’, because we can zoom out to a larger practice—one of our actual practices—that does satisfy the forward-looking criterion.

4.3 Rawls on Punishment

Perhaps the most suggestive discussion of bracketing emerges from the seminal work of John Rawls (1955), who distinguishes the justification of a practice from the justification of an action falling under that practice. Although Rawls puts this distinction to work primarily in order to defend utilitarianism from the objection that it cannot explain why we ought to punish the guilty or keep our promises, his recognition that the same person should at different times appeal
to distinct types of moral consideration gives us the clearest understanding both of the
deficiencies of an act-based utilitarian standard and of the need to bracket in certain contexts.

Rawls begins with an attempt to reconcile two competing justifications of punishment. The retributivist holds a backward-looking view, on which punishment is justified only in virtue of the fact that past wrongdoing merits punishment, typically in proportion to the amount of wrongdoing. The utilitarian holds a forward-looking view, on which punishment is justified only in virtue of the fact that it promotes good future outcomes, typically in proportion to the degree that it does so. Both views are plausible, but they are incompatible if applied to the same evaluand. Rawls’ claim is that they can be reconciled by distinguishing two evaluands: we should understand particular actions of punishment as justified by backward-looking considerations and the institution of punishment in general as justified by forward-looking considerations. The best way to promote good future outcomes for everyone, it turns out, is to hold individuals accountable for their past wrongdoing. If we ask why a particular person ought to be punished, we should make reference to their guilt, or the crime they committed. But if we ask why we punish people in general, we should make reference to considerations like deterrence or rehabilitation, which look to future outcomes.

Once we appreciate the distinction between these two levels of justification, Rawls claims, we can see why our institution of punishment has the features that it has. For one thing, it is vitally important that it be an institution that is stable, publicly acknowledged, and designates offices and roles for its implementation. A society without any stable set of practices for punishment—where individuals and crowds can choose how to mete out ‘justice’ at will—is not going to go as well as a society with well-defined practices. Further, the standard utilitarian
considerations are not, by themselves, sufficient to ground a successful institution. Rawls imagines a practice of ‘telishment’, in which officials have the discretion to condemn an innocent person whenever they believe that doing so is in the best interests of their society. But telishment would be highly problematic, insofar as it is unclear how the officials could be checked in the exercise of their power, especially given that the populace will be uncertain as to whether the ‘criminal’ in any case is being punished or ‘telished’. Perhaps most crucially, the motive not to commit crimes will plausibly be weakened, insofar as one could obey the law and still be subject to punishment. Rawls concludes that “as one drops off the defining features of punishment one ends up with an institution whose utilitarian justification is highly doubtful” (1955: 12).

On this view, which Rawls extends to promises and which, as I will argue in the next chapter, can be extended to partiality, instead of evaluating some action of punishment directly in terms of a forward-looking criterion, we consider that action as constituted by a practice that is itself justified by forward-looking considerations. It is this distinction, and not the justification of punishment per se, that primarily interests Rawls. But notice now how respecting this distinction generates a reason to bracket in some contexts. Even though I believe that punishment in general is justified by forward-looking considerations alone (we would not punish the guilty if it had no good consequences), I must not treat this as a reason for action in judging whether particular people merit punishment. Acting as a judge, or simply as a concerned private citizen evaluating a court case in the news, I must deliberate about and act on only the facts relevant to determining a person’s guilt, not the facts about whether punishment in this particular case is likely to deter other crime or rehabilitate the criminal.
Admittedly, there will be some instances in which forward-looking considerations can be directly appealed to in adjudicating particular cases. A judge might determine someone’s guilt but give her a lesser sentence because she wants to set an example for others to cooperate with police investigations. Or a promisor might deliberate on forward-looking grounds and decide not to keep a minor promise he made. But it is crucial to see that these exceptions are, typically, part of the practice itself. That is, forward-looking considerations can sometimes serve as motivating reasons to act, because they are sanctioned by the practice. In many cases, though, the practice itself forbids acting on the basis of forward-looking considerations. As Rawls puts it, “a practice necessarily involves the abdication of full liberty to act on utilitarian and prudential grounds” (1955: 24).

Although appealing to the rightness of following our actual practices can seem to push toward conservatism, this concern can be mitigated by appealing to the notion of a reformer. It is not the case that we always have to act in the ways constituted by our social practices, because we can and should employ a second-level criterion directly in order to improve the first-level practices. Rawls claims: “If one seeks to question these rules, then one’s office undergoes a fundamental change: one then assumes the office of one empowered to change and criticize the rules, or the office of a reformer, and so on” (1955: 28). As mentioned, it is not obvious how best to apply the utilitarian criterion, because among the considerations one has to weigh are the costs of implementation, the possibility of widespread acceptance, and the negative consequences to oneself of pushing for reform, but the criterion can nonetheless provide guidance in deliberating about when and how to reform. One should consider whether a society would be better off with a given practice rather than, say, whether a practice would be what a society deserves.
To summarize, on Rawls’ view we are retributivists about punishment because a practice of punishment is optimific: it promotes the most utility overall. The two levels could, of course, have been flipped: at the first level, we might have had a utilitarian practice of telishment, which is justified at the second level because that is what people deserve. But this is highly implausible, because it cuts against our understanding of desert. People do not deserve to be punished because punishing them will promote overall utility. But the reason for this is that the notion of desert is itself, according to Rawls, constructed by our practices. It is possible that other cultures could organize their practices in this flipped way, but we would find this difficult to understand given our own concept of desert.

Return to my three questions. What do we bracket, for Rawls? As a judge, I must bracket the aims of the larger practice in considering my reasons for action. That is, I must not treat forward-looking reasons as reasons in deciding whether to punish a defendant. Bracketing is especially salient when it seems as though I, as a judge (or rather, as a quasi-judge, since for Rawls it is constitutive of the role of a judge that I not punish the innocent), could get away with punishing someone whom I know to be innocent. That is, in those rare cases where it seems there would be overall positive consequences to punishing an innocent person—nobody would know she is innocent, she herself will not be harmed, etc.—I still have reason to bracket, therefore doing not what I have weightiest reason to do, according to the standard of utility, but what I have undefeated reason to do, since bracketing defeats the weightier reason to punish the innocent. In cases like this, to treat my second question, bracketing becomes much more psychologically difficult, especially since, as discussed above, some forward-looking considerations enter in at the first-order level anyway, and the best way not to act on a reason is
often not to think about that reason. So what justifies bracketing? As with Lewis’ answer in the case of toleration, our practices themselves prescribe bracketing: judges must bracket, though legislators need not. Zooming out to the larger practice—our actual, albeit non-ideal, practice—again allows us to see how the lower-order phenomenon, in this case punishing only the guilty, can be to some extent justified indirectly as part of a practice that meets the forward-looking standard. And we certainly do not have to claim the practice is ideal: the criminal justice system in the contemporary U.S. is deficient in innumerable ways, and reformers should concentrate their attempts on remedying welfare harms to inmates, their families, and the communities that are overwhelmingly, and disproportionately, swept up into that system.

5. Bracketing

In showing how to overcome the skeptical challenge to justify regret, toleration, and punishment in terms of a forward-looking act-based standard, I appealed in each case to bracketing. But one might find something puzzling about the demand to bracket. Even if one believes that we can have reasons not to act on otherwise valid reasons, one might still wonder how it is possible that we deliberate using bracketed reasons. The best way to argue for the psychological possibility of bracketing is, I believe, to demonstrate its actuality. I first offer a handful of examples of ordinary bracketing, then distinguish three options for how best to understand the reasons given by bracketing: as undercutting, overriding, or exclusionary reasons. I suggest that the best way to understand bracketing is in terms of exclusionary reasons, then respond to a recent skeptical challenge to the existence of such reasons and the possibility of deliberating in terms of them.
I start by motivating the notion of bracketing in ordinary contexts, hiving off the particular challenges of thinking in terms of ethical theories. When we bracket, we have a mental state, such as a belief, a desire or a commitment, and we set it aside in thinking about what to do. That is, we refrain from treating it as a motivating reason: a consideration for which we act. Importantly, the belief or desire does not disappear from our motivational set entirely, since we can ‘un-bracket’ it in a different context. The desire or commitment still represents something we want; the belief still represents something we hold to be true. Since the project of ethical theorizing concerns our beliefs, my examples will also discuss beliefs rather than desires or commitments. Consider an example in which we have to bracket our true beliefs in order to act successfully:

*Scientifically Literate Addict:* An addict knows that addiction is disease-like, in that there are biological determinants of his condition. But he also knows that studies have shown that people who believe their addiction is a disease have less ability to resist it.\(^\text{67}\)

Suppose the addict is deliberating about whether to sober up, or, having decided he should, about continuing to do so. This process takes place over a longer timespan, with complex motivational peaks and valleys. At some moments the addict is likely to think about the fact that his condition is indeed disease-like. But he also has a reason, given by his aim of sobriety, not to treat that fact as a reason to despair or even to abandon his goal, even though the fact bears on the likelihood

\(^{67}\) An interesting discussion of this phenomenon can be found in an interview with the writer J. D. Vance (2016), who says: “It’s this awful catch-22, where recognizing the true nature of the problem actually hinders the ability to overcome.”
that he will be successful. It might even be worthwhile to try to forget entirely that addiction is a
disease, or at least to refrain from reading or hearing about that fact.

Further examples are not hard to come by. Lawyers regularly ask jury members to set
aside certain facts in their deliberations, and deliberators in a democratic body are frequently
asked to bracket their own commitments to reason in a more neutral framework about the public
good. These considerations still bear on the case at hand—the DNA evidence, although
contaminated, still bears on the question of who committed the crime; the religious commitment
still bears on the question of the permissibility of the state’s policy—but they must not be acted
upon, for procedural or other reasons. Simply having to set aside a true belief would not give rise
to philosophical puzzlement, however, since the vast majority of my true beliefs do not bear on
any particular deliberative situation. What is puzzling is that it seems we have a reason, given by
some other end or aim, not to treat the bracketed true belief as a motivating reason, even though
what is bracketed is still a valid reason that speaks in favor of some option we are considering.68

The first challenge is to analyze that puzzlement and to understand more clearly what
bracketing is supposed to be. We can begin by distinguishing three ways in which reasons can be
affected negatively by other considerations. First, reasons can be undercut by other
considerations. If I promise to meet you for drinks, then I have a reason (on some views, even an
obligation) to show up at the agreed-upon time. But if you begin to express reservations about

---

68 Compare here scientific cases, in which investigators are deliberating about which conclusions their
evidence supports. No one thinks that, if the aim is truth, any inquirer ever has a reason to set aside
perfectly valid evidence; Galileo had no epistemic reason to bracket his observations of sunspots (though
he surely had some prudential reason to publicly proclaim that they were simply undiscovered planets
crossing in front of the sun). There may be cases in which one wants to find a more direct inferential route
to one’s conclusion and thus brackets some evidence; in many of these cases, however, the goal is to
persuade someone of a truth one already believes oneself (though not all cases, e.g., mathematicians
trying to find a simpler proof). The puzzle, then, is why bracketing can be legitimate in normative cases
as opposed to scientific cases.
your ability to make it, or your desire to drink, then, plausibly, your scruples undercut—reduce
the weight of—my original reason. At the limit, you might release me from the promise entirely,
in which case my original reason is disabled. Second, reasons can be outweighed by other
reasons. If I pass a drowning child on the way to the cocktail bar, then I have weightier reason to
stop and rescue him. I still have a reason to meet you; its force hasn’t been diminished or
canceled. But I have even more reason to help the child. Third, reasons can be excluded by other
reasons. The notion of an exclusionary reason was first formulated by Joseph Raz (1999). To
borrow one of his examples, suppose I am deliberating about where to send my daughter to
school. There’s a school which is conveniently located near my home and whose principal is a
friend of mine. But there’s another school which has slightly better educational outcomes and is
correspondingly more expensive. I have promised my partner that I will make the decision only
on the basis of what is best for our daughter’s education, and not on the basis of my own
interest. Such a promise, according to Raz, gives me a negative second-order, or exclusionary,
reason: a reason to do something not on the basis of some valid reasons. In particular, I have a
reason not to make my decision on the basis of self-interested considerations.

Clearly, we cannot identify bracketing with disabling reasons, because disabling reasons
affect the weight of other first-order reasons, whereas bracketing requires us to put aside valid
reasons whose weight is not affected. So bracketing must be identified with either outweighing

---

69 Dancy (2004) has a thorough analysis of the ways in which the weight of reasons can be affected by
facts that do not themselves count for or against the actions in question.

70 Supposing, for the sake of argument, that my daughter’s interests and my own diverge.

71 Each of these ways of affecting reasons has a positive counterpart: reasons can be intensified by other
reasons (making them more weighty), reasons can outweigh other reasons (without changing the weight
of the original reasons), and we can have positive second-order reasons (reasons to act on the basis of
other valid reasons). I focus on the negative sort because my interest is in bracketing.
or exclusionary reasons, since neither changes the force of our first-order reasons. The key difference between the two is that when deliberating in terms of outweighing reasons, I simply weigh the balance of all the reasons I have, whereas exclusionary reasons do not compete with first-order reasons; they tell us not to act on the basis of first-order reasons, and (if valid) they always win out. In particular, exclusionary reasons can sometimes tell us that we should not act on the balance of reasons: they can change the verdict about what we ought to do.\textsuperscript{72} Suppose that I have weightier reason to choose the first school for my daughter: it is not \textit{that} much worse, pedagogically, than the second, and we can give the money we’ve saved to charity or spend it on a vacation, and it will be more convenient and advantageous for me. If my promise to my partner generates an exclusionary reason, however, then what I ought to do is to choose the second, more expensive school, on the basis of my daughter’s educational interests, on the basis of the promise I made to my partner. Exclusionary reasons, as Raz admits, require us to give up the general principle that we ought to do what we have most reason to do, at least if ‘most reason’ is understood in terms of the weight of first-order reasons alone. Instead, we have to replace it with a related, but distinct, principle of practical reasoning: “It is always the case that one ought, all things considered, to act for an undefeated reason” (1999: 40), where a reason can be defeated either by being outweighed at the first order or excluded at the second order.

It is crucial to emphasize that exclusionary reasons are not reasons to perform some distinct mental act, such as not thinking about other first-order reasons. They are reasons to act on the basis of other reasons. Perhaps in some cases I do have additional reasons to try not to

\textsuperscript{72} Edmundson (1993: 333) emphasizes that there must be at least some cases in which the excluded reasons would tip the balance of first-order reasons, were they not excluded, if there is to be any practical difference between exclusionary and outweighing reasons.
think of the excluded reasons, because thinking about them makes it more difficult for me not to act on them. But in general, merely thinking about the excluded reasons is not ruled out: in the promise case, it is not that I promise not to think about the reasons given by my own interests (a difficult thing to guarantee, after all). I can think about them all I like, so long as I do not actually act on the considerations they represent. As Raz puts it, “So long as one knows that one’s reflections will not affect one’s action, the ill effects of such thoughts are avoided” (1999: 184). Perhaps Raz puts the point too strongly; the opacity of our motivational states suggests that it is difficult to know whether we have acted on one or another reason. But even if the distinction is messy in practice, it is clear, in theory, that what is required by exclusionary reasons is that we not choose to act on the basis of the excluded reasons. The following list compares the three kinds of reasons:

<table>
<thead>
<tr>
<th>Disabling</th>
<th>Outweighing</th>
<th>Exclusionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change first-order reasons</td>
<td>Do not change first-order</td>
<td>Do not change first-order</td>
</tr>
<tr>
<td>Weighed in the balance</td>
<td>Weighed in the balance</td>
<td>Not weighed</td>
</tr>
</tbody>
</table>

While I speak of reasons as entities, talk of reasons should be understood as shorthand for talk of processes of reasoning. And if this is so, then we can use the key difference between outweighing and exclusionary reasons as the basis of a test for whether or not some case of deliberation is best captured as one or the other. The test is whether we are, in deliberation, weighing up all the
considerations in deciding what to do. If we are not, then we are acting on second-order reasons; if we are, then we cannot be acting on second-order reasons. Consider Raz’s example of Jeremy, a soldier who is ordered by his commanding officer to appropriate a civilian’s van. This order, backed by the officer’s authority, generates an exclusionary reason for Jeremy not to act on the basis of his own deliberations. He must bracket his own assessment of the case. Suppose now that Jeremy actually believed, prior to the order, that he has weightier reason not to appropriate the van. Raz admits it is possible that someone could regard the order as an outweighing first-order reason but stipulates that this is not how Jeremy reasons. For Jeremy, it is not the case that “yet another factor is added to the balance of reasons”; the order is not “a conclusive reason for appropriating the van” (1999: 42). Rather, Jeremy reasons that it is not for him to act on his own assessment of the case.

