NORMATIVE PROGENY:
A MODEL OF MORAL JUSTIFICATION

Justin David Jeffrey

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Philosophy.

CHAPEL HILL

2007

APPROVED BY:
Geoffrey Sayre-McCord
Thomas E. Hill Jr.
Susan Wolf
Douglas Maclean
Bernard Boxill
ABSTRACT

Justin Jeffrey

Normative Progeny: A Model of Moral Justification
(Under the direction of Geoffrey Sayre-McCord)

Attempts to justify evidently important parts of morality regularly seem to succeed as justifications only by robbing these same parts of their importance. If asked to defend the goodness of generosity, for example, one might point to its salutary effects. But to the extent that this defense succeeds it seems simultaneously to imply that it is not generosity, but rather generating salutary effects, that is valuable. The “defense” of generosity’s value has apparently exposed it as “merely derivative”, a phrase that often amounts to a philosophical slur. But in this and other cases we might hope to do better.

What is needed, and what I offer in this dissertation, is a clear articulation of the problem’s structure and then, against this background, a specification of the criteria a proposed solution must meet in order to be successful. I argue that to solve the problem a kind of normative “transfer” must take place between objects of justification and their justifying ground, allowing the recipient to attain a kind of normative independence in spite of its obvious justificatory dependence. I call this phenomenon “normative inheritance”. If defensible, normative inheritance would allow for a proper vindication of rights, rules, virtues, and so on, whose normativity is apparently derivative but, hopefully, not merely so.
PREFACE

The title of my dissertation, “Normative Progeny”, refers to the central metaphor of my project; my dissertation is not literally about human parents and their children. As it happens, though, the seeds of my topic were planted by my frustration over a recurring argument I had with my father. I was in seventh or eighth grade, and I didn’t like to go to school. I probably didn’t like school for the usual reasons children don’t like school. But I also thought that time in class was spent inefficiently, and that I could learn everything I would learn in school by staying at home and simply doing my homework over the course of an hour or two. My confidence in this strategy rested in part on a belief that I was the smartest boy in the world.

I don’t think my dad agreed with me on this last point, and in any case he invariably rejected my weekly suggestion that I stay home from school and fake illness. Our argument concerning my proposal followed a predictable pattern. At the beginning of the car ride, I would present my case for skipping school, focusing on the inefficiency of the learning process and the incompetence of my teachers and peers. He would respond that even if my school did not use my time efficiently, pretending to be sick was dishonest, and being dishonest was wrong. End of story. But what’s wrong with lying, I would ask, when you can get away with it and everyone is better off, or at least no worse off, as a result? I told him that I wasn’t planning on making a habit of lying. And I would be discreet; I could keep a secret. Could he?
He said I was missing the point. Lying isn’t wrong just because of its bad consequences (which, he seemed to concede, were not *always* bad). Lying was wrong in itself (this, anyway, is how I interpreted him at the time; in reality, his views about morality were surely more nuanced than I have represented them here). Thinking that I had invented consequentialism, I felt that his claim about lying was inadequate and stood in need of defense—in particular, a defense that rested on the consequences of lying.

I was, it turns out, a muddled but committed consequentialist, and stayed this way from my middle-school years until the end of college. During this time I found consequentialism—more precisely, hedonistic act-utilitarianism—to be so obviously true that I was genuinely astounded that any (apparently) reasonable person could seriously entertain alternative moral viewpoints. But the dogma finally began to lose its grip on me as a graduate student at Chapel Hill. I still refused to accept the view that certain actions, like lying, were wrong in themselves. If lying was wrong, surely something else made lying wrong. I remained attracted to something like the utilitarian theory of value—the theory that happiness, and only happiness, is intrinsically good. But I became unwilling to accept the austere moral system that apparently resulted from this commitment.

I wondered if I could cling to my utilitarian moorings regarding intrinsic value but somehow broaden or enrich my account of what was “good in itself”. I was puzzled by the utilitarian claim that because virtues derive their value from utility, they therefore were “merely derivative”. Couldn’t the virtues be justified by their relationship to utility, and nevertheless have a robust normative life of their own? Or what about other parts of morality, like moral rules? I wondered about the rule against lying. Surely, I thought, if there were a justified rule against lying, it derived its force from its connection to utility; still, could its
normativity be freestanding, such that it carried weight even when following it didn’t promote utility? If this was right, though utility would ground the importance of truth-telling, truth-telling would—like a child sustained by her parents’ care who nevertheless possesses strength of her own—have a kind of normative independence. I began to conceptualize these questions in terms of this genetic metaphor.

I mentioned that my dissertation was not literally about parents and their children. It is also not about utilitarianism. Utilitarianism was simply my way in to what turned into a much broader, and I hope more interesting and deeper topic about morality and moral justification. A number of invaluable conversations with my dissertation director and graduate student advisor, Geoffrey Sayre-McCord, helped me to see that the problems I encountered in utilitarianism could be generalized to encompass a whole host of individual justificatory problems. The problem, it seemed, was not confined to one moral theory or another, but threatened moral justification in remarkably general way. Specifically, it appeared that attempts to justify evidently important parts of morality regularly, and unfortunately, robbed their objects of their importance. Morality—or at least many parts of it—demanded justification. And yet in answering this demand, justification seemed to undermine rather than support morality.

As I have already discussed, the utilitarian defense of the virtues trivializes their importance by making them “merely derivative”. But we can go beyond Utilitarianism. Kantians and neo-Kantians try to show that the things people rationally value are thereby valuable “as ends”, or “objectively” valuable. The effect here is the same: far from possessing a free-standing value, the value of such objects would be problematically tenuous, contingent on being actually valued. Virtue ethicists try to show that certain states of affairs
are valuable simply because they are the ends of virtues. Some argue, for instance, that happiness is good simply because it is the end of beneficence. This justification, though, is again undermining of its object. The value of happiness appears to be free-standing in a way that the virtue ethicist cannot account for.

I’m no longer impressed by the arguments for the fundamental role of utility in justifying the rest of morality. I have become extremely impressed, however, by the widespread even if only tacit recognition among moral philosophers that justification threatens to undermine the very parts of morality it aims to shore up. Even my dad and I seemed to be aware of the problem when we debated the morality of lying in order to skip school—it was a part of the background of our argument. We tacitly agreed that if I was right that lying’s wrongness depended on its consequences, then my dad was wrong to claim that lying was wrong in itself. We also tacitly agreed that, conversely, if my dad was right that lying was wrong in itself, then I must have been wrong to argue that something extrinsic to lying was what made lying wrong.

As impressed as I have become by the widespread recognition of the undermining effects of moral justification, and its undiscriminating affliction of nearly every type of moral view, I am equally impressed by the wealth of attempts to fight these effects and make justification work the way it is “supposed” to work. Though there is no general (or specific) approach to the problem I raise that strikes me as successful, I hope that the rough shape of what a solution would have to look like begins to emerge as I work in later chapters to eliminate various kinds of logical space. In the end, I am hopeful that that the stubborn but apparently rival intuitions that seemed to undergird the argument between my dad and I can be reconciled. That is, I am hopeful that—as my dad seemed to think—some elements of
morality are good or normative in themselves, but only—as I would (more or less) claim—because they are the fortunate recipients of moral justification.

I have aimed in my dissertation to discuss the structure and logic of morality and moral justification rather than defend or criticize a particular normative view. To the extent that I have succeeded in this aim, I must give my advisor Geoffrey Sayre-McCord much of the credit. I have had a tendency to get caught up in argumentative details of one philosopher’s theory or another that were not pertinent to my “meta-ethical” project, and Geoff has patiently and repeatedly managed to extract me from these unfruitful forays. I have treated individual substantive views with relative and I hope appropriate brevity, with an eye to the kind of structural justificatory relationships relevant to my thesis. Whatever might be lost in approaching things this way, I am optimistic that this method has allowed for an exploration into deeper themes surrounding justification than an alternative account that delved more into particular substantive issues would have allowed.

I am also deeply grateful to Thomas E. Hill Jr. for his careful reading and detailed feedback on multiple drafts of a topic that he only slowly warmed to. He helped to correct for another tendency of mine, familiar to graduate students who work on a dissertation over the course of a year or more. I tended to drift towards idiosyncratic terminology and obscure concepts not sufficiently anchored in the familiar substance and vocabulary of recognizable moral discourse. Not unrelatedly, he pushed me to continue to work to motivate my topic, encouraging me to discuss the kinds of substantive issues that were in fact relevant to motivating and supporting my thesis.

I thank Susan Wolf as well for her challenging but encouraging feedback on a paper I wrote for her class on Utilitarianism. This paper, a discussion of the relative justificatory
priority of utility and the virtues, was my first written attempt to articulate what eventually evolved into the topic of this dissertation. I don’t think that I have yet persuaded her to believe in “normative progeny”, but I am grateful to have been given multiple opportunities to try!

Conversations with Douglas Maclean made vivid to me the importance of my topic to interesting questions about the value of the natural environment. It turns out that much of the literature in environmental ethics struggles to show how moral justification (in this case, of the environment) is not undermining, but instead supportive of its objects. I thank Professor Maclean for his close reading of my writing sample, and for going over in fine detail various dubious or confusing claims of earlier drafts.

Bernard Boxill and Jan Boxill have been extraordinarily kind and helpful to me throughout my years in graduate school. Both have gone out of their way to advocate on my behalf in a variety of ways, especially as I have searched for an academic job in philosophy. I am grateful to them for their help. I must also thank Jesse Prinz for his efforts as placement director. He expressed indefatigable (and hopefully well-placed) confidence in my philosophical abilities, and maintained a seemingly unperturbable, and much appreciated, optimism that I would secure a fine job for myself in the end.