One complication that Raz does not discuss is that we can weigh whether or not to respond to an exclusionary reason. Jeremy could deliberate about whether or not to obey the order; he might suspect that his commanding officer is mentally unbalanced. But it is important to distinguish this deliberation about the authority of the officer from deliberation about the content of the order. The former deliberation might be a matter of weighing reasons only, but the latter, if Jeremy obeys the order, is a matter of not acting on his own assessment of the first-order reasons for and against appropriating the van. So long as we can distinguish these two kinds of

---

73 Chang (2016: 220-7) is especially emphatic on this point, although she would deny the usefulness of my test because she finds the notion of an exclusionary reason mistaken. Chang wants to frame purported cases of exclusionary reasons in terms of deliberation about which choice situation to be in, deliberation that is a matter of weighing considerations alone. This just moves the lump under the rug, however, since exclusionary reasons could surely be just as relevant to the question of which choice situation to be in!
deliberation, then we can apply the test to the latter and determine that it is better captured in terms of exclusionary reasons.

An additional complication in applying the test is that we may not ultimately be certain which reasons we acted on. Our motivations are notoriously labile and opaque, and we cannot always know what really motivated us. This is why the best way to bracket—to act not on the basis of certain reasons—is often to avoid thinking about those excluded reasons. If we have made it the case that we have not thought about those reasons, then we will be more confident that we have not acted on them. But the fact of our uncertainty is not sufficient to establish that ultimately we only have first-order reasons to try not to act on the basis of excluded reasons. Perhaps an order such as that of Jeremy’s commanding officer generates such reasons, but it also generates second-order reasons for him actually not to choose based on those reasons. Understanding bracketing in terms of exclusionary reasons gives us a vivid sense of the different orders of justification at play.

So which type of reasons, outweighing or exclusionary, should we employ to characterize my examples? Scientifically Literate Addict is plausibly described in terms of exclusionary reasons: the addict’s desire to overcome his addiction gives him an exclusionary reason not to avoid seeking recovery on the basis of his belief that his addiction is a disease outside of his control. But an objector might claim that it can also be analyzed in terms of competing reasons, if the belief that addiction is a disease is outweighed, in deciding whether to seek treatment, by the desire to get clean. So consider the three practices discussed above. In the case of regret, it seems most plausible that, in our deliberations about the things we value, we are not weighing up all the considerations relevant to our affective state. Regret passes the test for exclusionary
reasons, since it is not as though the sadness of regret is outweighed by the potential joys of valuing. Rather, we typically just experience the various dispositions and attitudes associated with valuing, and insofar as we deliberate at all, our deliberation concerns the objects we value, not our attitudes toward them. In the case of toleration, our practices clearly generate a reason not to act on the potential harm of others’ opinions; we simply must not weigh that harm in our deliberations about how to act. And punishment is perhaps the clearest case, given the separation of the roles of judge and legislator: the judge must not weigh considerations of overall utility, but must bracket them and deliberate about guilt alone, except insofar as the practice permits otherwise.

The vividness of the notion of an exclusionary reason—we have a reason not to act on the basis of certain considerations, even though those considerations retain independent weight—makes it attractive in thinking about bracketing. But the mere fact that we can make use of a concept to characterize some instances of reasoning is insufficient to establish the concept’s coherence. Daniel Whiting (2017) has recently issued a serious challenge to the existence of second-order reasons in general and exclusionary reasons in particular. Although his challenge explicitly concerns the existence of such reasons, as entities, it seems to me that the interesting question concerns whether or not we can deliberate and act in the ways described. In fact, his first two objections concern this more interesting question, charging that deliberation in terms of exclusionary reasons is impossible.\footnote{Whiting’s third objection is that second-order reasons are the ‘wrong kind of reason’. Reasons of the wrong kind are those that bear not on whether some object has some evaluative property (e.g., is desirable, credible, etc.) but on whether some attitude toward that object has that evaluative property. But although there may be theoretical reasons to deny that such entities exist, a wrong kind of reason may nonetheless be a good reason for action (Jacobson 2013). If a demon threatens to torture you if you do not desire to eat mud, that may not count in favor of eating mud ‘in the right way’, but it does provide an “exogeneous incentive” to eat the mud. As such, this challenge is orthogonal to my present purposes.}
1. Whiting’s first objection is that second-order reasons cannot *guide* us. One plausible constraint on reasons—one which I endorse—is that if you have a reason to \( \varphi \), it must be possible for you to \( \varphi \) on the basis of that reason.\(^{75}\) So if you have an exclusionary reason not to \( \varphi \) on the basis of some first-order reason, it must be possible to \( \varphi \) on the basis of that first-order reason on the basis of that second-order reason. But Whiting claims this is not possible: you cannot do something on a certain basis on a certain basis (2017: 404).

This claim is difficult to assess, because Whiting’s own examples do not adequately support his case. Notice at the outset that it is *prima facie* plausible to offer descriptions of motivation that fit the supposedly impossible form: I choose to go to rehab on the basis of my desire to get clean on the basis of bracketing my belief that addiction is a disease. It seems eminently possible to do something for a reason for a reason or, as in this example, for a reason *not* for a reason. But Whiting offers two examples to claim that this description of motivation is not coherent.

First, he considers the following description: Kelly goes to the pub for the reason that she promised to do so for the reason that she received a decent upbringing (2017: 404). This seems to fit the form that defenders of exclusionary reasons employ, but Whiting points out that ‘that she received a decent upbringing’ is not the kind of reason that could *guide* one; I cannot act for the motivating reason that I received a decent upbringing.\(^{76}\) ‘That she received a decent upbringing’


\(^{76}\) Plausibly, I could act on the basis of the *thought* that I received a decent upbringing, if, for instance, I imaginatively project a certain ideal of myself—as a well-raised, upstanding individual—and am motivated by that. But I cannot be motivated to act on the basis of receiving a decent upbringing alone.
is neither a motivating reason (a reason for which I could act) nor a justifying reason (a reason that counts in favor of going to the pub). But this objection misses the mark, because no one should think that the reason in question is exclusionary. That I received a decent upbringing is not a reason to be motivated or not motivated on the basis of any particular set of reasons at all; it is simply underspecified as a reason that could count in favor of anything.

    Second, Whiting attempts to extend this line of objection to the school case. It seems as though I can choose the more expensive school for my daughter on the basis of educational reasons on the basis of my promise to my partner. Whiting claims instead that my promise is better seen as an explanatory reason: a reason why I made my decision on educational grounds, not a reason for which I did so (2017: 405). But this is not so. For one thing, it seems to be a general fact about explanatory and motivating reasons that whenever I offer a motivating reason in terms of an aim my action furthers, I am thereby citing an explanatory reason. Whiting’s strategy could be extended to show that the fact that I made the decision on educational grounds is itself a reason why I chose the school I did. Indeed, I think this is true; that fact helps to explain why I made the decision I did. But it does nothing to exclude the possibility that the explanatory reason is also a motivating reason. So Whiting’s strategy proves too much. For another thing, when I imagine myself deliberating about this case, the promise seems to me to count in favor of making my decision on educational grounds; it doesn’t seem like an explanatory reason. I can think about the self-interested grounds all I like, so long as I do not choose the school on their basis. So the promise seems to give me a reason for choosing (what turns out to be) the more expensive school on the basis of educational considerations alone.

77 Skow (2016: 142) defends the claim that teleological answers to why-questions (i.e., answers in terms of the aims of action) report reasons why (i.e., explanatory reasons).
2. Whiting recognizes that his first objection rests on “an appeal to (so-called) intuition” (2017: 404), and I have tried to argue that his intuitions can be explained away by thinking more carefully about the cases in question. The next objection, however, he takes to be an argument proper, namely that we cannot receive credit for acting on second-order reasons. More specifically, he claims that “it cannot both be true that a person deserves credit if she does the right thing for the right reason and that it is possible to respond to second-order reasons” (2017: 407). Remaining with the school example, Whiting tries to argue that one cannot receive credit for acting on a second-order reason alone. He considers a version of the case in which the reasons in favor of the pedagogically superior school are weightier, and in which I would have chosen the other school if not for my promise to my partner. In fact, I am concerned about my daughter’s education only insofar as I am concerned with keeping my promise. In such a case, according to Whiting, I cannot receive credit for responding to second-order reasons, because I am not really acting for the sake of my daughter’s education.

There are several problems with Whiting’s version of this case. For one thing, here the exclusionary reasons lead me to choose the option that had greater first-order weight anyway. As Raz himself recognizes (1999: 41), this makes it difficult to assess whether exclusionary reasons are the best way to describe the case. For another thing, it may be that my lack of concern for my daughter is what is, distractingly, preventing me from receiving credit. In general, keeping a (valid) promise makes one creditworthy, and Whiting obscures this fact in describing the case as
one where I lack concern for my daughter.\textsuperscript{78} It seems to me that I deserve some credit here at least for keeping the promise.

Even if this is acknowledged, perhaps I will be pressed to concede that one always receives less credit for acting on second-order reasons than on first-order reasons. But this is not so. Suppose, to construct a better version of the example, that the first-order reasons in favor of one or the other school are tied; that is, there are equally weighty reasons to choose each school. Or, to paraphrase the ‘reasons as entities’ talk, I am torn between the two schools; each seems to me equally choiceworthy. But my partner really wants me to choose one school, although I do not know on what basis he wants this. Suppose that I have already factored this into my first-order deliberations: I think that my partner’s desire to choose the one school counts in favor of that school, and it is this fact that makes the two schools equally choiceworthy. That is, if my partner did not want this school so badly, I would have more overall reason to choose the other school. But suppose that now my partner extracts from me a promise that I will make the decision based only on educational grounds, and therefore I exclude other reasons and wind up choosing the school he prefers. It seems to me that in this case I deserve credit for keeping my promise, thereby satisfying my partner’s desires, and certainly not less credit than if I had simply flipped a coin or decided in some other way which of the (to me) equally choiceworthy schools to choose. Crucially, my partner’s desire for one school is not, for me, a first-order reason: it does not itself count in favor of choosing that school. But on the basis of my promise, I have an

\textsuperscript{78} Whiting does insist that we might be able to fill out the case such that I am blameless for my lack of concern for my daughter, and that I would still not be creditworthy for making my decision on the basis of my promise. But (a) I have a very difficult time imagining circumstances in which one is blameless for a lack of concern for one’s (young) child and (b) in any case, the defender of second-order reasons could continue to insist that this case would be better analyzed in terms of outweighing reasons and would therefore fall outside the scope of Whiting’s challenge.
exclusionary reason not to consider other kinds of grounds. So we can, contra Whiting’s challenge, receive credit for acting on second-order reasons.

At this point, I hope to have provided further defense of employing the notion of an exclusionary reason to characterize bracketing, and to have responded to a challenge that deliberation in terms of exclusionary reasons is impossible. To close, I attempt to draw together the strands of these first two chapters.

6. Conclusion

What unifies the phenomena discussed in section 4? In each case, we employ a plausible evaluative standard but find that the standard cannot justify some evaluand directly. In order to find the evaluand justified, we instead analyze it as a constitutive part of a distinct evaluand which can be justified directly. Regret, which appears unjustified from a forward-looking perspective, is analyzed as a constitutive component of valuing, which does appear justified. Toleration, which seems nonsensical when viewed in terms of the standard of preserving our own conceptions of truth and value, turns out to make sense when seen in the light of a treaty of toleration. Retributive notions of guilt and sentencing, which look confused when we consider that we cannot change the past, begin to seem justified when we see that a practice of punishing only the guilty makes things go better overall.

There is a second commonality to these phenomena, however, which is that the initial confusion arises from attempting to justify an act or attitude in isolation, considered apart from the psychological and social practices that render it intelligible. Rather than looking at the narrower evaluand alone, we can become more confident in its justification when we see it as
part of a broader one. As I suggested in section 3, our actions are often best evaluated not in
themselves, but as instances of habits or larger patterns.

A third commonality is the need for bracketing. As per the ethical psychology developed
in Chapter 1, most of the time, in ethical life, things go according to habit, and while our habits
are to a great extent repositories of practical intelligence, they are often suboptimal, due to path-
dependence and other contingencies. As we saw with Hare, it is a fantasy to imagine that we
chose our intuitive dispositions and reactive attitudes, and it is highly doubtful that we would
have chosen this set if we were designing optimal agents. Take the love of drunkenness, or
bashfulness, as James urges us in the epigraph: these phenomena cannot be “wholly explained,”
or justified, in terms of utility alone. But they go with other things that can be so explained. The
love of drunkenness often goes with a laudable conviviality; bashfulness often accompanies a
virtuous modesty. So we can sanction these phenomena, at least up to a point. And when we need
to change them, we can appeal to that other standard: to utility, or to some other second-level
criterion.

Still, we cannot always act on that second-level standard, for a variety of reasons. And
when that is so, we need to bracket. We sometimes have reason not to act on the reasons given by
the second-level standard, even when those reasons have greater weight. From the point of view
of the utility of actions, we sometimes have weightiest reason to punish the innocent, or to
suppress our opponents’ points of view, or not to regret the past. But those reasons can be
defeated by the practices that govern them and the exclusionary reasons that such practices
generate.
Return to Strawson’s ethical psychology. The objective attitude is a welcome resource. But our reactive attitudes, when understood in Strawsonian and not Harean terms, generate an exclusionary reason not to adopt the objective attitude on every occasion. If, in my dealings with some especially frustrating person, I find that viewing them from the perspective of the objective attitude is actively impeding our long-term prospects of continuing in a meaningful relationship —my first-level aim—then I have reason not to treat the objective attitude as reason-giving, even if doing so would make things easier for us both in various ways. Ethical theories, I claim, are similar. If ethical thinking comprises both levels, and not the second level alone—that is, if both levels are genuinely reason-giving—then we can sometimes have ethical reason not to treat the second-level standard as reason-giving.

Having motivated my favored theory, Non-Ideal Practice Utilitarianism, I will fill it out in greater detail in Chapter 3 by extending it to one of the most vexed issues in moral philosophy: our practices of partiality.
Chapter Three: Partiality as a Non-Ideal Practice

The art proposes to itself an end to be attained, defines the end, and hands it over to the science. The science receives it, considers it as a phenomenon or effect to be studied, and having investigated its causes and conditions, sends it back to art with a theorem of the combination of circumstances by which it could be produced. Art then examines these combinations of circumstances, and according as any of them are or are not in human power, pronounces the end attainable or not. The only one of the premises, therefore, which Art supplies, is the original major premise, which asserts that the attainment of the given end is desirable.

— John Stuart Mill, *System of Logic*

In this chapter, I extend the Non-Ideal Practice Utilitarianism (henceforth NPU) developed in the previous chapters to our practices of partiality. Partiality refers to the preferential treatment, in action and attitude, that we give to those with whom we share what are called *special* relationships, as opposed to the generic relationship we bear to fellow members of the moral community.