Finally, I need to thank my parents for their support through my seven years of graduate study. It is a bit remarkable to me that they were able to remain (at least outwardly) so encouraging and (again, at least outwardly) proud of me as I pursued one of the more impractical and monetarily unrewarding career paths open to college graduates. When times were tough financially for me—and, predictably, they often were—their generosity was extraordinary, unhesitating, and free of judgment. I can’t resist saying that through their
financial support, I had much more independence as a graduate student than I otherwise would have had. One wonders—and hopes—that morality could stand in a similar relationship of dependence with respect to its sources, sources that offer justificatory rather than monetary or emotional support. Can morality, in virtue of its (perhaps various) dependence relationships, be independent? What follows is an attempt to articulate and motivate this question, and to describe the logical space any successful answer must inhabit.
# TABLE OF CONTENTS

Chapter:

1. INHERITING NORMATIVITY

   I. Introduction
   
   II. Clarifying and motivating the model—first attempts
       A) Metaphysical worries
       B) Systematization
       C) Substantive theoretical commitments
       D) Theory-innocent intuitions

2. THE INDEPENDENCE PROBLEM EXPLAINED

   I. Complicating the model
   
   II. BP and PRR more closely considered
       A) The Bypass Problem
       B) Right Reasons
           i. normative independence
           ii. normative character

3. NORMATIVE PROGENY

   I. The Model of Normative Progeny: its design and function
      A) Normative Inheritance
         i. Normative Inheritance and the Bypass Problem

x
ii. Normative Inheritance and the ‘independence’ version of PRR……………………………………………………………42

B) Character Change……………………………………………………………………….42
i. Is character change ad hoc?...............................................................44

C) Preemption…………………………………………………………………………..45
i. Preemption and BP...............................................................46
ii. Preemption and the ‘character’ version of PRR........47
iii. Preemption and the ‘independence’ version of PRR…..48
iv. Preemption and ad hocness...............................48
v. Preemption and the preservation of consistency......49

D) Features of the Model Jointly Considered..........................50

II. A Promissory Argument.................................................................51

III. Initial Problems for the NP model........................................52

4. THE NP MODEL EMBODIED…………………………………………………………58
I. An intermediate level of abstraction................................................58
II. Taxonomy.................................................................59

A. Teleological, qualified source-commitment.......................63
   i. Rawlsian Rule-Utilitarianism........................................63

B. Teleological, source-committed indispensability.............72
   i. Hare’s two level theory..........................................72
   ii. Gauthier’s Contractarianism..................................74
   iii. Razian reasons and authority...............................83
   iv. Plato’s ‘goodness in itself’.................................87

C. Teleological, dispensability accounts..........................93
D. Non-hypothetical agent-election.................................96
E. Hypothetical agent-election........................................100

5. ANALYSIS.................................................................................104
   I. Overview and Methodology.............................................104
   II. NP accounts analyzed..................................................104
       A. Teleological, qualified source-commitment...............104
           i. Rawls’ Rule-Utilitarianism.................................104
       B. Teleological, source-committed indispensability.........111
           i. Hare’s two-level theory..................................111
           ii. Gauthier’s binding consequentialism...............113
           iii. Razian preemptive, dependent reasons............119
           iv. Plato’s ‘goodness in itself’.............................123
       C. Kagan-style teleological dispensability.....................124
       D. Non-hypothetical agent-election............................127
       F. Hypothetical agent-election....................................130
   III. Interpreting the results...............................................133
       A. Plato’s Principle..................................................133
       B. Preemption and Rigidity......................................138
       C. Character Change...............................................139
       D. Means and ends.................................................140
       E. Selectivity/promiscuity.......................................141
   IV. A final clarification....................................................142
   V. Conclusion...............................................................144
Chapter 1: Inheriting Normativity

“…value would seem always to be borrowed, and never owned; value would shine by a reflected glory....”

--W.D. Ross

“It may, therefore, be thought, that in the case of allegiance our moral obligation of duty will not cease, even tho’ the natural obligation of interest...has ceas’d...And indeed, to the force of this argument I so far submit, as to acknowledge, that general rules commonly extend beyond the principles, on which they are founded...”

--David Hume

I. Introduction

What is the status of actions, rules, principles, practices, virtues, states of affairs, and so on, whose normativity is “derivative”? Are derivatively normative things necessarily relegated to the status of, say, moral rules in act-utilitarianism? Rules for the act-utilitarian are famously only “rules of thumb”. These rules have no normativity of their own, so to speak, which is obvious from the way in which they are disregarded whenever they conflict with promoting utility. Is it possible, though, that moral rules and other things whose normativity is derivative could function differently, as though their normativity were underived? The suggestion sounds paradoxical, though we should hope this air of paradox

1 Ross (1988) 75. Ross here is rejecting subjectivism.
can be dispelled. For there is a vague but persistent demand that certain elements of morality must be derivatively normative, but nevertheless not “merely derivative” as this phrase is typically understood. This demand is stubborn, as I said, but stubborn with good reason: the defensibility of morality may depend up its satisfaction.

An example will help illustrate the demand I am thinking of, and the threat to morality posed by failing to satisfy it. Consider generosity: what makes generous actions good? If you’re a consequentialist, you’ll say that generous actions are good just because they promote good things. Generous actions might increase the welfare of their fortunate recipients, foster closeness between individuals, improve the character of the generous agent...and so on. What if generous actions never promoted these good things? If generosity were cut off from its usual effects in this way, it looks like it would never be a *good* thing to act generously. Acting generously in such a world not obviously be morally bad, but it would at least seem to be an imprudent waste of time. The goodness of generous acts can thus be thought of as derivative, or (the term I prefer) *dependent*—dependent, that is, on the teleological relationship generosity bears (in the real world) to things that are good.

Suppose one accepts this view of the goodness of generous actions. Are we thereby saddled with the conclusion that it is *never* good to act generously where a particular generous action is ineffective? This conclusion might seem unavoidable, but would be nevertheless unwelcome. Theory-innocent intuitions recognize a number of cases in which a generous but ineffective action ought to count as good. Suppose I donate a small amount of money to a relief organization, intending my donation to help victims of some recent disaster. But my check gets lost in the mail and so accomplishes nothing of any good for its intended

---

recipients. I don’t end up benefiting from the action either—I am financially a bit more stretched now, and annoyed that the relief organization hasn’t sent me a letter of acknowledgement. Yet for all that it seems that my act of generosity was in fact good, even if not as good as it might have been. The goodness of my generous act is not “merely derivative”, but seems good in some sense independent of its connection to other good things.

If all this is right, a tension has been revealed. There is an apparent incompatibility of the independently attractive claims that:

1) the goodness of generous actions is fully dependent upon the teleological relationship between generous actions and good results (I’ll call this the “dependence condition”); and
2) the goodness of generous actions is independent of this relationship (I’ll call this the “dependence condition”).

We might seem forced to choose between these individually compelling but evidently mutually exclusive views about the goodness of generosity. But the issue of compatibility between dependence and independence as I have expressed them is not obviously settled on conceptual grounds.

Consider a case in which dependence and independence can coexist in the same way that would be required to solve my moral case. Say my parents decided to give each of their children, on their tenth birthday, a part of their savings. So on my tenth birthday some of their money was transferred into my bank account. Given that my tenth birthday has passed, what is the status of this money in my bank account? It is clear I wouldn’t have the money if I were not my parents’ child. So in one sense my possession of the money depends upon my relationship to my parents. Nevertheless, now that I have received this inheritance my own
bank account balance cannot be affected by our parent/child relationship. If I were to officially “divorce” or disown them, or if they were to disown me, my financial balance would not reflect these unfortunate changes. The point is that my possession of the inherited money is in one way dependent on a parent/child relationship, but in another way independent of this relationship.

Let me be more explicit about the analogy I have in mind, returning to the example of generosity. The goodness generous acts promote is the analog of my parent’s money, and the goodness of generous acts themselves is the analog of my own money. The bearers of goodness—generous acts and the good things they promote—are analogous to me and my parents (the “bearers of money”), respectively. The hope is that generous acts could possess goodness in virtue of the (teleological) relationship they have to certain bearers of goodness, just as I could possess money in virtue of the relationship I have to my parents (bearers of money). Put differently, generous actions would depend upon the relationship they bear to the good things they promote, and nevertheless be good in some sense independently of this relationship. This normative analogue of financial inheritance is an example of what I will call ‘normative inheritance’.

Nothing about the example of financial inheritance is conceptually problematic or in any other way philosophically puzzling. This is encouraging, as it suggests that an analogue of inheritance in the moral context is at least not a conceptual impossibility. I think that if the analogy could be worked out in the right moral context, certain stubborn and important problems in moral philosophy—of which generosity is just one small instance—could be resolved. Put less positively, without a successful normative analogue to financial

---

3 I am aware that ethics does not have the market cornered on normativity, as it were, and that normativity is an important part of other areas of philosophy like epistemology and philosophy of language. But for simplicity’s
inheritance, morality itself may be literally indefensible—that is, incapable of being defended. We hope that morality can be defended as good in itself without by the very fact of that defense losing a legitimate claim to that status.

The problem I’m concerned with can be stated in somewhat formal terms. It is that of reconciling the apparently incompatible but individually compelling claims that:

1) the normativity of some x’s in morality are fully dependent upon the relationship between these x’s and some y’s that are also normative, and;

2) the normativity of these x’s is independent of this relationship.

What these claims mean is in obvious need of clarification, as is the claim that their apparent incompatibility poses a general problem for moral philosophy, one that goes beyond the case of generosity. Clarity on the first of these will not come easily, however, and the task of articulating more precisely the demand for the reconciliation of (1) and (2) will span a substantial part of the discussion.

An attempt at clarification and motivation of the problem, as well as a rough description of normative inheritance, will take up the first chapter.