The standard view of partiality treats partiality as a puzzle, given liberal Enlightenment morality’s commitment to *impartiality*, the idea that all the individuals who matter morally matter in the same way (Cottingham 1986; Kolodny 2010a; Scheffler 1982; Wolf 1982). Different ethical theories vary in which properties of individuals they take to be morally fundamental, but most agree with the core claim that no one matters morally more than anyone else. But if this is so, then what explains our giving certain people—those with whom we stand in special relationships—preferential treatment? As Simon Keller puts it in his book on partiality,
“If morality is all about promoting overall enjoyment and minimizing overall suffering, or treating all autonomous creatures as ends, or respecting everybody’s rights, then special relationships would not appear to hold any moral significance. … Why should the fact that someone happens to share some relationship with you make a difference to how you ought to treat her?” (2013: 5). The standard view takes impartiality as an ideal end-state standard for assessing our practices, and treats partiality as a puzzle accordingly. How do my objections to ideal theory from Chapter 2 apply to the standard view?

The primary motivation for going non-ideal was that we want theorizing to guide our action. One very basic objection is that it would be impossible to give up our partial motivations, and thus impossible to be guided by an ideal of impartial treatment (Goodin 1985). Sociologically, it seems that norms of limited altruism were crucial historically and are crucial today to the success of societies. Psychologically, it seems that we naturally empathize most with those closest to us, and that parental partiality is necessary for healthy child development: a child needs to perceive itself as especially valuable to somebody. And phenomenologically, it seems that our attachments to particular people are often the source of greatest meaning in life. I do not want to argue that ideal ethical theories cannot justify some partiality on impartialist grounds, but I do want to recall, from Chapter 1, that perhaps the most common objection to them stems from their alleged inability to guide action. This is the objection that acting on impartial motivations is self-defeating: if you visit your friend in the hospital just because you think that will promote the most aggregate good, then you are not really behaving as a good friend would (Stocker 1976).

Even if we thought that impartial treatment of others was the ideal, however, the fallacy of approximation implies that we might be unjustified in trying to directly close the gap between
our current and ideal motivations. And worries about false negatives recur here, too: why do we have to see partiality *per se* as a *problem*? Non-ideal theory proposes instead to consider what is good about partiality *now*, and what function it serves. And it is evident that extreme impartialism, on which *no* form of partiality is morally permissible, is neither endorsed by the vast majority of actual ethical practices nor reasonable or feasible to implement. So rather than begin with an impartialist criterion and attempt to justify partiality in those terms, I want to assume that our practices of partiality are to some degree reason-giving and then to apply my theoretical criterion to explain what is good about partiality, justify our current practices to some degree, and offer guidance for improvement. This will follow the model developed in Chapter 2 for regret, toleration, and punishment.

This chapter has two parts. In section 1, I return to an abstract account of two-level ethical theory, arguing against what I call the partialist constraint and against a constitutivist view of normative explanation. The partialist constraint, common in recent philosophy, states that in investigating what makes right acts right we must grant a great deal of authority to our experience of the reasons we act on. I argue instead that given the gaps between motivating and justifying reasons, the only match we need is extensional adequacy, which does not require that motivating and justifying reasons be identical in content. A constitutivist view of normative explanation states that the second level of ethical thinking offers an explanation of what it *is* for something to be a reason at the first level, without appealing to further reasons. I argue instead that normative explanation itself ought to be reason-giving.

In section 2, I develop my own account of partiality in terms of the facilitation of partial goods. Inspired by Mill’s procedure, quoted above from the *System of Logic*, I first interpret our
current practices to determine which goods are most valued in partial relationships, then try to identify what kind of value those goods have. Only then do I apply the second-level theoretical criterion, in order to investigate how such goods can be better distributed. The non-ideal criterion gives us new reasons to work toward improving our practices.

1. Two-Level Ethical Theory Reprised

In Chapter 1, I offered four examples of ordinary situations that seem to involve a split between motivating and justifying reasons. Characterizing these situations accurately, I suggested, requires an account of moral psychology in which agents can think in two different ways about ethics. At the first level are our ordinary moral thoughts: the properties of actions and attitudes that make them seem choiceworthy or appropriate, the reasons that we take ourselves to act on, and the kinds of considerations we might typically offer as justifications. At the second level are the more fundamental ethical considerations from ethical theory, fundamental in the sense that they explain why the first level is reason-giving. In Chapter 1, I argued that adopting a two-level picture of ethical thinking is the best way to respond to the deliberative role problem. But developing an adequate moral psychology is not the only motivation for a two-level picture of ethical thinking. There are at least three other reasons to go two-level.

First, one might be moved by curiosity about what makes right acts right. Why do we have a duty to keep our promises? Why should I try to be more generous? One way to hear these questions is as requests for explanation, such that even someone who is committed to promise-keeping and generosity might want to know more about why. This curiosity could be ethically motivated; for instance, I might believe that becoming more consistent will lead me to behave
more fairly toward others, and that I will be more consistent in my actions if I can develop a more systematic account of what makes them right. But curiosity can also be solely theoretical. Although it might be regrettable for someone to study ethics without becoming, or even desiring to become, a morally better person, such an outcome is surely possible. After all, one could develop a two-level theory to describe the ethical practices of some alien culture—giving an account of what its agents ought to believe about the right-makers of their acts—without endorsing its practices oneself. This quest for explanation does not commit one to moral realism, however, since the investigative procedure would be the same whether the second level was thought to be discovered or constructed. In either case, the only materials to hand would be our moral intuitions, our reflective judgments, various salient empirical facts, and our capacity to reason about all these things. I tried to remain neutral on such metaethical topics in previous chapters, and I intend to maintain that neutrality here.

Second, one might be drawn to a two-level picture of ethical thinking in order to legitimate both everyday morality and ethical theory. As Daniel Star (2015) has persuasively argued, a two-level picture can make sense of two plausible but apparently inconsistent claims: (i) people can be ethically virtuous without the aid of philosophy, and (ii) normative ethical theorizing, understood as the search for very general ethical principles, is not a waste of time. Behaving ethically might require knowledge only of the first level, but investigating the second level would nonetheless improve our understanding of the subject matter of ethics.

Third, a two-level theory seems better able to do justice to what we might think of as morality’s Janus face, which looks toward both individuals and social practices. The first level of ethics concerns how individual agents ought to think in order to act morally, holding fixed the
wider social milieu in which they live. But the second level provides the resources to evaluate
the broader social practices that make sense of our individual actions. As Simon Keller puts it in
discussing partiality, “To understand what you should do for someone with whom you share a
special relationship, you need to appreciate his value in isolation, but you also need to see him as
a participant in a relationship that has a certain social significance and function” (2013: ix).
These two views, as Keller’s quotation suggests, should both be treated as reason-giving, i.e., as
contributing to understanding what one has overall reason to do.

These three motivations for adopting a two-level theory can be brought together in the
following way: investigating our social practices can help to explain what makes right acts right,
which can in turn legitimate both ordinary ethical thinking (to some degree) and ethical
theorizing (in a certain fashion). In section 2, I investigate our social practices of partiality in
particular. In this current section, I focus on the explanatory function of two-level theory in
general.

Given our ordinary ethical thinking, including our understanding of the considerations
that count in favor of our actions and attitudes, we want to know why our thinking has the force
it does, and why those considerations do genuinely count in favor. On the first, ordinary level,
the fact that some action has been promised, or will help our friend, or will prevent causing pain
to a sentient creature, can all seem to be reasons in favor of performing that action; they can even
be the reasons on which we act, if we are sufficiently reflective before performing them. The
second level promises to tell us why these are reasons: for instance, because promise-keeping is
required by the Formula of Universal Law, or because beneficence is a virtue, or because pain is
intrinsically bad and always gives us a reason to avoid it (and so forth). Thinking at this second
level has the potential to loop back and put pressure on our first-level practices, at least to the extent that it implies that our other practices cannot consistently be willed as universal laws, or fail to demonstrate the virtues, or produce a net balance of pain over pleasure. This picture is illustrated in this simple diagram:

1: Ordinary moral thinking

explain

2: Fundamental right-makers

This understanding of two-level theory in terms of its explanatory role should be distinguished from others. First, I do not understand two-level theory in terms of the distinction between conditions of imperfect compliance and full (ideal) compliance. Christine Korsgaard (1986a), for instance, has argued for a Kantian two-level theory with ideal reasons at the second level, which set a standard beyond which we can never permissibly act, and non-ideal reasons at the first level, which tell us what we may do when we accept the responsibility of violating ideal standards. But I think of the second level as presenting not an ideal end-state standard, but rather an explanatory standard that can in turn be used to ameliorate our actual, non-ideal conditions. Second, as I argued at length in Chapter 1, I do not understand two-level theory in terms of providing guidance for how to implement a worked-out ethical theory. Hare, for instance,

---

79 As Star (2015: 17 fn. 16) points out, ‘right-maker’ is a term of art that can be used to refer either to one side of an explanatory relation or to one side of a metaphysical relation, such as the grounding or constitution relation. Here I understand right-makers in the former way, as parts of explanations that cite the properties and relations in virtue of which our actions and attitudes are justified.
understands the first level in terms of intuitive principles, derived from the second level, that approximate second-level principles under real-world conditions. For Hare, the existence of the first level is justified only in terms of its ability to help us to realize the demands of the second level (Feldman 2012; Forschler 2013). But I deny that the first level has value only derivatively of the second level. As Alex Worsnip (2016) points out, morality may be an instance of a general phenomenon in which ordinary people are largely attuned to fine-grained features of situations but very bad at articulating the general principles that govern their discriminations. If so, two-level theory would be a good way to model this.80

So much for a reprise of the two-level picture. In the remainder of this section, I first argue that there is no reason to treat our ordinary moral thinking as a constraint that, in itself, can rule out certain normative explanations. I then argue in favor of a particular kind of normative explanation, one that grounds our ordinary reasons in further reasons that we have.

1.1 Against the Partialist Constraint

Much of the recent literature on partiality contains an appeal to a methodological constraint of the following form: in investigating what makes right acts right, we should grant a great deal of authority to our experience of the reasons we act on.81 If it seems to Cordelia that what justifies her love for her father is the relationship that they share, and if Cordelia is not an

80 See also Kleiman-Weiner et al. (2017), who note that we learn abstract knowledge from observational data but often do not understand and cannot articulate how. Other examples might be the principles of grammar and the rules that govern perceptual recognition.

81 Often very little is said about the nature of this experience, but see Horgan and Timmons (2005: 60-1) for the claim that the phenomenal experience of moral judgment consists in (a) a felt demand evoked by (b) an experientially-present sense of the basis of that demand, namely its fittingness to certain features of the environment. The idea is that all direct moral judgments consist in part in a sense of some and only some features (the basis of the demand) calling out for some and only some responses (the demand).
isolated case, then parent-child relationships should be thought, *prima facie*, to be the fundamental right-makers of actions that fall within the compass of filial relationships. This constraint can come in a stronger or weaker form, depending on whether it is held that our experience of the reasons we act on must be the guide to fundamental right-makers, or simply one guide that we cannot discount. But in either form, this constraint is frequently wielded both (a) as an objection to any posited right-maker that does not show up in moral experience, notably the forward-looking claim that ‘making things go best’ is the fundamental right-maker that explains our reasons, and (b) as part of an argument for some favored account of what the right-makers are. Thus, Jay Wallace claims that the best argument for non-reductionism about the reasons provided by relationships—the idea that relationships themselves have fundamental moral significance—is that reductionism “amounts to a denial of the normative appearances” (2012: 185) and that the normative appearances reveal that relationships are in themselves reason-providing. Simon Keller similarly claims that it would be “both implausible and depressing to suggest that when we act well within special relationships we systematically misperceive our reasons” (2013: 27), and that the best argument for his own account, on which reasons of partiality are provided by individuals (and not relationships), is that it preserves the “phenomenology of partiality.” And Nick Zangwill claims that any ethical theory that posits a different right-maker from what we experience as the reasons we act on must be “committed to a massive error theory about ordinary moral thought” (2011: 164), and that we should therefore adopt what he calls a “common-sense morality” about reasons of partiality.

In Chapter 1, I argued that a Constructive view of ethical theorizing can meet the objection, raised by these writers, that two-level theories are committed to treating the first level
of ethical thinking as illusory. Here, I want to argue that the objections themselves are misguided insofar as they only get off the ground by endorsing the partialist constraint. Rather than assuming a two-level theory and showing how it can overcome the objection (as I did in Chapter 1), here I want to argue against the partialist constraint on independent grounds, based on more general considerations about motivating and justifying reasons.

One plausible response to Bernard Williams’ path-breaking article on “Persons, Character, and Morality” is to note that Williams fails to avail himself of this distinction between motivating and justifying reasons. Williams borrows from Charles Fried the thought experiment of a man forced to choose between rescuing a stranger and rescuing his wife. Imagining that the man opts for the latter route, Williams notes that for standard ethical theories, the consideration that ‘it was his wife’, although surely “an explanation which should silence comment,” would be thought insufficient for justification, in the sense of appealing to a moral principle that can “legitimate his preference” (1981a: 18). One such legitimating principle would be that ‘in situations of this kind it is morally permissible to save one’s wife’. But if that principle were to serve as the man’s ‘motivating thought’, then it would provide him with ‘one thought too many’. According to Williams, it ought to be motivationally sufficient for the man to have thought simply, ‘that’s my wife’.

As I said, a plausible response is to judge that Williams has not appreciated the force of the distinction he seems to assume. Motivating reasons are the reasons on which we act, and justifying reasons are those that, in Williams’ word, legitimate action. But once we have drawn this distinction, we need not imagine that the two should always coincide. In fact, we should expect that the two should become increasingly separate. Williams seems to think of motivating
reasons as the thoughts that immediately precede and causally initiate action. The motive for one’s action, which on this picture is always a conscious motive, is the consideration that springs to mind immediately before acting and that plays a causal role in bringing that action about. But notice now that even the thought that ‘that’s my wife’ might be one thought too many! Why should the man, faced with the choice of rescue, have to place this woman under the ‘wife’ concept in order to be motivated to save her? It might have been hoped (for instance, by his wife) that he simply has to recognize her in order to be motivated to save her, all thoughts of their relationship be damned. Or perhaps, given a different psychology of motivation, there is nothing thought-like at all that need pass through the man’s head before he acts: his love for his wife is so ingrained that he moves to rescue her with the force of a brute instinct.

But just as motivating thoughts can become increasingly spare, so can justifying thoughts become increasingly baroque. After all, the legitimating principle to which Williams appeals—“in situations of this kind it is at least all right (morally permissible) to save one’s wife” (1981a: 18)—might itself stand in need of further defense. Why is saving one’s wife morally permissible? What if there were two people, or five, or a hundred, who could be saved instead of her? Now we are back in the realm of ethical theorizing, and justifying thoughts, “fully spelled out,” might be wildly complex indeed. Thinking about the gap that can open between motivation and justification, I claim, puts pressure on the defender of the methodological constraint to explain why we should grant so much authority to our experience of the reasons we act on.

At this point, the defender of the partialist constraint might note an ambiguity in the phrase ‘the reasons we act on’. My argument, it is conceded, has force, but only if the reasons we
act on are understood exclusively as the motivating reasons we act on. But the phrase might also mean the justifying reasons we act on, such that our experience of those reasons would, even if it never showed up in our motivating reasons—because we acted entirely by habit, say, or because we were so fully virtuous that we never needed to deliberate even when confronting novel situations—need to be given some authority in investigating what makes right acts right. After all, as Marcia Baron puts it, “one’s action can be governed by a belief that one is acting fairly without its being the case that one thinks about fairness at the time of action” (1984: 212). In other words, I might have beliefs about justifying reasons of fairness without ever being motivated—in Williams’ sense of immediate, conscious, causally-effective motivation—by thoughts of fairness. And our investigation of what makes right acts right should, according to the objection, grant a great deal of authority to our experience of justifying reasons in this broader sense.