In chapter two I give a closer analysis of the problem, which reveals itself to be surprisingly complicated. The ‘independence problem’ I discuss in the first chapter is seen to possess both extensional and (multiple) intensional components. These various components of the problem are described and motivated individually.

The more complicated problem presented in the second chapter may seem to call for a more complicated picture of normativity as well; at least, more complicated than what is offered in the first chapter. In the third chapter, then, a more elaborate picture of normativity

sake, ‘normativity’ will be shorthand for normativity in morality (unless I say otherwise).
is proposed, a picture that is naturally informed by the discussion of the second chapter. I propose two potential theoretical outgrowths of normative inheritance that appear promising in solving this more complicated problem of normative independence. Strikingly, these “theoretical outgrowths” I propose are not obviously ad hoc additions; they are not obviously jury rigged additions to the inheritance model, concocted simply to provide a happy result. Instead these features appear to flow naturally from the structure of the more basic model. I conclude this chapter with a promissory argument for a working version of my model.

The full-blooded version the problem I’m concerned with and the model that aims to solve it—which I rename ‘normative progeny’ for reasons that will become obvious—are then illustrated in the fourth chapter. Here I survey a number of substantively divergent moral views and moral theories that are victims of the same version of the independence problem. I also note that they employ roughly the same strategy in attempting to solve this problem. Specifically, though these strategies vary in ways that are sensitive to their substantive context, they employ the basic structure of the normative model I have developed. The theories I discuss here include sophisticated forms of act-utilitarianism, rule-utilitarianism, Kantian and Hobbesian versions of contractualism, a theory of legal authority (and its moral analogue), Plato’s account of justice in *The Republic*, and a non-contractualist version of Kantianism. The widespread appeal to the Normative Progeny model, and its obvious usefulness, are suggestive. Of course, the popularity and extraordinary usefulness of the abstract Normative Progeny model does not entail that it can find its concrete counterpart. One could interpret the evidence skeptically, and see the widespread need for normative inheritance or normative progeny as an argument against the possibility of moral justification. Still, its recognized promise to solve a deep problem for otherwise plausible
views about morality gives the model some presumptive plausibility.

Unfortunately, an analysis of the substantive views canvassed in the fourth chapter does not reward us with a successful version of the model. In the fifth chapter, I argue each of the examined attempts to employ the Normative Progeny model fail, and for reasons that reflect a misunderstanding of the strange conceptual terrain of moral justification. From these failures a set of conditions for success is generated, conditions that I argue any successful model must satisfy. This analysis clarifies the avenues along which a successful account must proceed. There does indeed appear to be logical space, I conclude, in which the model can be instantiated. The intuitions that drive the development of the normative inheritance model are strong, and so my analysis—which rules out several generic strategies for making use of the model—should not be discouraging. Certainly one of my aims in this dissertation is to expose a deep and widespread problem for morality; but by the end I also hope to have identified logical space for a way forward.

II. Clarifying and motivating the model—first attempts

Some terms that refer to “components” of the normative inheritance model should be defined for the sake of clarity and fluency. A ‘normative dependent’ depends for at least part of its normativity on its relationship to some further bearer of the normativity, where it would not have this relational normativity if not for the normativity of this further thing. A ‘normative source’ is this further “bearer of normativity”. Normative sources possess normativity irrespective of their relationship to their normative dependents and irrespective of the normativity of their dependents. This definition requires that any normative source

---

4 I include this rider to make it clear that the normativity of sources is not incidental to the normativity of its dependents, which has not been made explicit until now.
have a normative dependent, and vice versa. Now, a normative source might be intrinsically
normative, or it might not be. That is, a normative source might, in a Moorean way, be
normative strictly in virtue of its intrinsic properties; or it might be normative in virtue of its
relationship to some further normatively charged thing. This latter possibility means that
something is a ‘dependent’ on a ‘source’ not absolutely, but instead relative to particular
normative relationships. Some x might be the normative dependent of y, while y could be a
normative dependent of z, which in turn might be a normative dependent of z’...and so on.
(Whether this chain of justification must ultimately be anchored in something intrinsically
normative is fortunately not a question I need to take a stand on here).

I should clarify at the outset a possible source of confusion. I have used, and will
continue to use, the term ‘normative inheritance’, a term that in most contexts implies a
process that is temporally extended. But this is not what I have in mind here, and any
confusion on this point can plausibly be attributed to my use of what I hope are otherwise
helpful physical metaphors to describe a sometimes elusive abstract model. I do not at all
mean to suggest that I am attempting an historical account of the development of
normativity. I do wish to suggest, however, a notion of an order or priority, but the order I
have in mind is justificatory rather than temporal.

Normative inheritance requires that normative sources and their dependents relate in a
conceptually puzzling way. As I am thinking of it, some dependent x has inherited the
normativity of its source y if and only if x has at least some of its normativity in virtue of its
relationship to y, x would not have this relation-dependent normativity if it were not for the
normativity of y, and x’s normativity is, in some (undefined) sense, nevertheless independent
of its relationship to y. The seemingly paradoxical requirements for this cohabitation of

8
independence and dependence should be clear enough. How can something be dependent for its normativity on its relationship to some further thing, and at the same time possess this normativity independent of this relationship?

Despite its prima facie conceptual strangeness, the model of normative inheritance is attractive. A number of deep and important problems in ethics could be elegantly solved by a working version of this model. Its manner of solving these problems makes it not only attractive, but also gives it prima facie plausibility. I will consider three different problems faced by moral philosophers with different theoretical commitments that motivate the appeal of, if not the need for, a working model of normative inheritance. I then consider a fourth problem that is less theory-committed than the others discussed, as it can be motivated without appeal to any particular controversial theory about morality. I argue that a working version of normative inheritance could solve this fourth, general problem as well.

A) Metaphysical worries

Some might be attracted to normative inheritance because they see intrinsic normativity as metaphysically problematic, but nevertheless want to capture the sense that there are things that have a kind of independent normativity.

The tolerance for intrinsic normativity might vary here. One “intolerant” view would be that intrinsically normative entities should be eschewed altogether. A more liberal view would be that such entities need not be rejected wholesale, but should be kept to as small a number as possible out of respect for metaphysical parsimony. Normative inheritance could help with either view.

---

5 This sort of parsimony would be ‘quantitative’ rather than ‘qualitative’. That is, the metaphysical excesses it wishes to avoid are not merely new kinds of entities, but rather gratuitous tokens of (certain) kinds of entities.
Normative inheritance could satisfy the intolerant view by allowing normative dependents to be (in some sense) independently normative, yet not the source of their own normativity. To take an example, some writers on environmental philosophy are for metaphysical reasons skeptical of intrinsic value, but nevertheless hope to show that certain things, like the natural environment, are “valuable in themselves” in some satisfying sense. In earlier and well-known work on the natural environment Thomas E. Hill Jr. took care to avoid the claim that the natural environment was intrinsically valuable, and argued for the more modest conclusion that a virtuous person would value the natural environment for its own sake (independently of whether it is in fact valuable for its own sake). But even if he were right on this point, something about the value of the natural environment would nevertheless have been missed—namely, that sense that the environment really is valuable in itself in some stronger way. (Incidentally, Hill now acknowledges the importance of capturing this sense). A philosopher with metaphysical worries along the lines of Hill but with a similar sense of the independent value of some normative dependent or other would welcome a way to accommodate these otherwise dueling intuitions. And so they would welcome—or more strongly, are in need of—a way to work out normative inheritance, since according to this model normative dependents can be independently normative and yet not be the (metaphysical) source of their normativity.

The more liberal metaphysical view—that intrinsically normative entities need not be completely rejected but rather kept to a minimum--could make use of normative inheritance in much the same way. For instance, consider the egoistic principle that the only thing with intrinsic normativity is the satisfaction of one’s preferences. Such a view has obvious

---

6 See Hill (2005)
difficulty accommodating the commonsense view that there are plenty of actions one should perform that don’t satisfy one’s preferences. If normative inheritance could be worked out in this context, the normative range that is otherwise limited to the satisfaction of preferences might be extended so as to satisfy the broader and commonsense view of what one ought to do. Some morally just action, say, might fail to satisfy one’s preferences, but because of its relationship to preference-satisfaction—perhaps just actions tend to satisfy preferences—it could nevertheless be morally right. In this way, the view that satisfying one’s own preferences is the only thing that is intrinsically normative could be maintained, while accommodating the sense that there are certain rules that ought to be followed regardless of their consequences in a particular case (Note that Hobbes and neo-Hobbesians in fact argue tirelessly for this point).

_B) Systematization_

Another reason for developing a model of normative inheritance would stem not from a fear that intrinsic normativity is metaphysically problematic, but from an unrelated commitment to systematization in moral theory. System is a virtue of any theory, the thought might go, and is in any case a virtue of a moral theory. Or maybe the commitment to system is motivated by a deeper commitment still, such as a commitment to completeness in moral theory^8_. In any case, a thorough systematization of morality would require frugality at the ground level of moral justification, which amounts to a kind of reductionism. As a result, 

7 Hill (1983)

8 I am referring to the ability of a moral theory to pronounce an unambiguous verdict, at least in principle, on any moral question. Systematic moral theory, it could be (and has been) argued, is required for this kind of completeness. (Particularist intuitionism along the lines of Dancy’s might be thought a particularly egregious example of an incomplete moral “theory”). See, for instance, McKeever (2003).
everything besides those things at this bottom justificatory level needs to be justified by appeal to their relationship to these more basic things. But this brand of reductionist may be reluctant to lose the commonsense idea that plenty of less basic elements of morality than those they “privilege” have some kind of independent normativity—even if they are not intrinsically normative.

Unfortunately, a slender justificatory base cannot in a familiar or obvious way sustain a richer, more commonsense view. Unless he is willing to run roughshod over commonsense moral intuitions, a reductionist would be grateful for an account of normativity that allowed him to retain a slim justificatory base, but also allow things external to this base to inherit, and so possess in an independent way, normativity from this base.