My response begins by conceding that there is a thin sense in which I accept the partialist constraint for our experience of our justifying reasons. This is the sense in which the posited right-makers must accord with what Brad Hooker calls “attractive general beliefs about morality” (2000: 4). The fundamental right-makers that we posit have to make sense to us; they have to have the right kind of connection with our self-understanding of what we are doing when we act. And so explanations such as ‘this action is right because it was chosen by a roll of the die from a predetermined option space’ or ‘because it is the will of our alien overlords’ will be ruled out by the thin sense of the constraint that I endorse. Those right-makers just aren’t plausible as explanations: we would not be willing to acknowledge them as right-makers. It is difficult to give a precise account of what is ruled out by this thin sense of the constraint, but the candidates
offered by standard ethical theories are certainly ruled in. This is why I find it odd that Zangwill (2011: 150) claims that the right-makers posited by consequentialism are not ‘intuitive’. It is precisely the fact that ‘because it makes things go best for everyone’ sounds so intuitive that makes consequentialism compelling for so many.\(^\text{82}\)

That being said, there remains another sense in which I do deny the partialist constraint. This is the understanding of the constraint in terms of the justifying reasons that we take ourselves to have when we act. These are the reasons that we would, if asked, cite as justifying our actions and attitudes. But I claim that even these reasons need not be expected to coincide with the fundamental right-makers. I can best illustrate this by reference to an analogous position, articulated by Susan Wolf in a recent exchange with Nomy Arpaly, about a certain class of non-moral reasons. Wolf proposes that, in addition to reasons of morality and reasons of self-interest, we also have reasons of meaning, which reflect our motivation to find and attach ourselves to fulfilling projects, relationships, and ideals. Such reasons, according to Wolf, cannot be captured by the domains of morality and self-interest alone. Arpaly worries, however, about one possible development of this line of thought: “Those who act, say, for the love of art, do not act for the sake of a meaningful life but rather for the sake of art. … imagine a person who thinks, ‘I am going to help my wife because she is my wife and love for my wife is among the things that make my life meaningful.’ That, too, is one thought too many” (2010: 90). Arpaly’s objection seems to be that ‘meaning’ cannot pick out a third category of reasons, because people do not act for the sake of meaning, whereas they do act for the sake of morality and the sake of self-interest.

\(^{82}\text{I am grateful to Michele Moody-Adams for discussion of this point. Philippa Foot (1985) refers in this context to the “spellbinding force” of consequentialism.}\)
Yet Arpaly, like Williams, fails to appreciate the distinction between motivating and justifying reasons. First, we need not deny that someone who had the motivating thought Arpaly imagines would indeed be “inappropriately agent-centered,” as she puts it. What is odd about the thought is that we do not typically conceive of the choiceworthiness of our projects, relationships, and ideals in terms of their being things that make our lives meaningful. Second, however, as with Williams’ legitimating principle, even this thought might not be taken as sufficient justification by a suitably hectoring interlocutor, and in any case it would certainly not be required in ordinary conversational contexts, where the consideration that it is one’s wife would be “an explanation which should silence comment.” Yet surely the thought Arpaly imagines could be part of a good explanation for why one ought to engage in such projects.

Wolf makes similar points in her response to Arpaly. She first discusses reasons of self-interest. When we do ordinary things such as preparing a meal or going for a walk, we are often motivated (at the first level of practical thinking, as I would put it) simply by the thought that these things will be fun or relaxing: “One may never class these acts under the more general and abstract heading of ‘self-interest’. But generally, one would be willing to acknowledge one’s reasons of pleasure as a species of reasons of self-interest” (2010: 117). I read this as passage as expressing a constraint similar to the thin constraint of sense-making I articulated above. Even if some agents never think in abstract terms about moral right-makers or the meaning of their lives, they would be willing to acknowledge such descriptions as intelligible and applicable to them. Wolf then extends this thought to reasons of meaning:

Insofar as there are such things as “reasons of meaning”—that is, insofar as facts to the effect that something will contribute to the meaning of a person’s life give
reasons to foster or promote that thing, those facts will rarely be directly available to the people whose lives are in question and are not likely to matter much, even when they are (2010: 118).

What kind of reasons could have genuine justificatory force and yet be unavailable to most people and not likely to matter much even where available? The answer is fundamental reasons, i.e., the kinds of theoretical considerations that figure at the second level of ethical thinking. These considerations are, often, not likely to matter much to the practices of those who reflect on them, when they are thought of at all. People are overwhelmingly more likely to act on reasons drawn from the first level, reasons that incorporate considerations about the things—the projects, relationships, and ideals—that they recognize as valuable. This is all perfectly in keeping with my Strawsonian ethical psychology. But at the same time, someone who wondered what it was in virtue of which all those things might be justified could reasonably think that part of the answer was their meaningfulness, even if a fuller answer would also make reference to the domains of morality and self-interest.

Yet these second-level reasons should not always be motivationally idle. They have the potential to loop back and change our first-order practices, as Wolf herself admits. Recognizing that meaningfulness is a dimension of a good life, distinct from happiness, “will make one want to ensure that all children are exposed to such things, and that our social, political, and economic institutions provide opportunities for all people to form relations and pursue interests that bring meaning to their lives as well as pleasure and comfort” (2010: 119). To some degree, “we already want these things,” Wolf notes—these motivations already belong to our practices—but the force of theorizing is to draw out and make explicit these previously unrecognized commitments in our
everyday practices, which can exert pressure to extend opportunities to others to participate in those practices.

In section 2, I will sketch a similar normative explanation for partiality, with a parallel structure to Wolf’s account of reasons of meaning. To summarize this subsection: the lesson I draw from my argument against the partialist constraint is that the explanation for why our reasons have the force they do need not coincide with the reasons we take ourselves to act on. Once we have made that separation, the normative explanation can be as egoistic or as impartialist as we like, so long as it is intelligible, applicable to us, and, of course, has good arguments on its behalf. We can apply this lesson to the debate about partial relationships. The fact that you are my spouse, or the fact that we stand in a certain relationship, or the fact that you are $X$, might all represent the content of my motivating reason or the justifying reason that I would cite if asked to account for my actions. Those are the considerations on which I act, and they might well silence comment in everyday conversation. But when, as ethical theorists, we ask the question of why those considerations provide reasons, we are offering an explanation that might appeal to distinct considerations: partial relationships promote my personal well-being, promote the well-being of our relatives, promote impartial well-being, constitute one way of legislating the Kingdom of Ends, serve our normative interests, exemplify the life of virtue, and so on and so forth. There is no reason to expect that the content of these reasons should coincide, given the vagaries of ordinary motivation. In fact, the only match we need between our ordinary reasons and the normative explanation of those reasons is extensional adequacy: the normative explanation should explain all the good reasons we take ourselves to have, although not
necessarily only the good reasons we take ourselves to have, since the normative explanation could itself, as I will argue next, reveal further reasons.

1.2 Against Constitutivism about Normative Explanation

The second level of ethical thinking is the level of normative explanation, which I have characterized in terms of right-makers. But there are two distinct views one could hold here, only one of which is compatible with my view that ethical theorizing can provide us with further reasons. First, one could offer a non-normative, constitutive explanation of what it is for something to be a reason at the first level, an explanation that does not appeal to further reasons that one has. Second, one could offer a normative explanation in terms of more fundamental reasons we have. In this subsection, I will argue that we should look for an explanation of the second kind, and in section 2, I will provide such an explanation for partiality.

The first way to think of the second level of ethical theory would be as providing an explanation of why reasons at the first level are reasons, an explanation that does not itself appeal to further reasons. This would amount to a kind of metaphysical explanation, one that tells us what it is for something to be a reason, or in virtue of what a particular consideration counts in favor. For instance, one might offer a constitutive consequentialist account on which reasons just are considerations which, when acted on, most promote the overall good. At the first level, the fact that some action aids another person does provide a reason to perform that act. But at the second level, the explanation of why that fact is a reason—because it contributes to the amount

---

83 I am grateful to Kieran Setiya for helping me to distinguish these two views of normative explanation.
of good in the world—is not itself reason-providing. The fact that some option promotes overall good is, on this view, *no reason in favor of it*.

There are three objections one could offer to such an account. One minor objection is that it incurs weightier ontological commitments than a non-constitutivist account, which can remain neutral on ontological questions. A more significant objection is that it renders our second-level explanation motivationally inert. Rather than pointing to further reasons on which one could act, thus offering the potential to change our practices, the constitutivist account is purely explanatory. I follow Star (2015: 30), however, in thinking that the account is problematic on its own terms. Star considers a constitutivist interpretation of rule-consequentialism, on which a rule directing us to keep our promises is explained in terms of the fact that it belongs to a set of rules that together satisfy the rule-consequentialist conditions. He imagines that we are speaking to a reliable guardian angel, who (i) tells us that rule-consequentialism is the best ethical theory, (ii) tells us that only one of the acts about which we are currently deliberating is permitted by a rule endorsed by rule-consequentialism, but (iii) does not tell us what that rule is. The key to Star’s objection is (ii). On the constitutivist account, we have *no reason* to perform the act endorsed by the rule; the information given to us in (ii) is, by the constitutivist’s lights, simply an account of what it is to be a reason, and not *itself* a reason. That is, for the rule-consequentialist constitutivist, to belong to the set of rules that jointly satisfy rule-consequentialism just *is* what it is to be a reason, and is not itself a reason (a consideration that counts in favor of something). But this is the wrong result. If we could possess such reliable moral information from a guardian angel, we would have a reason that counted in favor of performing the act, even if we did not know the content of the rule that permitted the action we had reason to perform. I take this
implication to be the most powerful argument against the constitutivist view of normative explanation.

The view of normative explanation that I endorse—call it the normativist view—claims that the fundamental considerations offered by ethical theory are reason-giving, however indirectly. If an action is what the virtuous person would do in the circumstances, or if it passes the Categorical Imperative procedure, or if it would make things go better overall, then that *is* a reason in favor of it. The constitutivist view of normative explanation denies this. And the constitutivist view does not sit well with the claim that criteria of right action express our fundamental commitments, since our commitments must, in order to count as commitments, be reason-giving for us. Of course, the fact that a consideration is reason-giving does not mean that we *always* have reason to act on it. As Chapter 2 discussed, sometimes the reason is overridden by another one, or sometimes we have an exclusionary reason not to act on a reason.

There are many accounts of our fundamental reasons that are consistent with the denial of constitutivism about normative explanation: egoist, eudaimonist, welfarist, perfectionist. My own preferred account, as per Chapter 2, is Non-Ideal Practice Utilitarianism. Having explained more fully the kind of two-level theory I want to offer, I now want to develop such a theory for partiality.

2. Justifying Partiality

In Chapter 2, I relied on Wallace, Lewis, and Rawls for their accounts of the justification of regret, toleration, and punishment, respectively. Here I offer my own account of partiality, which will proceed in three steps. First, I start with what we already value in partiality. What do
people want from their partial relationships (§2.1)? Then, I try to identify the distinctive aim, or aims, of partiality. What does this practice facilitate that other practices do not (§2.2)? Finally, I deploy impartiality as a non-ideal ideal—not a final, end-state standard that we must comply with, on pain of moral wrongdoing, but a hypothesis to guide proposed reforms (§2.3). In the end, we should be able to explain in more detail how relationship goods give us reasons to pursue them, and how a second-level standard gives us reasons to improve our practices.

The procedure I recommend can be described even more telegraphically: identify the goods (the Utilitarianism component) produced in our practices (the Practice component) as they currently are (the Non-Ideal component). Since the bulk of this section will consist in an elaboration of the three steps of the justificatory procedure for partiality, I first want to motivate this procedure by describing each of these components in more detail.

Goods: Deontologists hold that the right-making features of an act “include such things as whether the act involves keeping a promise, telling the truth, returning a favor, helping a friend, protecting the innocent,” and so forth (Hooker 2000: 106). These are the sorts of features that we cite, at the first level of ethical thinking, as counting in favor of our actions, and they are often the kinds of considerations that motivate us. Further, many deontologists hold that these right-making features are fundamental: there is no principle that explains and justifies these and only these features. But as Chapter 2 discussed, these features are not fundamental: their value derives from their tendency to make things better.

This component inspires the goods-based focus of my account, namely that the fundamental justification of partiality as a practice must appeal to the goods that relationships
facilitate, to their participants and to others. This is importantly distinct from the standard approach to partiality in the recent philosophical literature, which is to ask which relationship types are justified. Niko Kolodny (2010a), for instance, assumes that the question of when we have reasons of partiality is equivalent to the question of which relationship types justify partiality. On my view, however, the question of which relationship goods are worth producing is more fundamental than the question of which relationships justify partiality.

All parties to the debate agree that there is a challenge to determine when we have reasons of partiality. My claim is that the challenge cannot be met at the level of relationship types. This is because, in normative explanation, partial goods are more fundamental than relationship types. In other words, the value of a relationship type is best explained by the value of the goods it facilitates or instantiates, and not the other way around. Four considerations speak in favor of my view. First, there is an open-question argument in the vicinity. It is evident that to call a relationship a friendship or a sibling relationship is to say nothing about whether it is good, absent an account of its good-making features. Second, relationship goods are more fine-grained than relationship types, because any relationship is likely to be the site of multiple goods, such that an overall normative judgment can be rendered positive by the presence of multiple goods or rendered ambiguous by the presence of some negative features. Third, and because of the previous two points, explaining the value of a relationship type will not necessarily explain the
value of a relationship token.\textsuperscript{84} Fourth, the nature of a relationship type changes over time. Five centuries ago, a paradigmatic marriage was not as likely to be the site of mutual respect and equality that we take marriage to be today. This point can be supported by the stronger claim that relationship types do not exist independently of the social practices that individuate and regulate them, social practices that are not timeless or essential.\textsuperscript{85} All four considerations illustrate that the explanatory buck can and should be passed from relationship types to relationship goods.

I should emphasize from the outset that I am understanding relationship ‘goods’ in an expansive sense, to pick out all good-making features and not just discrete quantities such as money and other transferable resources. The list of relationship goods is potentially quite lengthy and includes goods of enjoyment, respect, attachment, vulnerability, trust, virtue, and so forth. In §§2.1-2.2, I will offer an interpretation of our practices that provides some structure to the list of goods.

To identify a good is not yet to license an inference to how we ought to act. In general, of course, one has reason to promote goods.\textsuperscript{86} But the denial of standard maximizing act-

\begin{itemize}
\item Kolodny occasionally seems to recognize this point, but it does not affect the way he frames the challenge: “it is far from obvious that there is the same reason for partiality within every relationship that is, in some recognizable use of the terms, a relationship between ‘parent’ and ‘child’, sibling’ and ‘sibling’, and so on. The challenge ‘Why do we have reason for partiality in some relationships, but not in others?’ seems to recur within kinds of relationship that, loosely understood, call for partiality” (2010b: 55). Seidman rightly points out—and I would agree—that a better answer to the question of when we have reasons of partiality would tell us not which \textit{relationships} are reasons of partiality, but which \textit{considerations} are reasons of partiality (2013: 124).
\item As Raz puts it, “for friendship to be possible it must be predefined. It is predefined by social practices which establish and mark out patterns of interaction between people. The ability of people to have a particular relationship depends on its being established by social practices known to them, and which they share, at least to some degree. People nowadays cannot establish the relations between master and apprentice which existed in mediaeval society” (1989: 19-20).
\item Maguire (2017: 687) calls this claim \textit{VALUE SUFFICIENCY}: “If some state of affairs S is valuable, then the fact that some option φ would promote S is a reason for you to φ.” This is consistent with there being overwhelmingly weightier reasons against φ.
\end{itemize}
utilitarianism implies that there is no entailment from the claim that a state of affairs is valuable to the claim that one has most reason to bring that state of affairs about. Furthermore, just as, in identifying various goods that relationships provide, we do not yet say anything about how to act or feel in any particular situation, neither, in identifying various reasons that we have, do we yet say anything about the weight of those reasons, which would be required to render a final deontic verdict on how to act or feel. But the inquiry into the goods we value in relationships is still relevant to real-world ethical deliberation, because it identifies the kinds of considerations that we ought to be weighing; we are identifying the raw materials for deliberation, as it were.