C) Substantive theoretical commitments

A third reason one might be eager for a working version of normative inheritance would stem neither from metaphysical concerns about intrinsic value nor from a general commitment to reductionism in moral theory, but rather from an unrelated commitment to a particular substantive view that happens to be lean at the ground level of justification. Such a view will struggle to support commonsense views about morality.

For instance, one might be committed to a value theory familiar from hedonistic act-utilitarianism, namely the theory that pleasure and only pleasure has intrinsic value. Someone committed to this theory may nevertheless be troubled by its inability to accommodate a number of widely shared intuitions about morality, intuitions according to which regardless of consequences, certain things are morally normative. I have in mind intuitions like: promises are prima facie binding; lying is prima facie wrong; the natural
environment is valuable in itself, and not just because of the instrumental relationship these things bear to pleasure. Utilitarians struggle heroically to save these intuitions, and their resourcefulness to this end is remarkable. But these efforts are not ultimately convincing. If the normativity that attaches to pleasure could somehow be inherited by promises, lying, the natural environment, and so on, the hedonistic utilitarian account of value could be maintained without having to dismiss or explain away otherwise incompatible but strong moral intuitions. Rule-utilitarians, as we will see, offer one account of how normative inheritance might work within the framework of a hedonistic utilitarian view of value.

To take another example, one might be committed to a Kantian account of the scope of our moral duties, according to which we have direct duties only to rational beings because only rational beings are intrinsically normative (since only they are ‘ends in themselves’). But then what about our evident duties to preserve the natural environment, or to treat animals humanely? The Kantian strategy here is embarrassingly inadequate. Regarding our moral duty to treat animals humanely, Kant argues that we morally ought to treat animals humanely only because failing to so treat them will make us more likely to mistreat rational beings (like humans). Even if (implausibly) the alleged causal connection between mistreating animals and mistreating rational beings could be shown to hold in every case, the strong sense that mistreating animals is wrong in some way independent of our duties to rational beings is not at all captured by Kant. If, somehow, the normativity of certain non-rational beings could be inherited by the normativity of rational beings, the Kantian account of what is intrinsically normative might be maintained without losing the sense that our duties regarding some non-rational things are (in some sense) independent of the intrinsic normativity of rational beings. In later chapters I discuss how Christine Korsgaard and other
neo-Kantians offer a version of normative inheritance designed to resolve some of these concerns.

D) Theory-innocent intuitions

These three types of reasons might well be sufficient to motivate the need for a working version of normative inheritance. But each is more or less heavily committed to a particular theory about morality that plenty of apparently reasonable people reject (reasonably or not!). That is, plenty of apparently reasonable people find the concept of intrinsic value metaphysically unmysterious, or reject the proposal that systematization is a virtue of a moral theory, or have no particular substantive commitment that requires the help of normative inheritance. It would be nice, then, if we could locate a motivation for normative inheritance that does not depend upon controversial views about morality. I think there exists such a motivation, and because of its initially promising profile it will be the focus of the rest of this chapter.

There is a stubborn sense that, irrespective of any particular (controversial) theory about morality, certain things are obviously not the source of their own normativity. Instead, their normativity must depend on the normativity of and their relationship to other normatively charged things. But there is an equally theory-innocent--and equally stubborn--sense that many of these same things are nevertheless normatively independent.

A descent from this abstract characterization will be helpful. Reflection on the value of the non-sentient natural environment immediately raises this fourth version of the problem. (As should become clear, this is not a rehashing of the example described in (A) above, which turned on metaphysical worries not at issue in the present case). When
considered with theory-innocent eyes, the non-sentient environment does not appear to be the source of its own value. Environmental philosophers have a notoriously difficult time defending the claim that the non-sentient natural environment could be its own normative source. Some have argued that all living things are valuable in themselves, but this view faces devastating counterexamples. The AIDS virus, cancer, and the bacteria that caused the bubonic plague are all living, but it is hard to insist on their intrinsic value without blushing. Attempts to avoid this problem by drawing distinctions between non-sentient things in the natural environment that are intrinsically good and those that are not smack of ad hocness. Arguments for the usefulness of the non-sentient environment can only go so far, and in any case fall far short of a successful defense of the environment's independent value. We seem stuck with the view that the value of the non-sentient natural environment depends upon its relationship to sentient beings that can appreciate, value, or benefit from it. Nevertheless, to claim that the value of the non-sentient environment is “merely derivative” fails to capture the strong sense that it has value that is in some way independent of sentient beings. If the value of the non-sentient environment were explained in the same dismissive way act-utilitarianism, say, explains the derivative value of truth-telling or promise-keeping, its full value would not be accounted for. Many would welcome an account of normativity that would allow retention of the seemingly conflicting but, again, independently compelling views regarding the dependence and independence of the value of the non-sentient environment.

Moral rules face this fourth, “theory light” version of our problem as well. A rule that states something like “everything else being equal, one ought not break one’s promises” seems like it should be a normative stopping point in some manner. That is, if I am debating
whether or not to break a promise, the very fact that in breaking a promise I would violate the
moral rule proscribing promise-breaking seems like it should itself count as a consideration
(even if not an overriding one) that weighs into the moral calculus of the case. However,
even if I am not a consequentialist, I must grant that this rule has a point (such as to
maximize goodness, or respect rational beings, etc.). It seems I must further that if following
the rule didn’t generally serve its point there would be no reason to follow the rule. It seems,
then, like the normativity of such a rule is dependent on the normativity of and its
relationship to some further thing. In spite of this dependence, there is an enduring sense that
if the rule against promise-breaking does not, in a particular case, promote goodness or
respect rational beings or successfully serve some other point, its normativity in this case is
not wholly vitiated.

Rule-Utilitarians try to reconcile the conflicting intuitions about the normative
dependence and independence of moral rules with respect to consequences. Most rule-
utilitarians claim that the normativity of moral rules comes from the instrumental relationship
of certain rules to utility⁹. Following, promulgating, or internalizing certain rules (most of

---

⁹ With at least one prominent exception in Brad Hooker. See his (2003). Hooker’s theory will not concern us
directly, simply because it deliberately does not attempt to accommodate the competing intuitions about the
independence and dependence of morality. But Hooker’s rule-utilitarianism is fascinating as an example of a
theory that, instead of construing the demand for normative dependence and independence as a problematic but
unitary desideratum, views the demand as requiring a choice between normative dependence and independence.
Put simply, he chooses independence. He argues that various principles of morality are free-standing, justified
independently of some further end. (He does offer epistemic justification of his principles—their fit with
considered intuitions, for instance, tells in their favor—but normatively these principles are basic). Hooker
rejects the demand for normative justification, at least with respect to the basic principles of morality. In
digging in his heels and insisting on the plurality of self-standing principles, Hooker is of course criticized for
failing to provide these principles with a proper defense—and it is hard to deny that these criticisms are fair.

If Hooker chooses independence with respect to moral principles, the act-utilitarian chooses dependence.
Hooker denies that the principles of morality must be normatively justified; the act-utilitarian denies that the
principles of morality must be self-standing (with the solitary exception, of course, of the principle of utility).
The act-utilitarian must concede that keeping promises, telling the truth, being kind, and so on, are not good or
normative in themselves, but only derivatively normative.

Hooker and the classical act-utilitarian illustrate the alternatives to satisfying both aspects of the demand for
“dependent independence” in the context of utilitarianism. Both alternatives are problematic, and suffer for
their unwillingness to try to meet the demand in its entirety.
the time or all of the time), it is argued, will lead to good consequences. As a result, moral rules that meet the appropriate standard ought to be followed. It is granted, then, that if these rules did not in general lead to these consequences there would be no reason to follow them. So these rules are seen as dependent on their relationship to the good they usually promote. But in particular cases where following a rule does not promote good consequences, the rule still ought to be followed. The normativity of moral rules, then, are supposed to be in some sense independent of their relationship to good consequences, even though in another sense dependent on this relationship.

Unfortunately, versions of rule-utilitarianism that have the structure I’ve described are notoriously unstable and threaten incoherence. They threaten incoherence because, without further argument, there seems no way that rules could be dependent on yet independent of their relationship to good consequences. If the only thing of value is utility, how could moral rules—which are justified by utility, just like everything else that is justified—stand on their own, let alone compete with the principle of utility in cases of conflict? Many would be happy to bring together normative dependence and independence of rules in the way attempted by rule-utilitarianism. And for good reason. For then the apparently conflicting intuitions many of us share regarding the dependence and independence of moral rules could be saved.

Other examples are available. Certain moral virtues, and valuable works of art, or other products of human ingenuity also intuitively call for the problematic union of normative dependence and independence. This union would be effected through normative inheritance; it is not clear that union could be achieved in any other way. Note also that the importance of the problem I claim to have identified does not turn on any one example
successfully illustrating the demand for normative inheritance. If a particular example can be explained away, others are there to take its place. The call for moral justification cannot be ignored, even for those elements of morality that have an independent or self-standing quality; positing “self-justifying” or “intrinsically normative” entities is at best a last resort, at worst the construction of a fiction. And the demand for independently normative entities likewise must be satisfied; but, as I hope to have shown, the satisfaction of this demand sometimes requires that normativity be inherited.

In the next chapter, what I initially characterized as a “vague, persistent” demand for the union of normative dependence and independence will be further examined. Just what is it that we (or some of us, anyway) want when we demand the union of normative dependence and independence in the ways I have been gesturing at? The answer reveals the demand to be rather complex. Uncovering its complexity will lead to greater clarity on what an account of normativity sensitive to these demands would have to look like.
Chapter 2: The Independence Problem Explained

I. Complicating the Model

To sharpen our understanding of the ‘independence problem’ I have introduced, it will be useful to focus on one particular case in which this demand familiarly arises. I suggest we focus on moral rules, where it is granted for the sake of discussion that moral rules are normative dependents of the consequences they produce. (If this particular example does not move you, I suggest substituting, mutatis mutandis, another of the examples I have discussed). Critically pressing on the ‘independence problem’ causes it to splinter into two distinct primary components, one of which contains its own sub-components.

One of the reasons for wanting moral rules to be both dependent on and independent of the consequences that give them their normativity may already be clear:

1. If moral rules were “merely dependent” or (synonymously, for my purposes) “merely derivative” and so merely rules of thumb, they could be dispensed with or “bypassed” whenever doing so better promoted the normative consequences they are supposed to serve.

There is another, perhaps less obvious reason to seek this union of dependence and independence:
2. There is a sense that, for certain moral rules at least, one of the reasons for complying with them is *just that they are moral rules*. When asked, for instance, why one ought not break some promise or other, it seems fine to cite the good consequences that might come from keeping the promise. But even if these good consequences are sufficient to show in a given case—or even in every case—that one ought not break promises, it seems that there is yet another reason one should not break one’s promise. Namely, one should (in general) not break promises simply because a justified moral rule prohibits doing so. Such a reason for keeping promises appears underived or primitive, and in a way that would be spoiled by an immediate and exclusive appeal to the good consequences of following the rule. An account of the reasons to keep one’s promises that cannot accommodate this sense of the underived normativity of certain moral rules is incomplete in an important way.

The two problems just described are instances of two more general problems. The first, of which (1) above is an instance, is what I will call the ‘bypass problem’ (BP). The problem is this:

(BP) where some A is normatively dependent on B, it is too often permissible to ‘bypass’ or dispense with A and respond instead to the normativity of B alone (e.g. by promoting, respecting, or complying with B alone.)

The second general sort of problem, what I will call the ‘problem of right reasons,’ (PRR) is illustrated by the case described in (2) above. PRR will prove itself to have more than one aspect, but an early formulation of the problem will be helpful (even if incomplete):
(PRR) even where some dependent A is not subject to BP, we seem to be given the wrong reasons to respond to A, seemingly because A, in virtue of its normative dependence, is not normative in a free-standing way.

These two problems might be thought specific to instrumentalist moral theories and justifications, and my examples so far do not offer evidence to resist this suspicion. But, as I will argue below when I consider BP and PRR in more detail, “deontological” theories, just as they were shown vulnerable to the relatively unanalyzed ‘independence problem’, are also (unsurprisingly) not immune to a more nuanced version of the problem.

After fleshing out BP and PRR, I will explain in the next chapter how normative inheritance, if it could be worked out, could help solve these problems. I will also catalogue two additional features of moral approaches that try to make use of the normative inheritance

---

1 I should explain what I have in mind when I speak of "instrumentalist justification." An instrumentalist justification of morality attempts to show that some part of morality, such as rights, virtues, states of affairs, duties, and so on, should be (respectively) exercised, inculcated/exhibited, promoted, or fulfilled because doing so promotes some other thing of value, or enables us to avoid something of disvalue. Philosophers often speak of Kantian hypothetical imperatives as imperatives of instrumental rationality, though this is not what I have in mind by "instrumentalist justification." A familiar fact about hypothetical imperatives is that what distinguishes them from categorical imperatives is not their linguistic "if/then" form, but rather that they, unlike categorical imperatives, are not unconditionally binding upon us. An imperative can be stated "hypothetically", and for all that be categorical: "If you are going to avoid killing innocent human beings, you ought not overindulge in alcohol before driving." Because we are unconditionally required to refrain from killing innocent human beings, a Kantian would not consider this imperative to be one of "instrumental rationality." In Kantian parlance, in such an imperative reason is not merely an instrument (as it is in hypothetical imperatives), but rather sets the end of the action. However, as I am using "instrumentalist justification," the end towards which some instrumental good aims need not be a mere preference or desire; rather, the end could be intrinsically good, or the sort of thing whose promotion is unconditionally required. Unlike for the Kantian, what makes a defense "instrumental" for this discussion has not to do with the unconditionality of the ends, but rather that it enjoins us do something as a means to bring about some other end.

Crucially, what makes a justification of morality instrumental is relative to its context--specifically, relative to the feature of morality it is called upon to justify. For example, though act-utilitarianism gives an instrumentalist defense of truth-telling (e.g. "one ought to tell the truth most of the time because doing so will maximize happiness over the long run") it does not offer an instrumentalist justification of pleasure. Pleasure is taken by the theory to be basic, and so cannot receive an instrumentalist defense. So, a defense of morality is instrumental relative to some feature F of morality where F is shown to be useful in bringing about some desired or desirable end E, and where E is not the same as F.

2 The overwhelming focus in the literature on consequentialism’s vulnerability to one or both of these problems suggests, at least, a failure to appreciate the vulnerability of other sorts of theories to these problems. See, for instance, Prichard (1912), Ross (1988), Rawls (1955), Scheffler (1984), all of whom finger instrumentalist theories alone as having something like the susceptibility I have been discussing.
model, features that seem designed to solve the frequently “co morbid” extensional and intensional problems of BP and PRR. As I will try to show, despite appearances these features are not obviously *ad hoc* theoretical salvage jobs, but plausibly flow from the basic structure of normative inheritance.

II. BP and PRR more closely considered

A. The Bypass Problem

BP is a worry about a theory’s *extension*. The worry is not about how a theory picks out the things it judges right and wrong or good and bad, but only with the sorts of things that are picked out as right and wrong or good and bad. BP threatens wherever there is a derivatively normative entity that ought to be treated as though its normativity were *underived*.

The most general sort of bypass problem--of which every other bypass problem is a species--is simply the problem of immorality: in many circumstances derivative justifications of morality permit us, require us, or otherwise provide reason to be immoral.

A familiar example of the problem of immorality is what has come to be called the ‘problem of injustice’. Perhaps the most famous versions of the problem of injustice are put forward by Plato’s Lydian Shepherd, Hobbe’s Foole, and Hume’s Sensible Knave. Though these problems are different in some of their details, they share a common basic structure. This version of the problem of injustice arises on egoistic defenses of morality, where the rationale for acting justly is the advantage it will bring to the individual agent. The thought of both the Humean knave and Hobbes’ Foole is that only by acting justly can one make oneself fit to partake in, and benefit from, the mutually advantageous convention of justice.

---

3See Gauthier (1990).
(The Lydian Shepherd, once he has acquired the ring of invisibility, does *not* act justly of course. But his immoral behavior [and its apparent reasonableness] following the acquisition of his great power is part of an argument for justice based on individual advantage.) If one acts unjustly, the argument goes, inevitably one will earn the reputation for being unjust, and will be excluded from mutually beneficial interactions. So one had better act justly. The perennial problem with argument is that even if acting justly is often or usually productive of an individual's good, it may not *always* be so productive; what if one can act unjustly without anyone finding out about it, thereby keeping one's reputation for being just intact? Or what if one is powerful enough to act unjustly without suffering the usual consequences? Put in the terms I have been using, the worry is that there will arise many instances in which it will make sense--given the derivative status of justice--to bypass its requirements in favor of promoting one's own good directly.

Plato, Hobbes, and Hume posed the challenge of injustice against an egoistic, subjectivist defense of justice. A parallel problem of injustice can arise, though, also where the reason given for being just is agent-neutral. For instance, the problem of injustice is a notorious stumbling block for classical act-utilitarianism. A well-worn act-utilitarian strategy is to argue that justice and utility will coincide in the long run: acting justly in a given case is more likely to produce the best consequences than acting unjustly. However, various compelling counterexamples to this claim can be generated. For instance, it is certainly unjust to kill one relatively healthy patient and harvest his organs in order to save the lives of five others, or to push a corpulent gentleman in front of a trolley in order to derail it and save one hundred others--however, we can imagine that these actions could be "optimific". To give less fanciful sorts of cases, act-utilitarianism seems like it could be used
to justify discrimination based on race, sex, or class, and perhaps could even be used to justify slavery, since, despite obvious negative consequences, systematic sorts of oppression might produce better results on the whole than any other social arrangement. Here again, the problem is usefully understood as a bypass problem. For act-utilitarianism, too often it makes sense to bypass the merely derivative rules of justice in favor of promoting general utility.

Again, these are all examples of teleological accounts of justification, which to some would appear no accident. The thought would be that teleological theories are uniquely susceptible to BP in virtue of the status that an instrumentalist mode of justification confers upon their objects of justification. Since normative dependents there are so often more effective ways of responding to the normativity of a given instrumentalist source than by taking the prescribed channels, much of morality can apparently be bypassed if doing so will just as effectively or more effectively promote this source. But it is not clear, the thought runs, how this same problem presents itself for deontological theories.

The first thing to note is that, whenever some x is defended as normative because of its relationship to some further justifying ground y, x is derivatively normative. An act that is right just in virtue of being permitted or required by a deontological principle, then, would be a derivatively normative act. Such an act derives its normativity from its controlling principle. So instrumentalist justifications are not alone in their capacity to bestow a derivatively normative status on their objects. Still, perhaps only instrumentally “derivative” objects are vulnerable to BP. Can deontological justification block the bypass threat?

Like (many) consequentialist theories, many deontological theories would like to save our commonsense moral intuitions as much as possible. Sometimes this is done in a
'deontological' way: morality, or some part of morality, is justified because it is elected or chosen by an appropriately situated moral agent. But it is not clear how this sort of election is supposed to confer a robustly independent sort of normativity upon its objects such that these objects cannot be bypassed. Such value “conferral”, as discussed by, for instance, Christine Korsgaard⁴, appears vulnerable to the bypass threat. If the source of the value of conferred objects is some moral agent, why not just respond to her normativity (whatever that may require) instead of the objects she has chosen? Whatever value an elected object might have on a deontological view, it is likely to have less value, or more generally less normativity, than its source. Kant explicitly claims, for instance, that rational nature is never to be sacrificed in favor of the objects of rational choice⁵; and many deontological theorists would make a similar claim.