*Practices:* I take a practice to be a shared form of activity constituted by norms that define roles, moves, penalties, and standards of right and wrong behavior. A practice of baseball defines the role of the batter, the pitcher, and the umpire, and standards of correctness for how they play. An economic practice defines the roles of buyer and seller, and sets standards that govern what counts as a transaction versus a robbery. A practice of voting defines the roles of voter and observer, and distinguishes between a valid and an invalid vote, or a free choice and a coerced choice. Practices are something more than recurring patterns of individual action: they are constituted by the norms that govern such actions and the roles that relate the people who occupy them. So why take partiality to be a practice? First, partiality clearly has structured roles: becoming a friend or a mother or a citizen all carry with them evaluative standards for how to behave in that role and how to relate to those who play complementary roles. Furthermore, these roles are both variable and to some degree optional. They are variable because, as mentioned above, different social groups organize these roles in distinct ways: relationship types do not
exist independently of the contingent social circumstances that regulate them. Role play all you like, you and your friend simply cannot become a medieval master and apprentice; that social form is no longer available to us today. And these roles are optional because people deliberate about whether or not to take them on. Consider marriage or parenthood. It’s noteworthy that many people deliberate whether they wish to inherit the baggage of these practices. Do they really want to be married in a culture where marriage is so politically fraught. Do they really want to raise children in a social group where parenting takes on outsize importance? When people do opt in, however, and they do so consciously, it is typically because they expect to be able to facilitate certain goods for themselves and others. So while it may be true that partiality has less structure and fewer institutions than other canonical practices, such as punishment, this is no mark against understanding partiality as a practice in my sense.  

As They Are: Understanding partiality as a non-ideal practice means starting with our current habits of behaving partially to one another. Once we have a sense of which goods are actually valued (§2.1), we can try to identify what kind of distinctive value those goods have (§2.2). These steps have to go in order; it is not as though we have a list of our antecedent values, specifiable in advance, and then work out practices that best realize those values. As Elijah

---

87 My definition is inspired by Rawls, for whom a practice is “any form of activity specified by a system of rules which defines offices, roles, moves, penalties, defenses, and so on, and which gives the activity its structure” (1955: 3 n. 1). A distinct definition of a practice, which would also make room for partiality, comes from MacIntyre: “a coherent and complex form of socially established cooperative human activity through which goods internal to that form of activity are realized in the course of trying to achieve those standards of excellence which are appropriate to, and practically definitive of, the form of activity, with the result that human powers to achieve excellence, and human conceptions of the ends and goods involved, are extended” (2007: 187). This definition, like Rawls’, emphasizes the cooperative and normative elements of practices, but, unlike Rawls’, adds an emphasis on the excellence involved in achievement internal to the practice and conceptions of the good external to the practice. Although I agree with MacIntyre that self-understanding is one good of practices, I follow Rawls by not making self-understanding partially definitive of a practice.
Millgram puts it, “typically, it is only after one has a great deal of experience with a … practice that one is in a position to do a reasonable job of picking out organizing goals for it” (2011: 190 n. 32). And it bears emphasizing that identifying the organizing goals or values of a practice is a matter of interpretation. We do not deduce the goals of a practice, but infer them from observing how people act within the practice. What are we aiming at when act partially toward each other in various relationships? The key move, as my epigraph from Mill’s Art of Life describes, is to assume that whatever our aims turn out to be, those ends are valuable.\(^8\) This generates a rational explanation of a social practice, in which some of the reasons that agents have to act anyway are reasons for them to act in ways that conform to and support the practice. Of course, there is no expectation that each agent acting within the practice has formulated such a rational explanation in order to act well within the practice. As Lydia Goehr writes, “The way practice works is not necessarily identical to how people think it works in their day to day activities. Persons tend to think globally only when encouraged or forced to do so” (1992: 104). The interpretation of how practice works, at this global level, is a job for theory.

It is not the only job for theory, however. Once we have identified the organizing goals of a practice, we are in a position to evaluate it (§2.3). In particular, this is where we can apply the theoretical criterion of rightness given in Chapter 2: An act is morally right if it is permitted by a justified actual practice in one’s society, where a justified practice is one that better promotes the happiness of those affected by it than the viable alternatives do. We want to know not only which

\(^8\) As a recent trio of commentators puts it, “For such [practices], ascertaining their respective axiological first principles is a fairly straightforward matter of interpretation rather than a matter of affirming a value judgment. For each such art, the end at which it aims can be identified by interpreting the actions understood to be guided by the art in question, and then its axiological first principle can be formulated as the proposition that that end is desirable” (Eggleston, D. Miller, and Weinstein 2011: 5).
acts are permitted but whether there is feasibly a better distribution of the goods of partiality that could increase the happiness of those affected by the practice. If, as I will argue, the organizing goal of partiality is to provide certain goods, then we can see how it falls short where those goods are poorly distributed. So the theoretical standard can give us a reason to change our practices where possible; it is not normatively idle. As with Wolf’s reasons of meaning, the widespread distribution of the goods of partiality is already an element within our practices, but the force of non-ideal theorizing is to serve as a diagnostic tool to make explicit this previously unrecognized commitment and suggest hypothetical solutions for improvement. My hope in pursuing the three-step justificatory procedure is not only to show what NPU has to say about partiality, but to offer further illustration of how two-level ethical theorizing can be reason-giving for us.

2.1 Valuing and Relationships

Begin, then, with the goods that human beings actually value in their partial relationships. I understand relationships broadly as relations individuated by the identities of their participants and the history of encounter between them.\(^89\) An exhaustive list of what we seek and benefit from in relationships would perhaps be impossible to provide, but we can make some progress by distinguishing two kinds of goods. *Special* goods can by provided only by a particular person with whom we share a functionally irreplaceable relationship, while *generic* goods could be

\(^{89}\) I take relationships between persons as my paradigm, but the account can in principle extend to cover relationships between human beings and animals, artworks, the natural world, and so on.
provided by others, in principle if not in practice. There are difficulties in spelling out the distinction precisely, but we need only the basic idea.

Any token relationship may be the site of goods of both kinds. For instance, a (suitably reflective) child might value her relationship to her parents for the educational opportunities and material comforts they provide her, as well as for their emotional support and approval. The former are generic goods: if her parents genuinely could not afford tuition and someone else had to pay, the child should not be upset at the need for substitution. But the latter are special goods: what the child desires could not be provided by any other adult, and if her parents were cold and withholding it would be appropriate for her to feel bereft even if she were smothered with affection by other family members and friends. It is a significant fact about human valuing that many of the goods we most desire and pursue are special goods: the love of a particular individual, the approval of a certain friend, and the way that just one person makes us feel can all take on outsize importance in our practical deliberation. And when we are lucky enough to receive these special goods, they will often be among the most highly valued elements of a relationship. Our relationships to others give our lives meaning: they have the power to make our lives go well or poorly, and as a result we very much hope that they go well.

Believing that relationships are the source of these goods does not mean that we value our relationships only instrumentally. Following Rae Langton (2007), I distinguish two basic

---

90 I borrow the terms ‘generic’ and ‘special’ from Simon Keller (2013: 106-11), although my usage is different from his. Keller defines a good as special “if a full description of its nature implies that there is only one person, or only a few people, by whom it could be provided,” but this account produces false positives. I may want a recommendation for membership in a certain club, which has to come from a current member. Although there are only a few people who could provide it, the recommendation should be classed as a generic good, even though the recommendation could only be provided by my sponsor, because the relationship of sponsorship is replaceable. (I might, however, also have an irreplaceable friendship, or other relationship, with the person who serves as my sponsor. We can bear multiple relationships to the same individuals.)
ways in which we value things: intrinsically (for their own sake) and extrinsically (for the sake of something else). Valuing something instrumentally (for the sake of its effects) is an instance of extrinsic valuing, because one way to value something for the sake of something else is to value it for the sake of its effects. It is not obvious how we typically value relationships, in part because the psychology of intrinsic valuing is so obscure. But it seems fairly clear that we do not value relationships only instrumentally.\[91\] We value something only instrumentally, I claim, when we would abandon it without loss of value if we could acquire the good to which it is a means in another way. To borrow an example of Samuel Scheffler’s, suppose I value my membership in the American Association of Retired People for the sake of the discounts on prescription drugs that membership provides (2010b: 50). I would value my membership only instrumentally if, could I get the discounts in some other way, I would do so without seeing it as a loss. Yet if I experienced loss at the lapse of my membership, that would suggest that I value it in non-instrumental ways, such as for the sake of the camaraderie it provides me. The breakdown of important relationships often produces a sense of loss, even if we believe that they end for good reasons (more on this in Chapter 4).

Even if we deny that we value our relationships only instrumentally, however, I claim that we do value our relationships in part instrumentally, i.e., for the sake of the goods they provide. To many, this will sound problematic, especially if it is taken to imply that we value relationships, as, in Scheffler’s words, “a means to some independently specified end” (1997: 196). Two points can be made in response. First, although relationships can be thought of as

\[91\] Railton has a nice discussion of this point: “Friends are not ‘things that make one achieve friendship’—they partially constitute friendships, just as particular happy experience[s] partially constitute happiness for an individual. Thus taking friendship as an intrinsic value does not entail viewing particular friendships instrumentally” (1984: 149-50 fn. 21).
means to ends, in the case of special goods relationships can only be *constitutive*, and not merely productive, means. That is, because the relationship is irreplaceable, it would be the only way of attaining the good in question, and so might well be valued in its own right as a constitutive part of the end.\(^92\) Second, and in any case, reflection on the psychology of valuing suggests that relationships are not typically thought of as ‘independently specified’ goods or ends that we pursue. Although people often set out in search of friends or romantic partners, what they are usually looking for is someone to play that role generically, not someone who answers to a fully-specified description. And even if we were to embark on a relationship seeking an independently specified end, it is often the case that successfully maintained relationships come to seem valuable in themselves.\(^93\)

The picture of valuing given by John Dewey lends further support to the claim that insofar as we value relationships at all, we value them in part instrumentally. Dewey holds that ends and means lie along a continuum. By this, he means two things. First, we do not evaluate ‘ends-in-themselves’ independently of means, but (at least in part) in terms of whether or not we can—whether in a practical, prudential, moral, or some other sense of ‘can’—take the means to attain them.\(^94\) Second, and more relevantly for present purposes, the ends we seek, the desires we have, and the consequences we aim at are all “valued in turn as means of future consequences” (1988 [1939]: 229). For Dewey, the means-ends distinction is only clear-cut at a

\(^{92}\) A more precise definition comes from Setiya: “A productive means is an efficient cause [of an end]; a constitutive means is one that is an instance of, or part of, the relevant end” (2007: 99).

\(^{93}\) Consider the description offered by Proust in his Preface to Ruskin’s *Bible d’Amiens*: “A man becomes acquainted with a woman because she may help him reach a goal other than herself. Then once he knows her, he loves her for herself and sacrifices to her without hesitation that goal which it was merely her function to help him attain” (1987: 58).

\(^{94}\) This implies that, for Dewey, there can be deliberation ‘about’ ends. We decide which ends are choiceworthy in part on the basis of whether they are feasibly realized through the means available.
single timeslice, which is where most philosophers focus. But over time, what was once valued as an end can become valued as a means. I might aim to talk to you at a party in order to invite you to collaborate on a business venture, but later come to value the end of collaboration in turn as a means to befriend you. My conception of my ends shifts as I respond to new experiences, thereby shifting what counts as a means. In sum: there is no reason to think that conceiving of our valuing of relationships in terms of the goods they provide implies that we do, or should, value relationships only instrumentally, even if there are good reasons to think that we often do value our relationships in part instrumentally.

2.2 The Aim of Partiality

Having briefly surveyed which goods are most valued in various partial relationships, the second step is to identify what is distinctive about partiality. I argue that relationships facilitate distinct special goods, and not only distinct forms of generic goods. If this is so, then facilitating special goods is the distinctive aim of partiality as a practice. And then, following Mill, we can assume that facilitating such special goods is itself valuable.

Thomas Hurka (2013) distinguishes three different kinds of value that close personal relationships may have. First, relationships may be a place where generic goods are realized to a greater degree than usual. Second, relationships may be the site for distinct forms of generic goods, with additional value that flows from their location. Third, relationships may instantiate distinctive goods not found in any form elsewhere. Hurka does not make a claim about which, if any, of these values are facilitated by relationships, although he notes that the three kinds of value are not in conflict; relationships could be the site of all three. I want to argue that
relationships can in fact be the site of all three kinds of value, although I will understand these values somewhat differently than Hurka does.

Return to my (incomplete) list of relationship goods: enjoyment, respect, attachment, vulnerability, trust, virtue, being known. It is certainly true that relationships are a place where some generic goods are realized to a greater degree than usual. The time I spend with my friends, for instance, is often much more enjoyable than time spent alone, or with strangers. Being a good parent demands many exercises of virtue, and additionally allows me to pass on virtues to my children. Marriage and romantic partnerships require enormous amounts of trust and vulnerability. And all of these relationships may involve the distribution of material goods such as gifts and money. But because these are generic goods, they could be acquired in other forms. I could respect, trust, be emotionally vulnerable, and give money to someone with whom I do not share a relationship, just as a stranger could respect, trust, be emotionally vulnerable, or give money to me (as in Dickens’ *Great Expectations*).

Relationships can also be the site of distinct forms of generic goods. Hurka thinks that these generic goods attain an additional, distinctive value in virtue of the shared history that we enjoy with our relatives. Just as my furniture or clothing would have additional value, beyond their fine intrinsic qualities, in virtue of being treasured hand-me-downs from beloved ancestors, so do our relationships attain a distinctive value, beyond the intrinsic qualities of their participants, in virtue of the past experiences that have been shared.\footnote{The furniture example is Hurka’s (2013: 206-7). Kolodny (2003) holds, implausibly in my view, that relationships (and not individuals) constitute the sole normative ground of love, while in (2010a) he more plausibly claims that histories of shared encounter give us additional reasons for partiality over and above the discrete encounters of which they are composed.} I agree that shared history provides one source of additional value over and above the value of generic goods, but I believe
that Hurka overlooks the fact that some goods attain additional value just in virtue of being produced by a specific person. To return to an earlier example, even if a child’s relationship to her neglectful parents generates no additional value, given the negative quality of their shared history, there will still be a distinctive value to the first instances of emotional support and affection that the parents do show her. In general, though, we should class many special goods in this second category of value, so long as we think of it more capaciously than Hurka does.

When it comes to the question of whether relationships might facilitate distinct goods, and not merely distinctive forms of other goods, Hurka tentatively suggests that there may be a special value of reciprocity. This would arise from thinking of the relationship as a Moorean organic unity: “The whole composed of each of your loves for the other is, because those loves are returned, a separate intrinsic good” (2013: 213). Like Hurka, I find this possibility difficult to assess, because it is difficult to understand how reciprocity or mutuality could make one’s life go better without having any further good features beyond those the relationship already has. Fortunately, I believe that relationships facilitate a less controversial, and less puzzling, distinct good, which I call **epistemic intimacy**.

Epistemic intimacy refers to the condition of being known intimately by another, in a way that is manifest to oneself.⁹⁶ There are many phrases to reach for in describing this condition, such as being known as one’s inmost self, understood at the core of one’s being, or seen for who one is. I find that last phrase in particular suggestive, with its connotations of **recognition**, 

---

⁹⁶ I add the modifier to rule out cases where we might want to say that one party is known intimately by another, perhaps through observation from afar, but does not know this. The definition is also intended to rule out illusory cases in which one party merely believes that she is intimately known. See Cocking & Kennett (1998) for a view that emphasizes the importance of secrets in friendship. Nehamas rightly points out that although I might share the same secrets with others (such as therapists), “when I tell my friends, both of us treat them [i.e., the secrets] as material that can help or hinder the formation of our character and personality” (2016: 223).
although this should not be understood as mere political recognition. Epistemic intimacy, like many other rich goods, is often well-described by novelists. Richard Russo, in his novel *Straight Man*, writes: “Which is why we have spouses and children and parents and colleagues and friends, because someone has to know us better than we know ourselves. We need them to tell us. We need them to say, ‘I know you, Al. You’re not the kind of man who’” (1997: 374). Similarly, Willa Cather, in *O Pioneers!*, has her protagonist Alexandra say: “It’s by understanding me, and the boys, and mother, that you’ve helped me. I expect that is the only way one person ever really can help another” (1992 [1913]: 27). Many, probably all, of us desire to be able to agree when someone says they know or understand us in this way. But this is the kind of good that brooks no substitutions, because the exact way a particular person knows us could not, even in principle, be replaced. This is the phenomenon that the sociologist Georg Simmel has in mind when he writes that the knowledge some person A has of person M is different from anyone else’s knowledge of M, since that knowledge “contains as an integrating, form-giving precondition the psychological peculiarity of A and the particular relation into which A and M are brought by their specific characters and destinies” (1950: 309). Because epistemic intimacy is a distinct good not analyzable as a distinctive form of any generic good, some special goods do belong in this third category of value. I suspect that most of the remaining special goods can be accounted for in the second category of value, as distinctive forms of generic goods, whether their goodness is to be explained in virtue of shared history or not. To summarize:

**Relationship goods**

I. Generic goods (e.g., enjoyment, virtue, trust, gratitude, material resources)

---

97 Philip Kitcher has objected that others don’t necessarily know us *better* than we know ourselves, just *differently*. I agree completely, but think this leaves the core of Russo’s observation untouched.
II. Special goods:

a. Distinct forms of generic goods (in virtue of shared history or not)

b. Distinct special goods (e.g., epistemic intimacy)\(^{98}\)

The procedure I recommend for justifying partiality takes what has been said in this section so far to be sufficient for meeting the second step. Having understood more clearly the kinds of value that relationship goods can provide, and having interpreted the distinctive aim of the practice of partiality to be facilitating these goods, we can assume that the aim is valuable before going on to evaluate how well our current practices are fulfilling that aim and where reforms are required.