Of course, another option for deontological theories is to employ instrumentalist justification in some limited role, but this just puts deontologists in the situation of their teleologically-minded counterparts. Kant’s previously mentioned defense of the humane treatment of animals, where we are not to treat animals inhumanely because doing so increases the probability that we will treat (rational) human beings badly, is a classic case in point. The alleged connection between inhumane treatment of animals and humans is purely contingent, and unlikely to hold in every particular case. And even if it did, a more effective way of treating humans humanely might—who knows?—require the brutal treatment of animals. Where treating animals inhumanely does not dispose us toward cruelty toward human beings, why not bypass the general injunction against inhumane animal treatment?

⁴ See Chapters 4 and 5 for a fuller discussion of Korsgaard’s view.

⁵ Kant (1998) 4:397.
B. Right Reasons

i. normative independence

Unlike BP, PRR is a problem concerning the intension of moral justification—how they pick out what counts as right and wrong or good and bad as opposed to what sorts of acts, policies, states of affairs and so on get counted as such.6

I have claimed that PRR is the problem that derivative justifications of morality give us the wrong reasons for acting morally. But when formulated in this way, there remains an important ambiguity. By “reasons” I might have in mind either ‘normative reasons’ or ‘motivating reasons’.7 A ‘normative reason for some action (say) is a reason that justifies, or goes some way in justifying, that action, whether or not we are in fact moved by that reason. A motivating reason for some action, on the other hand, need only serve as the volitional principle upon which we act, and so explains why that action was performed—whether or not it helps to justify that act.

Modern moral theories (like utilitarianism and Kantianism) have been indicted for providing us with the wrong sort of motivating reasons.8 It is argued that to be motivated in accordance with what many modern theories take to be genuine reasons alienates us from our attachments and projects, making, for instance, true friendship and true love impossible. For to truly love or have a friend, requires that we value the object of our affection for her own sake, rather than impartially, where they are viewed only as one part of a giant moral matrix.

6 Some might contend that a theory’s intension is right if and only if its extension is right, and that the latter determines the former. On this view, a moral theory’s intension does not matter in its own right. This is a tough position to maintain, however, and the reluctance of philosophers to embrace it is not a presumption in its favor. I have in mind the myriad attempts by moral theorists who are apparently committed to the derivative status of intension, but who nevertheless struggle against this commitment. A small sample of these “myriad attempts” will be discussed in chapters 4 and 5.

7 Michael Smith makes this useful distinction in his (1994) 95.

8 Williams (1973, 1981) at points; also Stocker (1976).
To escape this problem, it is said, one could try to trick oneself into caring about friends and lovers for their own sake by erecting a kind of cognitive firewall between one's normative reasons and one's motives. One would train oneself to care about others for their own sake, even though one knows, in some insulated part of one's mind, that these friends and lovers are not really the sorts of things that are desirable for their own sake. But this solution—which requires a kind of moral schizophrenia—9—is argued to be unacceptable. And since so is the alternative of friendlessness, modern, impartial moral theories, the argument concludes, cannot be right.

However, the problem of “right reasons” I am concerned with is not that of motivating reasons, but rather of normative reasons. Normative and motivating reasons are conceptually distinct, and can pull apart in practice as well. Though one can imagine that an incorrect account of normative reasons would also provide its adherents with the wrong motivating reasons, a theory that struggles with providing the right motivating reasons could be perfectly correct in its account of normative reasons. It is important, then, to maintain this conceptual distinction.

To see why wrong motivational reasons do not entail wrong normative reasons, consider an example. An instrumentalist moral theory, when used as a procedure for making a moral decision—and so when motivating moral thought and action—may be morally disastrous. This is a familiar point to utilitarians. Mill, for instance, was aware of the problem10, but also realizes that the problems with embodying utilitarian motives does not in

---

9 Stocker’s felicitous term, as used in his justly celebrated article (1976).

10 In Chapter 2 of Utilitarianism Mill considers arguments against utilitarianism that were based on the claim that utilitarianism provides the wrong motivating reasons. Examples include the charge that utilitarianism is wrong because it is impractically time-consuming, or requires access to facts to which we cannot possibly have access.
any way impugn utilitarianism as an account of normative reasons (though he did not speak in quite these terms). His practical recommendation was to cultivate the love of things other than happiness (such as virtue), since in this way we would be most likely to maximize utility. And this practical approach seems like the right sort of response. As a criterion of moral rightness (in other words, as an account of the right normative reasons), a moral theory is not obviously undermined by a demonstration that, as a decision procedure, it would lead to disaster. Darwin’s theory of natural selection is no less plausible as a biological theory because it makes us less cheerful than a creationist account, and it is hard to see why in the case of ethics things should be any different.

I will give some examples of intensional problems for moral justification, examples whose intensional shortcomings should be fairly uncontroversial. These examples will provide the “data”, as it were, for developing a more nuanced account of PRR.

Allow me to revisit Kant’s defense of why we should treat non-human animals with respect. Now, there are surely some good instrumentalist reasons for not abusing animals, even on Kant’s focus on rationality. Perhaps, as Kant says, we ought to treat animals well

---

11 Mill (1979) writes: "Whatever may be the opinion of utilitarian moralists as to the original conditions by which virtue is made virtue, however they may believe (as they do) that actions and dispositions are only virtuous because they promote another end than virtue, yet this being granted, and it having been decided, from considerations of this description, what is virtuous, they not only place virtue at the very head of the things which are good as a means to the ultimate end, but they also recognize as a psychological fact the possibility of its being, to the individual, a good in itself, even although, in the individual instance, it should not produce those other desirable consequences which it tends to produce, and on account of which it is held to be virtue. This opinion is not, in the smallest degree, a departure from the happiness principle" (35, emphasis added).

12 One might think that the problem with moral theories whose normative and motivating reasons naturally come apart is not that they lead to disaster, but that they lead to a particular self-effacing sort of disaster. If one objects on these grounds to my example of Darwin’s theory, consider a variation on the example. Suppose one is an archeologist and one’s most productive field assistant believes in creationism. Her zeal to vindicate her creationistic beliefs is part of what makes her such an important contributor to your team: she works longer hours than anyone else, excavates with particularly great care and efficiency, etc. Suppose that if she were to be convinced of the truth of Darwinism, her productivity would decline as a result of her deep disappointment. If we assume that the point of Darwin’s theory is to shed further light on the natural world, then her holding Darwinism to be true would be antithetical to this point--Darwinism (in this instance, at any rate) would be self-effacing. But this says nothing, of course, about the truth of Darwinism.
because mistreating animals increases the likelihood of mistreating humans, who are rational beings. Suppose this response as a matter of fact provided us with extensional results that were consistent with our actual duties to animals—in other words, suppose it is not susceptible to BP. Even granting this, an important set of reasons against mistreating animals seems to have eluded this justification. The pain and suffering of animals seems like it should count in its own right as a reason against their mistreatment.

Instrumentalist defenses of environmentalism provide another nice example of intensional incompleteness. A utilitarian can give us compelling reasons for preserving the environment: nature provides us with aesthetic pleasure; plants are needed to replenish the atmosphere; the rainforests contain plants and animals that could be very useful for fighting disease; and so on. But even if we can be convinced that we ought to preserve the environment for these reasons, we reasonably hope that there is an additional reason to preserve the environment. Namely, that that preserving the environment was important in its own right.

So far I have only considered PRR as it applies to instrumentalist defenses of morality. Even the Kantian example focused on an instrumentalist defense within Kant’s theory. But the strictly “deontological” aspects of moral theories are vulnerable to PRR, just as they are vulnerable to BP. Consider a Kantian defense of our duty to promote the happiness of some rational person r. According to Kant, we are required to treat humanity always as an end. This means that we are required to promote or foster the (rational) ends rational beings. Each rational being, according to Kant, has his or her own happiness as a (presumably rational) end. Let us grant that Kant is right about this. Let us also grant that Kant is right that r’s having happiness as a (rational) end provides a reason to promote r’s

---

happiness. So, if we buy Kant’s premises, we have a reason to promote r’s happiness. But is that the end of the story? Surely there is more to say here. There seems to be another reason we should promote r’s happiness that has been missed: namely, that happiness—or at least the happiness of a good person—is a good thing in its own right!

These examples suggest a diagnosis. In each example, the relevant part of morality being justified or defended was normative in a dependent way. The normativity of each object of justification depended, that is, upon something beyond (or other than) itself. When we think that some x is independently normative of some y, we are plausibly claiming that x is normative independently of its relationship to y. If truth-telling is good independently of its consequences, this seems to mean that truth-telling is good whatever consequences might come from it. But in each of the examples considered, the objects of justification were clearly not normative independent of their justifying ground.

ii. Normative character

Some derivative justifications are susceptible to the problem of right reasons along more than one dimension. The version of the problem I have just discussed I will call the ‘independence’ version of PRR. But there is another version as well. This is because derivative justifications threaten not only to rob the normative independence of their objects, but also because they often fail to offer justification of the right ‘normative character’.

I need to explain what I have in mind by ‘normative character’. When we ask of some principle, say, “Is that a moral principle? Or is that a principle of etiquette? Or is it just a principle of prudence?” we are asking for what I’m calling the character of the principle. Another, maybe somewhat more illuminating way of describing what I mean:
when we say, critically, of some reason to do such and such “well, that’s not a moral reason to do such and such” we are objecting to the character of the reason. Most straightforwardly: the normative character of some x is the kind of reason there is for conforming with, promoting, avoiding, minimizing, or otherwise responding to x (where ‘kinds’ are restricted to those of prudence, morality, etiquette, aesthetics, rationality, legality, and other domains of normativity).