2.3 Impartiality as a Non-Ideal Ideal

The preceding two sections investigated the goods that people pursue in being partial to one another. Enjoyment, respect, attachment, vulnerability, intimacy, and so on are the kinds of considerations that motivate us at the first level of ethical thinking. Stepping back from our everyday immersion in these practices, however, we can ask what makes actions motivated by and in pursuit of such goods right. Here we are looking for more fundamental considerations that explain why we have reason to participate in these practices. The answer that NPU offers is that

\(^{98}\) Of course, like almost all goods, epistemic intimacy can suffer from oversaturation, as illustrated by this passage from Hipps’ novel *The Adventurist*: “I would have her remain inscrutable. I would never arrive at that moment a month or year hence when, watching her fuss over one of her eccentricities (she dislikes the color palette assigned to traffic lights, say), I catch myself thinking, Now that’s Madison for you! Thinking this and knowing too she is not truly fussing, that it is a hamming kind of fussing, distracted and fluttering—is in fact a code for us to celebrate our odd-couple pairing, the mercantile rationalist and artistic neurotic; cue for me to sigh and tell her to let it go. It isn’t much of this before a person finds himself fitted to persona, all promise extinguished, all mysteries solved, his ‘style’ set down as surely as an epitaph” (2016: 89-90). In short, we can know someone too well.
these actions are part of a practice that makes things go better by facilitating such goods. And we act rightly when we act in a way permitted by a justified actual practice.

In evaluating partiality, then, we would like to know which actions are permitted by our current practices. To a large degree, this is not so much a matter for philosophical analysis as sociological observation: we have to interpret our practices to determine exactly which norms and rules they embody. Moreover, it is an open-ended question, since our practices are constantly changing as groups work out and revise the norms that govern their attitudes and conduct, sometimes even under the influence of an ethical theory. Though some actions are uncontroversially permitted—comforting a friend, giving care to an aging parent, providing for a child’s education—there are innumerable hard cases as well.

We would also like to know whether there are viable alternatives to our current practices. I said above that I do not take extreme impartialism, on which no form of partiality is permissible, to be a viable alternative, let alone a desirable one. That would be to fall into the trap, diagnosed in Chapter 2, of wrongly identifying false negatives by peering through the lens of ideal theory. NPU takes impartiality to be an ideal in the non-ideal sense: not a final, end-state standard that we must comply with on pain of moral wrongdoing, but a hypothesis to guide proposed reforms. Impartiality does not have the status of an ideal principle but of a generalization about the direction ethical progress has taken in the past. As the quotation from Keller at the beginning of this chapter suggests, impartiality can be understood in a variety of

---

99 One example that springs to mind is the influence, far beyond university campuses, of Peter Singer’s theoretical arguments for animal welfare.

100 There are difficult questions, which I hope to address in future work, about what exactly constitutes a viable alternative to a current practice. A full answer would make reference to the expected value of a positive change weighed against transition costs, with checks in place to avoid status quo bias.
ways; NPU takes impartiality to consist in improvements to overall well-being. Before I discuss one proposal for moving our practices in a more impartial direction—one that could increase the happiness of those affected by partiality—I first want to compare NPU to the most influential ideal utilitarian view in the philosophical literature, the one offered by Peter Railton (1984). Though my view has much in common with Railton’s, there will be several crucial differences.

Imagine that you are happily married, perhaps even a model spouse. You display love, attentiveness, and concern for your partner. The two of you are tolerably well-off, living in an affluent Western society, but not lavishly wealthy. You are generous toward friends and family, but also mindful of the wider world around you, donating around 15% of your income to the charities that you judge most effective. Suppose now that you have a friend named Juan, who has a similar profile. You are both out to dinner one evening (nowhere too fancy) with one of Juan’s friends, a would-be provocateur whom you’ve just met. This gadfly is wondering about marriage as an institution, its goods and especially its evils. He knows that Juan is a principled individual, and asks how Juan’s marriage fits into that larger scheme. Marriage is all well and good, he says, but what about all those other, needier people whom Juan could help if he broadened his horizon still further? Juan thinks for a while, and then says the following:

Look, it’s a better world when people can have a relationship like ours—and nobody could if everyone were always asking themselves who’s got the most need. It’s not easy to make things work in this world, and one of the best things that happens to people is to have a close relationship like ours. You’d make things worse in a hurry if you broke up those close relationships for the sake of some higher goal. Anyhow, I know that you can’t always put family first. The
world isn’t such a wonderful place that it’s OK just to retreat into your own little circle. But still, you need that little circle. People get burned out, or lose touch, if they try to save the world by themselves. The ones who can stick with it and do a good job of making things better are usually the ones who can make that fit into a life that does not make them miserable (1984: 150).

In hearing Juan answer in this way, you might have three thoughts. First, you might appreciate that there is more to be said about your moral choices, and in particular more to be said about why they are right. When asked to justify your love for your partner, you can go beyond the simple response, ‘Of course I take care of my partner—that’s just the right thing to do!’ Constructing this kind of explanation is, of course, one of the principal reasons to adopt a two-level theory as I have characterized it.

Second, you might believe that Juan is on to something when he claims that “it’s a better world” where there exists a practice of marriage than where there is no such practice. This is especially true given the current existence of the practice: the transition costs of ending it (“if you broke up those close relationships”) would be far too great to make that a viable option. This makes Juan’s a forward-looking justification, since his explanation of why marriage is good appeals to its outcomes. Notice also that Juan has (correctly, I believe) left the scope of his standard somewhat indeterminate: there is no implication that in order to meet the standard, marriage as a practice needs to make things go better for each particular person, although clearly marriage needs to make things go better for some number of people (and certainly not worse for a larger number of others).
Third, you might nonetheless worry that something is off about Juan’s answer, in that it seems to imply, first, that his marriage is good only if it is instrumental to producing better consequences for everyone ("you need that little circle") and, second, that marriage in general is good only because it allows people to “do a good job of making things better” for others. It is here that NPU differs from Juan’s answer. Adopting its second-level theoretical standard means that there is a constraint on our first-level practices: if they fall below some threshold of value, then they must be altered. But this does not imply that being instrumental to that other value—happiness—is the only source of their goodness, because we can preserve, at the first level of ethical thinking, all our everyday beliefs about the special goods that marriage and other partial relationships facilitate. At the first level, relationships facilitate various goods for us as individuals in relationships. At the second, theoretical level, we see these relationships as constitutive elements of a larger practice that makes things better overall.

NPU therefore dissents from Juan’s answer in at least three respects. First, it does not apply the theoretical standard directly to individual relationships, but to a particular practice of organizing marital relations. Second, it does not require compliance with impartiality understood as an end-state ideal toward which we must progress, but as a way of making piecemeal changes toward greater overall well-being, without thinking we have to get to the greatest possible aggregate well-being. This is not to say that NPU denies that there are moral requirements—some practices are not justified, so revising them is required—but rather to shift the threshold at which reforms are required. Third, in taking impartiality as a hypothesis to be tested in practice, NPU is open to the idea that we might need to revise away from impartiality in the future. Moving in the direction of greater impartiality might well bring new problems in its train,
problems that we cannot even conceptualize in the present. And if our practices were no longer
able to facilitate special goods such as epistemic intimacy, that would constitute a grave moral
loss. All three of these features make NPU less demanding than Railton’s view, both
psychologically (in the sense in which Reconstructive theories are psychologically demanding)
and practically (in the familiar sense in which utilitarian views require more of agents than other
views of morality).

To this last point, the gadfly issues a skeptical challenge: what about all those other,
needier people that Juan could help? After all, NPU endorses the axiological principle that
everyone’s good, if equal in quantity and quality, is equally valuable. But how could the special
preference that Juan shows for his wife be justified in light of commitment to that principle? Juan
and his wife exchange various goods; couldn’t some of those be better distributed elsewhere
(especially the material goods)? The response is to zoom out from an evaluation of a particular
action—giving charity on some specific occasion—to an evaluation of the larger practice,
thereby justifying the action indirectly as falling under a (justified) practice. Furthermore, the
practice itself prescribes bracketing thoughts of impartial utility, i.e., partiality generates an
exclusionary reason: we should decide to help those with whom we share partial relationships
not on the basis of their impartial moral value but on the basis of our practices of partiality. In
particular, we should choose to help our relatives on the basis of the motivations sanctioned by
the practice—love, concern, reciprocity, and so forth—and not on the basis of a commitment to
axiological impartialism.

How do we know this is an exclusionary reason and not an overriding reason? In this
case, NPU grants that the reasons given by the impartial value of all people are indeed weightier.
We do not have to deny, and can in fact insist, that on any particular occasion, we have weightier reason, from the point of view of impartial utility, to distribute some financial resources to those who are in greater need. But that weightier reason is excluded by the practice, just as, with Rawlsian punishment, a weightier reason to punish the innocent would be excluded by the practice. And since that weightier reason is excluded, it is not the case that, as Hare believes, partiality is a regrettable side-effect of our not being able to be impartially benevolent. Rather, partiality is a practice that is itself (to some degree) justified.

But what justifies the exclusionary reason generated by partiality? This is where the special goods of partiality carry a significant justificatory burden. If the organizing aim of partiality were simply to provide various generic goods, such as financial benefits and occasions for virtue, then we can imagine other practices that could do so more efficiently. But I argued that among the goods facilitated by partiality are special goods, including distinct special goods such as epistemic intimacy. Such goods could be provided in no other way except through something approximating our practice of partiality. And providing special goods typically requires the provision of some generic goods as well. Part of what enables us to develop relationships of epistemic intimacy with others is the trust and vulnerability, and the exchange of material resources, that accompany long acquaintance. Most intimate relationships begin as relationships with a greater degree of interchangeability, and it is only on the basis of providing generic goods that they are in a position to be sites of exchange of special goods.

NPU does not have to admit that our practices are completely justified in their current form, however. Rather, taking impartiality as a non-ideal ideal yields two broad classes of

---

101 I argue for this claim in Chapter 4.
reasons for improvement. The first set relates to distribution of goods within the practice. Insofar as material resources are components of our practice of partiality, then we can ask how better to share them. And here it seems overwhelmingly clear that if the aim of partiality is the beneficial distribution of various partial goods, including generic goods, then the practice suffers from severe shortcomings. Affluent parents often take themselves to have reason to give their children every possible opportunity for advancement, at the expense not only of other children in their social milieu but in the world more broadly. Economically well-off people often think nothing of spending their discretionary income on expensive and time-consuming hobbies that are inaccessible to those in impoverished communities. But the second-level theoretical standard recommended by NPU need not approve of such reasons, since it is implausible to insist that a practice in which people feel free to spend their resources indiscriminately is better than a practice in which people are at least strongly encouraged (if not mandated) to devote at least some resources to beneficial causes. I do not have anything like a full-dress proposal here, but I believe that the structure I have outlined represents the best way to think about these matters, which has been my main goal in this chapter. Rather than deliberating on each particular occasion whether or not to donate one’s resources to someone with whom one stands in a partial relationship or someone with whom one does not, such deliberation should take place at the level of practice-reform. And if we want to move in the direction of improving overall well-being, then we might fruitfully consider work in the standard consequentialist tradition in search of proposals for reform, where these proposals, again, are understood not as moral requirements but as hypotheses to be tested in practice and which we could walk back if necessary.
The second set of reasons relates to expansion of the practice. Unlike those philosophers who claim that relationships themselves have moral importance, and that there is nothing more to be said by way of normative explanation, NPU can account for those who are not currently in relationships, or not part of our practices of partiality. Adopting a more impartial ideal implies that we also have moral reasons to *form* special relationships. These reasons could apply at the individual level, the institutional level, or both (cf. Collins 2013). At the individual level, we have reasons to increase our chances of forming relationships, and to promote others’ relationship-formation. For instance, we have reason to make new friends, to help others become friends, to try to set people up as romantic partners, to adopt children, and to promote adoption. Officially, these are reasons to increase our chances because most relationships of course require the cooperation of the other party. At the institutional level, states should work for beneficial redistribution, even across national borders, and to promote policies for adoption and child-rearing that improve access to resources among all children, regardless of background. But much of the work has to be done at the individual level, because better relationships arise when people feel free to care deeply about their relatives, a state of affairs that cannot be externally prescribed. At a minimum, states should maintain policies of benign non-interference with many of our more ‘private’ relationships, especially given that non-interference is among the conditions under which epistemic intimacy is more likely to arise.

These new reasons may seem controversial, especially at the individual level. But there are three things to say in response. First, they are only *pro tanto*, and thus could be outweighed by other reasons, especially if part of what makes our current practices function successfully is the sphere of autonomy and privacy that they create. Second, these reasons are not totally
disconnected from our current practices. Many people serve as matchmakers, advocate for adoption, and try to make more friends. What NPU implies is that people have moral reasons to do these things. Third, and finally, if non-ideal theory more generally is to have any practical impact, we ought to expect that it gives us novel reasons to improve our practices. These would be reasons that we do not ordinarily take ourselves to have but which, as Wolf puts it, we would nonetheless be “willing to acknowledge” as reasons upon reflection (2010: 117). Again, although it is not my intention to propose a full package of reforms, I believe that the procedure recommended here is the best way forward. My aim in this chapter has been to illustrate how best to think about partiality on a non-ideal two-level picture, and to outline how a second-level standard gives us reason to pursue relationship goods.

At this point, I have outlined my three-step procedure for justifying partiality. I investigated our practices of partiality to determine which goods are most valued in various relationships, I provided an interpretive taxonomy of the kind of value those goods have, and I argued that seeing partiality’s organizing aim as facilitating those goods allows us to justify it to a certain degree but also to construct a non-ideal standard for improvement. NPU justifies particular actions against skeptical challenges by zooming out to see them as falling under our admittedly non-ideal practices of partiality. But whether for partiality or for other practices, we shouldn’t see ourselves as making progress toward some distant end-state ideal so much as progress away from our current non-ideal circumstances.
Chapter Four: Objections and Replies

In this final chapter, I discuss various objections and replies: seven, to be exact. My goals are to tie up loose ends in my treatment of NPU and to relate my discussion of partiality to recent work in the philosophical literature. I conclude by considering NPU as a Constructive theory.

1. Objections and Replies

**Objection 1:** The formulation of NPU in Chapter 2 was motivated in part by the idea that theoretical criteria of rightness should apply in the first instance to practices, and not to individual actions. But surely not all aspects of our ethical lives are regulated by practices. What, if anything, does NPU have to say about these aspects?

Reply: As formulated, NPU indeed applies only to those aspects of ethical life that are regulated by practices. In this it might be regarded as structurally similar to Scanlon’s contractualism, which purports to account only for “a narrower domain of morality having to do with our duties to other people,” as opposed to the broader domain of morality that includes forms of behavior such as idleness or wastefulness, which “are often considered immoral even when they do not harm other people or violate any duties to them” (1998: 6). Although NPU does not draw the distinction in the same place that Scanlon does, I might well be forced to concede that there is a narrower domain to which NPU applies and a broader domain that includes aspects of ethical life that are not regulated by practices.
That concession having been made, though, there are two things to be said to mitigate concerns that NPU is now insufficiently explanatory in scope. First, it seems to me that almost all of ethical life is structured by social practices. Even our ‘private’ lives are, while isolated from the gaze of others, often judged with reference to their standards; it is arguably built in to the notion of a conscience that we enforce internally what would otherwise be imposed externally. And social practices themselves often determine where the line of privacy is to be drawn: a certain practice of etiquette might insist that we should not pry into the health or income of others, but allow us to judge how well an individual holds their fork; a general practice of toleration allows that some matters are left to an individual to decide, while others should be publicly regulated. In that sense, therefore, even private actions can fall under practices in the sense of being permissible or impermissible by their lights. NPU does not, however, claim that we cannot or should not try to step outside of our practices at times; it is at such times that we ought to employ a second-level theoretical standard such as impartiality, particularly in our capacity as reformers. Practices do not themselves do all the justifying work.

Still, someone may continue to maintain that there are hard and fast lines to be drawn here between the practice-governed and the purely individual. So, second, a non-ideal individual utilitarianism could be formulated to account for any remaining ethical spheres. Such a theory would take as its primary target of assessment individual actions, but it would preserve the teleological and non-ideal components of NPU, thereby retaining its commitment to a moral epistemology on which we do not need knowledge of the ideal or best in order to acquire knowledge of the better. This picture, which is inspired by Dewey, maintains that we are mostly guided by unreflective habit until jarred into reflection by confronting a problem, whether at the

158
social or the individual level (Sen 2009; Anderson 2010; Appiah 2017). So whether or not all aspects of ethical life are ultimately thought to be regulated by practices, a non-ideal utilitarian theory will have something to say about how to solve the concrete difficulties we face, as opposed to finding a final, finished, complete way of living.

**Objection 2:** This moral epistemology surely requires further defense. Why think of ethical reflection in terms of problem-solving at all? As Hilary Putnam puts it, “The very words *solution* and *problem* may be leading us astray—ethical ‘problems’ are not like scientific problems, and they do not often have ‘solutions’ in the sense that scientific problems do” (1990: 181).

Reply: This objection turns on understanding problems as finally solvable, or as having definite resolutions, and my reply turns on denying this understanding. Instead, I take problems as requiring continual *re*-solution in the sense that any implementation of a solution should be regarded as local and provisional as we await new problems and new understandings of how to resolve them. To return to the health analogy, briefly mentioned in Chapter 2, when I recover from an illness, I should have no expectation that the condition will never recur, let alone that I am in a state of perfect health. Further, ‘solving’ the problem that a chronic condition poses often requires techniques of management and learning how to live with the condition; here the resolution does not even consist in the removal of the underlying causes so much as a series of strategies for coping with the symptoms.\(^{102}\)

---

\(^{102}\) The analogy comes from Dewey’s “Theory of Valuation” (1988 [1939]: 209-11). I note that for Dewey, the notion of a problem encompasses not only conflict but also unfulfilled needs or lacks.
To take a paradigmatically individual case from ethics, I might discover that I have a tendency to become impatient, whether because others point this out to me or because I find myself frequently rushing to get through things in my life. However this comes to my attention, once I have noticed this habit, it has become a problem to me; rather than unreflectively continue, I must decide what to do about it. One resolution might be to decide not to worry about it, but that path is likely to lead to further problems down the road, such as overly hasty decisions or the irritation of others; a more satisfactory resolution might be to cultivate techniques of breathing and mindfulness, though this is not a permanent solution either; if I travel too far along this path I might find myself becoming irresolute or complacent, which would in turn require further problem-solving.

Indeed, Putnam’s own example is well-understood in this sense of problem-resolution as provisional. He discusses the Supreme Court’s decision on abortion, which held that “a first-trimester fetus does not have legal protection; that abortion of a second-trimester fetus is something to be regulated, primarily in the interest of the mother’s health, though not forbidden; and that a third-trimester fetus must be amply legally protected” (1990: 182). Putnam rightly points out that there are many issues here that have not been ‘solved’, such as the question of when personhood begins or the extent of women’s right to privacy over their bodies. But a non-ideal theorist can agree with Putnam that the Supreme Court’s decision embodies a successful adjudication of a difficult dispute without claiming either that it is the ‘last word’ or that it is the ideal end-state solution.
Objection 3: NPU clearly has some affinity with actual rule utilitarianism, the theory on which agents have a permission to follow the rules of their current societies, so long as those rules are not structurally incompatible with happiness-promotion (Eggleston and D. Miller 2007; R. Miller 2009). But what if the rules, or practices, are unjust in other ways?

Reply: I begin by rehearsing the two main attractions of actual-code theories. The positive attraction is that actual codes, or practices, have instrumental advantages over ideal codes or action-based codes: they lessen the likelihood of individual rationalization, promote the goods of social coordination, facilitate belonging to a moral community, and secure freedom from undue blame. The negative attraction is that actual codes are non-ideal: they do not offer timeless standards of rightness, but focus ethical evaluation on transitions. Yet the primary objection to actual-code theories is that they are too conservative, producing nothing more than a fancy defense of the status quo. In fact, this worry is only heightened by the focus on viable transitions between practices rather than practices as such, because expected transition costs will often be a legitimate reason not to alter our practices too much. For instance, if a proposed reform to our practices of partiality—raising children communally, for instance—required breaking up many existing parent-child relationships, that would be a strong reason not to adopt the proposal.

The best response to this objection is to recall that actual-code theories are importantly distinct from conventionalism, the view that we have most reason to follow the actual conventions of our society, full stop. A theory like NPU places an important moral condition on a society’s conventions, namely that they do a better job at promoting happiness than viable
alternatives (where the viable alternatives include past versions of the practices, which are known to have been actual at some point in time). So the theory need not endorse all our current practices, and is thus not identical to conventionalism. In fact, it recognizes two sources of moral requirements: (a) those internal to existing practices, such as requirements to care for one’s children or not to punish the innocent, and (b) those external to existing practices, such as requirements to reform, or act in such a way as to reform, practices that fail to do better than viable alternatives.

Still, the present objection makes reference to other kinds of injustice. Perhaps certain expressive harms do not affect individual happiness, even unknowingly, or perhaps an entire society is mired in false consciousness about its habits of economic exchange, though no one member can articulate this. The response that NPU offers here leans heavily on its teleological component: if some putative injustice does not ultimately make a difference to individual quality of life, then there is no reason to strive to change it. But nearly any diagnosable injustice does have consequences for individual welfare: race-based housing discrimination, lax workplace harassment regulations, and even voter suppression laws arguably cause or constitute welfare harms. The onus is on any would-be reformer to show how a proposed remedy answers to what William James calls “the cries of the wounded” (1956 [1897]: 210), where people can be wounded in a variety of ways.

The objection can be answered even more strongly by noting that the capacity to reform is built into our practices, since acting as a reformer is permitted by a justified actual practice. Animal-rights enthusiasts, LGBTQ+ activists, effective altruists, and other movements for ethical change are all, by the lights of NPU, behaving permissibly. Some might find reformers
obnoxious, but that does not mean they are morally wrong, so long as the means they employ are not forbidden by other justified actual practices. The theory does not tell us precisely which movements we should support. But that is one of its strengths, because the question of whether these movements are in fact happiness-promoting is one to be settled only by experiment over time.

**Objection 4:** Your standard of evaluating partiality is goods-based. But aren’t goods-based standards susceptible to the ‘trading-up’ objection?

Reply: The objection to applying a goods-based standard to some evaluand is that it requires trading up when there is even a slight increase in expected value. My account of partiality avoids the trading-up objection, however, because it is highly indirect. It does not evaluate actions directly in terms of the second-level standard, as act-utilitarianism does.  And it does not assess actions indirectly in terms of the dispositions that govern them, as motive-utilitarianism does. It assesses actions indirectly in terms of the practices that govern them, where those practices themselves recommend certain dispositions. This makes the theory even more indirect than most, and it thus approves of actions that other standards might not.

---

103 This is Railton’s view, and it requires him to say that there can be acts of objective wrongdoing that are nonetheless in keeping with the character that it is best to have: “Juan should have (should develop, encourage, and so on) a character such that he sometimes knowingly and deliberately acts contrary to his objectively consequentialist duty” (1984: 159). NPU can maintain, *contra* Railton, that many such actions are right.

104 This view is developed, in somewhat different ways, by Adams (1976) and Mason (1998).
**Objection 5**: That seems like a dodge! Insofar as agents apply a goods-based standard not just to their practices of partiality but to particular relationships, then they would still face the trading-up objection. So what about particular relationships?

**Reply:** This is a useful place to emphasize an important feature of my theory, namely that NPU is officially *silent* on the reasons internal to our practices themselves. As a highly indirect theory, it limits the kinds of questions we can reasonably ask about our practices from the point of view of ethical theory, since it holds that the primary locus of theoretical assessment is practices, not particular actions.\(^{105}\) Yet as Chapter 3, §§2.1-2.2 suggests, there is still work for philosophers to do that is not at the second level of theorizing *per se*. In particular, there is the interpretive work of identifying and then evaluating the goals, aims, and reasons internal to our practices, as with, for example, the first two steps of my account of partiality. So my answer to this objection will take place at that interpretive level, the first level of ethical thinking.

Clearly, many people do not deliberate about whether to opt into a practice of partiality or not; rather, we find ourselves already behaving in partial ways. But we do deliberate about whether to form or maintain particular friendships, romantic relationships, or family relationships. And here, I also hold that what explains our beginning or continuing some relationship is whether it is expected to facilitate certain relationship goods. Again, however, this does not follow from NPU, which is compatible with some other account of agential deliberation.

---

\(^{105}\) Jordan MacKenzie (2017) emphasizes this point in her recent account of agent-regret as a constitutive element in a social practice of moral luck. For MacKenzie, it is not reasonable to ask whether agent-regret is, in and of itself, rationally justified; we can only ask (1) whether a particular instance of agent-regret is justified according to the practice and (2) whether our practice of moral luck is, as a whole, rationally justified.
within a practice of partiality. But since I hold that a goods-based standard of evaluation is doing some work even at the first level of ethical thinking, I want to address the trading-up objection at that level, too.

It is tempting to think that the trading-up objection is especially worrisome if our goods-based standard is maximizing, aiming for the highest expected value. But even if the standard is less demanding, such as a satisficing standard that only requires the relationship to meet some suboptimal level of value, the trading-up objection would still apply so long as a given relationship fails to meet the standard.106 The objection is most fundamentally about thinking in terms of the provision of goods at all, no matter at what threshold a particular relationship is assessed as suboptimal. As Joseph Raz puts it, “Being engaged in a pursuit or a relationship includes belief that certain options are not comparable in value. Abandoning such beliefs is therefore one way of abandoning the pursuit. Regarding a particular relationship as a proper subject for an exchange damages or even destroys it” (1986: 356).

The first thing to be said in response is that some people simply deny the force of the trading-up objection. Emerson believed that we should periodically replace all our friends, a policy justified by his views that friendship is ultimately a means to self-cultivation and that self-cultivation requires regular infusions of new ideas and stimuli. Going further back, Plato’s Symposium can be interpreted as an argument in favor of trading-up: as we ascend Diotima’s

---

106 In any case, I am inclined to think that worries about maximizing are, when applied to deliberation, a side-show: most often we do not know what to maximize (how do I balance short-term hedonic pleasure and long-term satisfaction in a romantic relationship?), or how (even if I know that I want to maximize my own hedonic pleasure in a romantic relationship, how much do I need to care about my partner’s pleasure in order to do so?). This is not to deny that maximization cannot be the basis of a successful objection to the criterion of right action offered by a theory; in fact, I think that such objections tell against maximizing criteria in general (cf. Norcross 2006). See also my discussion of practice-based theory in Chapter 2.
ladder, we should no longer form relationships based on certain kinds of properties of the beloved (e.g., physical beauty), but based on their having different, better properties (e.g., beauty of character, beauty of belonging to a city that instills such a character). If that is the case, then we would have reason to replace our beloveds if they did not possess the better kinds of properties.

Since such views may have only niche appeal, let me instead try to understand what makes Raz’s objection compelling. First, it could seem compelling if the goods-based standard is thought to be egoistic. If I am just out to maximize my own pleasure, then I am indeed likely to damage our relationship sooner or later. But I deny that our standard, within partiality, should be egoistic. Instead, it should concern the goods produced for both parties to the relationship. Perhaps I think that you could find a more suitable partner, and that our incompatibility is preventing you from attaining various relationship goods. There need be nothing egoistic about ending a relationship so that you can trade up. Second, the objection could seem compelling if a loving relationship is supposed to be unconditional. But a loving relationship should not be unconditional, because everyone should agree that there are some reasonable grounds for ending some relationships (cf. Edyvane 2003). If your partner, or your parent, or you yourself, underwent a drastic personality change, such that you became uncontrollably violent, or wildly prejudiced, or viciously hateful, then you would no longer have reason to pursue the relationship: the expected goods it provides would have dropped below whatever threshold of value is set by the standard. Perhaps you might still have all the attitudes and emotions of love toward such a person, in virtue of your past history, but an ongoing relationship would no longer be feasible.
Dispelling these two misconceptions allows us to see what is misguided about Raz’s objection, and in particular why it rests on a faulty picture of deliberation. The fact is that everyone should regard their particular relationships as proper subjects for exchange, at least conditionally.\footnote{As I argued in Chapter 1, §2.} Part of what it means to have a commitment is to have conditions under which one would reasonably revise those commitments. But the mere thought that there are conditions under which one would end a relationship certainly does not imply that one intends or plans or even considers doing so now. As Raz himself insists in discussing exclusionary reasons, the mere thought of acting on a reason need not in itself have negative consequences for action.\footnote{As I quoted in Chapter 2, “So long as one knows that one’s reflections will not affect one’s action, the ill effects of such thoughts are avoided” (Raz 1999: 184).} Certainly it is true that intending, planning, or even seriously considering ending a relationship means that the relationship is damaged or destroyed, but this is not because we consider ending it. Raz gets things the wrong way around: in cases where considering the end of a relationship is a live option, that is because the relationship itself is already damaged.

In short, we need to have some sense of when it is reasonable to form or maintain a relationship. But again, NPU is officially silent on that topic, leaving that question to our current practices to decide. In fact, the theory forbids the scrutiny of particular relationships in terms of its more impartial utilitarian standard, since such scrutiny is precisely what gives rise to the skeptical challenge to partiality.\footnote{This point is well emphasized by Tedesco (2006), who notes that, from the perspective of a consequentialist standard, most particular acts of, say, friendship, and most friendships themselves, will be found unjustified because suboptimal. If the standard is applied to the pro-friendship disposition, as Mason (1998) urges, or to our practices of friendship, as I urge, then we must not also apply it to particular acts or particular friendships.} So we should not reject or end particular relationships because they fail to meet the utilitarian standard, but because they fail in other ways according to our
practices. Different practices may have different senses of when it is reasonable to end a relationship—compare medieval Christian and contemporary secular views on divorce—and in striving to alter our practices, we should deliberate in terms of the utilitarian standard—doesn’t the contemporary secular view prevent years of embittered unhappiness, even if it may be imperfect in other ways?—but the trading-up objection itself should not be a cause for concern.

**Objection 6**: One recent debate about partiality concerns the ground of partial treatment, and whether our reasons to behave partially are ultimately grounded in facts about (the value of) our own projects, the relationships we stand in, or the individuals with whom we share relationships (Keller 2013; Lord 2016). Does your view take a stand here?