Some examples might help clarify what I have in mind. Intuitively—and here I will need to lean heavily upon common intuitions—a principle that requires we always treat others with respect has a moral character; a principle that requires we always take the most efficient means to our ends is apparently prudential in character; a principle that requires dressing according to the formality of the event we are attending is a principle of etiquette, and as such has its own distinctive normative character.

Various things can have a normative character, simply because there are many things to which one can have a reason—of one kind or another—to respond. Virtues can vary in their normative character, for instance, as can rights, duties, good and bad states of affairs, and so on. Most generally, then, we might say that reasons can have different normative characters.

The distinctions between normative characters are difficult to articulate at an abstract level, and there is no philosophical consensus regarding the full set of conditions a reason must meet to be a moral reason, say, or a prudential reason.

One might resist this skepticism, believing that the hallmark of a moral reason is its impartiality. Moral impartiality is complicated, and a moral theory or view is impartial only

---

14 That is, as long as one accepts the popular view that all talk of normativity can be translated, without loss, into talk of reasons. See, for instance, Raz (1999) 67, one of the view’s leading proponents.
relative to the various considerations that it refuses to count as morally relevant. One theory might be impartial with respect to merit or desert, while another might be “partial” with respect to merit/desert but impartial with respect to gender and race. Utilitarianism achieves its own severe sort of impartiality through indifference to every factor besides the net sum of happiness.

Though impartiality might differentiate the moral from the prudential, it seemingly cannot differentiate between the moral and the rational. Rationality appears binding in an impartial way that mimics morality’s supposed impartiality. The universally binding nature of rationality gives it its impartiality—the norms of rationality do not “play favorites” and apply to some people or groups but not others.

One might attempt to distinguish normative “kinds” by appealing to categoricity: in cases of conflict, moral reasons, they are widely presumed to trump all other sorts of reasons. Prudential reasons, on the other hand, are almost never considered categorical. Further, prudential reasons, if they can rightly be called impartial at all, are not impartial in the manner of moral reasons—they are explicitly self-regarding rather than other-regarding. But even morality’s categoricity can be called into question when considering certain cases. Suppose I have the option to save my wife or the lives of five strangers (but not both). Morality would seem to require that I save the five strangers, but reasons that apparently stem from my personal relationship with my wife appear capable of trumping morality here, and permit—even if not require—saving her alone.

The deontological/teleological divide seems another promising place to look for demarcating the moral from the non-moral. Interestingly, however, the normative character of a thing seems to be separate from, and not obviously determined by, its status as
teleologically or deontologically justified. For consequentialist and deontological principles alike appear to admit of normativity of various characters\textsuperscript{15}. I point this out primarily because a naïve view of moral justification relegates consequentialism to the non-moral (even if expedient and practical), and where deontology is held up as moral (even if overly rigid). Every part of this stereotype is probably unfair, but I will focus on the claim that deontological justification is inherently moral, and consequentialist justification inherently non-moral.

Consider the principle that one ought always to maximize the state of affairs in which persons are treated with respect. This is a teleological principle, but it has a fairly uncontroversially moral character. Why do I say this? Simply because the reason for following the principle is itself apparently moral: treating persons with respect is considered valuable, and as such ought to be maximized. Consider alternatively the principle that one ought to act in a way that would be allowed by principles agreed upon by rational beings aiming to construct principles of polite conduct. This is evidently a deontological principle of etiquette. Deontology and Consequentialism do not carry their normative character with them, as it were.

If the deontological or consequentialist ground of a reason does not, by itself, determine a reason’s normative character, then what does? Consider yet another alternative: perhaps extension, and only extension, determines intensional character. If a reason requires prudence, then this reason has a prudential character—or so the thought would go. Likewise, if a reason required morality, then this reason has a moral character. But does this

\footnote{\textit{Pace} etiquette columnist Miss Manners, evidently. In a recent column she suggests that the rules of etiquette do not admit of a consequentialist justification: “Etiquette rules are supposed to be followed because that is the right and expected thing to do, not because of proven consequences in disobeying them.” (http://women.msn.com/1213436.armx)}
make sense in light of how normative character has been defined? If the normative character of some x is whatever kind of reason there is for responding to x, then the extensional view being considered here cannot work. To see why, consider being told that you ought to treat certain guests to your home with the greatest politeness you can muster. So far, the reasons for being polite—on the extensional view, at any rate—appear grounded by considerations of etiquette alone. However, in a private moment you are informed that if you are not polite to your guests they will kidnap and kill you—this is the reason you should be polite. Prudential considerations, then, ground your polite actions. As such, your actions, though superbly polite in their extension, are solely prudential in their character.

Given the absence of attractive alternatives, I suggest a different account of normative character, one that is sparse in its details but hard to deny at the structural level. Very simply, I propose that the normative character of a normative source determines the character of its dependents. Since the normative character of a thing is the kind of reason one has for responding to it in various ways, and since normative sources provide reasons for responding to their normative dependents, the normative character of normative sources presumably determine the normative character of their dependents. So, for instance, on this view the character of any particular justified act within act-utilitarianism will be determined by the character of the principle of utility that justifies it (or the principle that, in other words, gives one reason to perform the act).

If I am right, we can see why the problem of ‘normative character’ arises in the context of normative inheritance. It appears that a normative source could, in virtue of its own normative character, infect its dependent with its own character, and so provide the wrong kind of reason for responding to its dependent. If the default view of what gives
something its normative character is right—and we should assume that it is—then this second version of PRR threatens in a fairly obvious way.

The ‘normative character’ version of PRR will appear especially dangerous for a moral approach that attempts to ground morality or some aspect of it in terms of non-moral considerations. Such justifications are desirable, according to their proponents, because they can provide non-question-begging reasons for acting morally to the moral skeptic. This ambition is admirable, but one that creates a special vulnerability to PRR. Hobbesian contractarianism, for instance, faces the character version of PRR. The normative source of the Hobbesian rules of justice are prudential—if Hobbes is right, the good for each person is his or her self-preservation, and to serve this good self-interested individuals should obey the rules of justice (provided a sovereign punishes those who disobey these rules). The good of self-preservation is the normative ground of principles of justice that require, for instance, the keeping of covenants (or promises). But a prudential reason such as that suggested by Hobbes looks to be of the wrong character to ground the moral obligation to keep promises—even if Hobbes could allay extensional concerns by showing that we ought to always keep our promises.

One might also worry that a moral theory that is grounded on the normativity of rationality would provide justification of the wrong character for being moral. In rationality philosophers often seek a ground of morality that is universal and categorical, but one wonders if rationality, even if extensionally adequate, is of the wrong character to ground morality. One might think that the answer to any moral question such as “why should I tell the truth?” must be answered at least in part by appeal to the morality of telling the truth, not, or not just, its rationality.
Interestingly, the character version of PRR may also apply to moral theories that take the justifying ground of morality to have a *moral* character, such as familiar deontological theories. Such theories can give an overly moralistic justification of certain of their objects. There are some things that clearly have value where this value is nevertheless not naturally described as moral in character. Pleasure, for instance, is valuable, but it is not obviously *morally* valuable, at least not primarily. And music is likewise valuable, but not obviously for *moral* reasons, or, again, at least not primarily for moral reasons. Worries about the separate normative character of rationality aside, Kant argues that we ought to promote happiness since the want of happiness might engender a dissatisfaction that would lead to the transgression of duty. Kant may have more to say about reasons to promote happiness. But if this were all he could say, his arid account of the reasons for promoting happiness would ascribe to happiness the wrong normative character. A moral theory that justifies its objects in terms of their conformity with some moral or moralistic principle, then, can infuse these objects with a moralism that misdescribes the character of their normativity.

Now that we've formulated PRR a bit more clearly, we can appreciate its depth. The bypass problem is formidable, but at least we can without much difficulty *conceptualize* what a satisfactory solution might look like. For instance, if, however implausibly, an act-utilitarian succeeded in showing that in order to maximize happiness it *never* made sense to be unjust, this would count as a utilitarian solution to BP. Or if the Kantian could show that human beings as a matter of fact rationally choose only moral ends, and that in so choosing provide others with reason to promote these ends, then BP could likewise count as solved for the Kantian. But it is hard to similarly conceptualize what a solution to PRR, on either of its versions, might look like. PRR is seemingly brought upon morality by the very attempt at its
defense. Once morality is defended, such that some part of it that we would like to be independently normative is defended as normative because of something normative beyond itself, it looks like the independence condition is ipso facto unsatisfiable.

There is likewise no easily imaginable solution to the character version of PRR. One thought would be to abandon non-moral defenses of morality altogether, and concede that such an attempt to ground morality rests of a mistake; the proper reasons for acting morally cannot be given by non-moral considerations. But to abandon this project is to abandon the hope that we can supply the moral skeptic with the appropriate sorts of reasons for acting morally, and also to give up on a number of otherwise attractive moral theories (such as Hobbesianism, various neo-Hobbesianisms, and various versions of utilitarianism). And in any case such a retreat may not be enough. As I have already suggested, a defense of morality that takes its justifying grounds to have a moral character might lend out its moral character too liberally, giving certain valuable goods an unnaturally stilted, moralistic character.
Chapter 3: Normative Progeny

With BP and PRR explicated in some detail, I am in a position to articulate a more elaborate version of Normative Inheritance that responds to these problems. Once the model has been articulated, I will offer a promissory argument for the viability of the model and consider initial objections.

There are two additional characteristic features of the moral views that respond to BP and PRR, and that embody the more elaborate version of Normative Inheritance I am developing. (Perhaps “developing” here is misleading, since I generate the model only by uncovering various features of different theories that attempt to address BP and PRR, and highlight their common structure. A wealth of appealing but substantively divergent moral views of justification reveals striking structural similarities that display a commitment to this model). The model I’m articulating builds on the discussion of normative inheritance\(^1\), and includes it as a central aspect. I will call this expanded version a model of ‘normative progeny’ (NP), a model according to which certain normative dependents take on a normative “life of their own”. I refer to these dependents as the ‘progeny’ of their sources.