Reply: Yes! Non-Practice Utilitarianism claims, as a teleological theory, that all of our ethical reasons are ultimately justified by considerations of goodness. In particular, its criterion of rightness makes reference to whether our practices better promote the happiness of those affected by them than the viable alternatives do. This is the happiness of individuals, and thus my theory does take a stand on what ultimately grounds partiality. The value of relationships and the value of personal projects must be cashed out in terms of how they affect individual quality of life.

Yet at the first level of ethical thinking, any of these three kinds of facts might serve as motivating reasons. In being motivated to behave partially toward someone, I might permissibly concentrate on their effects on me (‘I would hate to be single’) or on the mere fact that she stands in a certain relationship to me (‘she’s my mother!’), without focusing on properties of the individual him- or herself. Our current practices don’t seem to forbid thinking in these ways,
though perhaps on some interpretations it would be problematically selfish to be motivated in these ways *alone*. But most of us, I take it, have had motivating thoughts of each of these three kinds.

Furthermore, my theory also has the resources to respond to Keller’s objections to what he calls ‘a consequentalist version of the individuals view’. Keller’s first objection is that consequentialism gets our reasons wrong: though he concedes that the view “has the resources to explain the form of reasons of partiality—how we could have (derived) agent-relative reasons and how they could be morally significant—it will struggle to explain their content” (2013: 127). Because my view is non-ideal and practice-based, however, I am happy to allow that, to borrow Keller’s example, a parent has stronger reason to give her children treatment for severe asthma than to make a donation that will save multiple children from blindness. Our current practices would even forbid acting in such a way as to save other children at the cost of one’s own child’s life. The constraint on our first-level reasons from the point of view of theory—the ‘derivation’ of our ordinary reasons from a theoretical standard—takes place at the level of practices, not at the level of individual actions. Keller’s second objection is that consequentialism is wrongly committed to condemning our ordinary reasons to act: “the consequentalist will have to say that when we take ourselves to have reasons that are not answerable to assessments from the impartial point of view, we make a mistake” (2013: 129). But the Constructive element of my theory means that there is no commitment to seeing our second-level reasons as fundamental or our first-level reasons as illusory because non-fundamental.
Objection 7: Another recent debate in the literature on partiality concerns the value of relationships, and whether they have fundamental or only derivative moral value (as briefly mentioned in Chapter 3, §1.1). Does your theory take a position here?

Reply: Yes! I hold the latter, reductionist position, and in the course of showing why NPU is committed to reductionism, I hope to offer one final consideration in favor of the practice component of the theory. I begin with some terminology. Non-reductionism about relationships is the claim that relationships have fundamental moral importance, while reductionism is the claim that relationships have at most derivative moral importance.\footnote{I adapt this formulation from Wellman, who notes that reductionists and non-reductionists (whom he terms associativists) can “agree about the content and stringency of any given obligation among intimates” (2000: 540). Agreement at the first level of ethical thinking does not preclude debate at the second level about the ground or normative explanation of these obligations or, more broadly, reasons.}

Something that has fundamental moral importance has unconditional reason-giving force, meaning that thing itself is a source of reasons independently of its meeting other conditions. For classical ideal utilitarians, pleasure (and only pleasure) has fundamental moral importance: the fact that something will produce pleasure is itself a source of reasons, even absent any other facts about it. Something that has derivative moral importance is a source of reasons only conditionally on its also having something of fundamental moral importance. A given practice will have derivative moral importance for ideal utilitarians if and only if it is likely to produce pleasure. Partial relationships, for instance, would have at most only derivative moral importance for such utilitarians.\footnote{As a matter of terminology, I understand the claim that relationships have intrinsic value to entail non-reductionism, and reductionism to entail the denial of the claim that relationships have intrinsic value.}
Inspired by my discussion in Chapter 3, §2.1, we might pursue this latter position and observe that if relationships have value only because of the goods they instantiate or facilitate, then reductionism about relationships is true. This can be motivated by the thought that many relationships are abusive, destructive, or dysfunctional. Some marriages are sites of radical inequality at which one spouse benefits greatly to the detriment of the other. Some friendships are occasions for vice in which both friends enable each other’s bad habits. And some familial relationships produce nothing but anxiety and ill will for all parties involved. We might well judge that these relationships are not, on the whole, valuable. This is not to say that they could not contain valuable elements: an abusive marriage may result in happy, well-adjusted children; a vicious friendship may provide companionship and joy; and a dysfunctional sibling relationship may yield opportunities to deploy one’s devastating wit. But on the whole, the burdens of the relationship would outweigh the benefits, leading us to render an overall negative value judgment.

At this point, however, the defender of non-reductionism about relationships could object that these examples have no force, because they are not examples of the kind of relationship that has fundamental moral importance. Raz, for instance, who holds that relationships have intrinsic value, claims that there is a strong normative component to friendship, such that a relationship between people who simply enjoy amusing themselves in each other’s company is not actually a friendship (1989: 19). Having a friendship, according to Raz, involves seeing yourself as having duties to your friend. This just pushes the problem back a step, however, because some such duties will not be valuable. There is honor (and duty) even among thieves, and the overly servile wife is servile in part just because she sees herself as having extensive duties to her husband. The
mere existence of a duty does not establish that the duty is valuable. So it seems that reductionism is true.

Consider, then, an objection to reductionism, intended to show that relationships have value over and above the goods they facilitate. Imagine a world in which there are no relationships as we understand them. People do not regularly give preference to certain friends over others. There are no long-term romantic couples. Children are raised in common and not by the same guardians. Nonetheless, all the goods that relationships facilitate in our world are enjoyed in this imagined world. It’s just that the goods are dispersed randomly and not in the patterns that partially constitute relationships. I help someone who is carrying her groceries, and then later a stranger helps me carry my groceries. I give money to a child for her education, and then later a different child calls me up to ask about my day. Call this, following Philip Kitcher, the “Secret Santa” world, or the “Random World” (2011: 200). It seems that the reductionist is committed to denying that there is any loss of value in the Secret Santa world. But, intuitively, there is something wrong with such a world.

The reductionist, however, need not deny that there is a loss of value in that world, because she can deny that the thought experiment is plausible: if goods really were distributed at random, then the Secret Santa world would not promote the same goods as our world does, because it would exclude special goods. If I am in need of comfort, a stranger may be able to listen to my woes, but she will not understand me as well as a friend or family member might. At this point, then, the defender of non-reductionism about relationships can try to strengthen the

112 Friedman makes a similar point against the claim that all relationships have intrinsic value, saying that it “fails to take account of the differing moral value of different personal relationships, and it mistakenly presumes that whatever sustains any personal relationship is a moral good without qualification” (1993: 42).
objection. Imagine, *per impossibile*, that all the special goods are facilitated. When you need a shoulder to cry on, your friend will (at random) appear. When you want to reminisce about your childhood, someone who knew you back then will (at random) be there to talk with you. Exactly the right person will be there at exactly the right moment. But then, I claim, the world will begin to look a lot more like ours. In order for these special goods to be facilitated, then normative expectations will need to develop. If I am to enjoy the good of feeling secure in our relationship, you would need to be available to me at times and places we previously agreed upon, such that I could reasonably criticize you if you failed to show up. Even if we can imagine, as Kitcher does, that the Random World were possible through massive amounts of false belief—“that the guilty suffer *because* of the suffering they inflicted and that the generous rejoice *because* those they help express gratitude” (2011: 200)—we can see that it has no bearing on the actual world, in which such a scheme could not be set up. In our actual, non-ideal world, we can safely conclude that reductionism is true.

This thought experiment illustrates two points. First, it is impossible to evaluate a relationship apart from the goods and ills it brings about. We are simply not able to assess whether a token relationship has fundamental moral importance independently of knowing which features of that relationship make it good or bad. Thus even if we had a good, non-stipulative answer to the question of whether the relationship *itself* is valuable, we would not be able to assess how it gives us reasons. The Secret Santa world is therefore additional evidence for passing the explanatory buck from relationship types to relationship goods. Second, facilitating relationship goods depends on our social practices. As my definition insists, partiality requires a system of rules that define roles, expectations, penalties, and so forth. *A practice* of partiality, as
opposed to one-off acts of charity, vulnerability, and suffering, demands stable, repeated patterns of action that sufficient others take themselves to have reason to conform to. Reflection on the Random World thus provides one final consideration in favor of appealing to practices as the primary target of ethical theorizing about partiality.

2. Conclusion: Non-Ideal Practice Utilitarianism as a Constructive Theory

The Reconstructive view of the role of theory in everyday deliberation, as I characterized it at the beginning of this project, is committed to three major claims: (1) we should be able to use ethical theories to derive what to do in any conceivable situation; (2) we should have an overriding commitment to following our favored theory; and (3) because of this commitment, our everyday modes of deliberation are justified only to the extent that and because they are sanctioned by the theory. Constructive views deny each of these three claims. Here I explain how NPU is a Constructive theory by showing, first, how it denies the three Reconstructive claims, and, second, how it responds to the instability objections from Chapter 1.

As a reminder: NPU is non-ideal, in claiming that in order to be justified, a practice needs only to facilitate more good for those affected than the viable alternatives; it is teleological, in appealing to the happiness of those affected by a practice to rule out some practices and rule in others; and it is practice-based, in evaluating actions not directly in terms of happiness-promotion but indirectly in terms of compliance with happiness-promoting practices. As with other indirectly utilitarian accounts, the criterion of right action that emerges flows from the rightness of the broader evaluand, in this case a practice:
Criterion of right action: An act is morally right if it is permitted by a justified actual practice in one’s society, where a justified practice is one that better promotes the happiness of those affected by it than the viable alternatives do.

(1) It is clear that NPU cannot be used to derive what to do in every conceivable situation, because it insists that the ethical status of actions depends in part on practices. All ethical theorists agree that some empirical facts are necessary in determining what to do—the standard utilitarian who holds that an action is right just when it brings about the best outcome, or the virtue ethicist who holds that an action is right just when it would be performed by the fully virtuous agent, would both need facts about the outcome space or circumstances in question—but once these facts are taken as inputs, these Reconstructive theories themselves label which actions are, say, supererogatory or obligatory. But on my Constructive view, these other deontic categories are not derived from the theory; rather, our practices determine the deontic status of most actions, and NPU is used only to check whether a practice is itself justified. Unlike the maximizing act-utilitarian theory that would label as wrong giving away, say, only 20% of one’s income to effective charities (as opposed to some larger percentage), NPU would, in my current social context at least, label that action as supererogatory—though the status of that action might well change over time as our practices do. Thus, most of our ethical deliberation takes place within and is structured by our practices—the judge who is deciding how to punish, or the partisan newspaper editor who is deciding how to respond to an offensive letter—rather than in terms of NPU itself.

(2) In fact, there would be a further obstacle to using NPU to derive the single right action, which is the same reason that an agent’s commitment to NPU is not overriding: there are
(at least) two independent ways to comply with it. Standard act-utilitarians are motivated by an overriding commitment to promote the good, and rule-consequentialism can be motivated by an overriding commitment to do what is impartially defensible (Hooker 2000: 101). But on NPU, we can either do what is permitted by our practices, or we can seek to reform them. Neither commitment is automatically overriding, since on the Constructive view, we are independently committed to our practices, and not only on the grounds that they are justified in the light of theory. NPU isolates one component of those practices—the fact that they can promote the good—and uses that as a standard for revision. But, following Mill, the content of morality cannot be derived from the good. As John Skorupski puts it, glossing a Millian view,

the utilitarian theory of the good is the appropriate ‘controller’, in his word, of all our principles, ideals and ends. It is not their generator. They are not derived from it, or reducible to it, even though they are defeasible by it. They have relative autonomy.

According to this utilitarian view, we have many normative principles governing action that are justified in their own terms, without derivation from the final good; but they must give way if they turn out to be systematically [i.e., as a system or practice] incompatible with that final good (2006: 37).

(3) It is the fact that the good is the controller of our practices, and not their generator, that explains in turn why, finally, NPU denies that everyday deliberation is justified only to the extent that and because it is sanctioned by the theory. On Reconstructive two-level views, the

---

113 There is a question, probably more difficult than I would like to admit here, about what to do when practices conflict. To adapt the simple answer typically given by actual rule-utilitarians, agents are obliged to conform to all the rules that flow from the legitimate practices to which they belong, where rules are to some degree weighted by the practices themselves (R. Miller 2009: 22-3). Though I think this answer will address many conflicts between practices, I doubt it will address all. But perhaps this implication is a feature of the account, which stresses the benefits, for deliberation, of not being able to derive what to do in any conceivable circumstance.
only value of everyday deliberation is instrumental, as an implementation of the ideal criterion of
rightness, and thus the only value everyday deliberation can have is derived from the theory. But
on the Constructive view, everyday deliberation, and our practices more generally, have value in
their own right. Certainly they can have instrumental value, and NPU exploits this fact in
guiding efforts to reform—how could our practices of blame, for instance, better satisfy our
general aim of preventing future wrongdoing?—but they do not have only instrumental value.

At the end of my lengthy Chapter 1, I only mentioned, with reference to Strawson, why
the Constructive view of the role of theory is not in general subject to the various instability
objections lodged against Reconstructive views as a whole and Hare’s two-level utilitarianism in
particular. Here I spell out why NPU can avoid those objections as well.

Clearly bifurcation, in Bernard Williams’ sense, is no concern, because there is no “willed
forgetting” (1985: 109) in having the two or more commitments that accompany NPU. An agent
does not understand her dispositions in purely instrumental terms, but “sees the world from the
point of view of [her] character” (1985: 108), whether she is immersed in the norms of her local
practices or reflecting on how best to reform them. Although I conceded that even a
Reconstructive theorist such as Hare can perhaps overcome the bifurcation objection, the other
two instability charges were more serious.

The alienation objection holds that there is normative failure in being made to feel
alienated from one’s psychological makeup simply on account of coming to adopt an ethical
theory, as Harean utilitarianism entails. The Constructive theorist, by contrast, is able to insist
that there is no reason to regret one’s intuitive dispositions as a class. Again, the independent
commitment to one’s practices that NPU endorses does a lot of work here, since while it is
perfectly possible to regret individual instances of one’s dispositions—in considering one’s blaming practices, one might strive to become less judgmental, for instance—and while adopting a conception of how to reform one’s practices in the light of the second-level utilitarian standard might give one reason to regret larger swathes of one’s dispositions, there is arguably no viable alternative to our current practices on which regretting one’s intuitive dispositions as a class would become appropriate.

Neither version of the inconsistency objection can be made to stick, either. First, NPU denies that the reasons on which we act when immersed in our practices are illusory, inaccurate, or in global error, as has been charged against indirect act-utilitarianism in particular. Second, NPU does not claim that we selected our current dispositions or practices, or that these are the optimal dispositions or practices to have. That falls directly out of the non-ideal component of the theory.

I am sure that there are further objections one could raise to NPU in its own right, and in future work I hope to draw out its implications in more detail. But the question with which I began this project was how a commitment to an ethical theory could guide an agent’s ordinary life. Though I said from the outset that I would not have a full answer to that question, it has been my contention throughout that we should assess ethical theories not only in terms of their epistemic merits, but in terms of the effects that adopting them would have on our deliberative lives. Pursuing that line, I therefore argued that not only the role of theory, but its form as well, should be constrained by facts about our limited deliberative capacities. Going non-ideal can put to rest many of the older objections that critics raised to ethical theorizing.
Though different value-orientations will produce different flavors of non-ideal theories, I motivated one such utilitarian-style theory, in the first instance in order to illustrate in more detail my primary claim about the relationship of theory to deliberation. But I also showed that NPU has substantive and attractive normative implications for our practices of regret, toleration, punishment, and, most of all, partiality, which gives us independent reason to consider it further. Understanding ethics as a matter of improving our practices in piecemeal fashion, guided not by a distant end-state ideal but by our diagnoses of current problems and attendant hypotheses about viable solutions, produces not only a healthier picture of ethical theorizing, but hopefully a better ethics as well.
Bibliography


Eggleston, Ben and Dale E. Miller. 2007. India House utilitarianism: A first look. *Southwest*


Macleod, Christopher. 2016. Mill, morality, and malleability. Talk delivered at the International