Normative Inheritance by itself could go a long way in solving BP and PRR. I will discuss how below. But in many cases (that is, in many moral circumstances, and in the context of many different normative theories) normative inheritance needs to be supplemented in order to handle BP and PRR adequately. These additional features I will

\(^1\) Cf. Ch. 1
call ‘character change’, and ‘preemption’. These features do not always accompany
Normative Inheritance, though perhaps, as I will suggest, they should. Their additional
presence might be warranted simply by their function; though possibly these features are a
natural—and even unavoidable—outgrowth of the structure of Normative Inheritance itself.
I will develop these possibilities below as I describe what I have in mind by ‘character
change’ and ‘preemption’. Later in the chapter I offer a promissory argument in favor of a
working version of Normative Progeny (the name I give to this more complicated version of
Normative Inheritance), and finally consider objections to the model.

Some of the ground covered in the next sections will inevitably have a familiar feel to it; the unanalyzed ‘independence problem’ has been discussed, and the uncomplicated
Normative Inheritance model has been argued to show promise in solving the problem.
Since the ‘independence problem’ is essentially an ill-articulated version of BP and PRR, and
since and Normative Inheritance model is central to the Normative Progeny model, part of
the NP model has already been applied to vaguely characterized problems that properly fall
under the scope of BP and PRR. But it is worthwhile to see how remarkably well each
aspect of NP model appears to deal with the various elements of the ‘independence problem’
once it has been unpacked.

I. The Model of Normative Progeny: its design and function

A. Normative Inheritance

i. Normative Inheritance and the Bypass Problem
The logic of Normative Inheritance has already been described\(^2\). And its usefulness in satisfying the “vague, persistent” demand for normative independence and dependence persistent has been discussed. Just how is Normative Inheritance supposed to solve this demand, though, now that the vagueness of this demand has been analyzed away and its persistence made even more evident? Just how, in other words, is normative inheritance supposed to help with BP and the two versions of PRR?

On the most familiar sorts of moral theories that invoke derivative justification, normative dependents are considered “merely derivative”. This usually entails—together with certain facts about the actual world—that in some circumstances it makes sense to bypass derivative dependents in favor of responding to their normative source. Dependents for the act-utilitarian, for example, should be responded to as normative only when they bear the proper relationship to their normative source. And at least in a number of familiar theories—including AU—the “proper relationship” does not hold in every case in which, intuitively, the dependent \textit{ought not} be bypassed. This seems to make sense: since dependents receive their “funding”, so to speak, from a source \textit{beyond} themselves, why think their normative source stands in the right relation to its dependents to avoid objectionable bypass? Though it plausibly makes sense to lie on some rare occasions, AU permits lying on \textit{too} many occasions—and this is simply because lying is often optimific. The model of normativity I’m articulating aims to fix this problem, not only for AU but for other theories subject to BP. It aims to do this by allowing normative dependents to behave as though they were normatively \textit{independent}, which would seem to require that their normativity would not depend upon an unreliable connection to some normative source.

Rule-utilitarianism (RU) provides an example of how the Normative Inheritance is

\(^{2}\) Ch. 1, p 7
supposed to function with respect to BP in a consequentialist context. According to RU, which should by now be familiar, though a set of rules is normatively dependent upon the principle of utility, these rules *themselves* seem to provide reasons for action. Just like normatively basic or primitive principles, justified rules are normative even when they are not conducive to utility in a particular instance. A rule prohibiting lying might be justified because of its general utility-conduciveness, and yet these rules may not permissibly be bypassed as mere “rules of thumb” on RU when following them does not maximize utility. It is as though the robust sort of normativity that is possessed by the principle of utility itself—the normative source of justified rules—has been inherited by the moral rule that prohibits lying.

To take another example, some highly utility-conducive rule that proscribes murdering innocents might on AU not make sense on some occasion. Perhaps the situation is like the one Dostoyevski’s Raskolnikov believed he perceived: the victim of the murder has no friends, is herself miserable, and her death will profit others. The permissibility of such a murder on AU is an embarrassment for the theory. The rule-utilitarian can relieve this embarrassment by applying a version of normative inheritance. RU can admit that the wrongness of murder depends, and depends entirely, upon its relationship to bad consequences. If it were not the case that murder reliably results in horrific states of affairs (the horror of which could, on a sufficiently generous account of value, include the state of affairs in which an innocent person is murdered) it would not be wrong. While admitting this dependence, though, RU can judge a utility-maximizing murder to be wrong, as though the rule against murdering innocent persons were free-standing or self-justified rather than dependent upon its relationship to the principle of utility.
Generally, normative inheritance could help to solve BP by giving a certain robust, seemingly independent normativity to features of morality that should not be cast aside as often as a familiar understanding of derivative normativity would permit.

ii. Normative Inheritance and the ‘independence version’ of PRR

If normativity could be inherited as I have been describing, a now (hopefully) familiar aspect of PRR could likewise be dealt with. This is the problem that, even if a theory succeeds in *extensionally* defending morality or some part of it using derivative justification (and so solves BP), certain things seem like they should have the *intensional* status that is typically thought of as the exclusive domain of things independently normative. (This aspect of PRR should be the most familiar of the various aspects of the “vague” demand for independence/independence, as it has figured prominently in the examples I have discussed [such as the example of generosity in the first chapter, and the discussion of the environment’s value in that same chapter]).

The sense in which normative dependents must be independent of their source in order to satisfy this aspect of PRR is still, however, not clear. Greater clarity on the demand will come only through an analysis of attempts to satisfy it. Certain approaches will be ruled out by their failure, and their failure will help reveal the conditions for success.

B. Character Change

Fascinatingly, though normative inheritance appears unfriendly to the suggestion, many theories that attempt to make use of the model claim that the *character* of the normativity possessed by dependents is different than that of their normative source.
Normativity is supposed to somehow change its character—not over time, of course, but across justificatory generations.

If this were possible, the ‘character version’ of PRR might admit of a solution. This would be great news for many. The Hobbesian, for instance, would be delighted if character change could be worked out in the context of his theory: he would be able to grant the non-moral character of “morality’s” source, while insisting upon the moral character of Hobbesian moral principles. For according to Character Change, a source can serve as the sole justificatory source of its dependent, and yet its dependent can acquire a normative character entirely distinct from that of its source. The idea is simple to state, though its defense is likely more complicated.

The ability of normativity to vary in its character in this way “across generations” may seem puzzling, but the success of many moral theories and views of justification depend on it. Unfortunately, concerns about the possibility of character change are not hard to motivate. One wonders what could be responsible for this change in normative character. There are no immediately obvious answers. For recall that according to the most plausible, presumptively true view, the normative character of some thing—the kind of reason there is for responding normatively to some x—is determined by the character of its normative justification. If a moral principle justifies some virtue, then this virtue ipso facto would acquire a moral character. So far, character change looks like a non-starter.
i. Is “Character Change” ad hoc?

One might hope, though, that help would come from normative inheritance, even conceding the truth of the default view of normative character. Recall that for normative inheritance to work, dependents must “inherit” normativity from their source, such that dependents function as though both dependent on and independent of their source (in a way that helps solve BP and PRR). Perhaps the normative independence gained through inheritance can help make sense of character change. The hope would be that a thing’s normative character can, like normativity itself, possess a certain independence of its source. Since the normative character of some object appears given by its immediate justification, the presence of a justification that is independent of its justifying source may seem to provide the crucial justificatory stopping point—and so provide for the possibility that the object in question would be protected from automatically imbibing the character of its source. For example, since justified rules on RU are supposed to be in some robust sense independently normative of the principle of utility, it might be hoped that these rules could have a different character (hopefully, a moral character) than their source (the seemingly non-moral principle of utility). Just why a justified rule would have a moral character as opposed to some other character, or even no character at all, needs to be explained. If a justified rule is normatively independent (of its source) in the same robust way an intrinsic good is normatively independent (of everything), then it might have no normative character whatever. On the default view of normative character, such a rule would either be like every other dependent and inherit its source’s normativity, or it would be independent in such a robust manner as to leave it without normative guidance.
But maybe this default view of normative character isn’t right, or maybe it is substantially right but incomplete. After all, intrinsically normative things have no normative source, and they nevertheless have determinate normative characters. Maybe this comparison reveals that there is in fact some wiggle room for character change. I will examine this suggestion later⁴. If this suggestion could be developed and defended, the apparently *ad hoc* notion of Character Change would be revealed as a natural consequence of Normative Inheritance. Normative Inheritance would give normative dependents the kind of freestanding status needed to resist taking on board the normative character of their source.

Alternatively, maybe I prematurely dismissed the proposal that the character of a thing is determined by its extension. If I am guilty of rushing to judgment on this, normative inheritance would again be of help, and character change would again be one of its natural consequences. Here’s why: In solving BP, normative inheritance would solve an extensional problem for derivative or dependent justification and the theories that employ it. More specifically, it would enable such theories to capture morality’s proper extension by giving its dependents a certain normative independence. These dependents—or progeny—would then be of the right character simply because they would get morality’s extension right.

*C. Preemption*

The third and final characteristic feature of NP—one that is (like character change) commonly conjoined with normative inheritance—is the ability of progeny somehow to insulate individual actions, motives, virtues and other elements of morality from the normativity of their source. Instead of a source giving independent normativity to its dependents, and then *sharing* the domain over which they might jointly exert their normative

---

⁴ Cf. Ch. 5
influence, a normative source is, in the very fact of bestowing normative independence upon its progeny, exiled from the possibility of this kind of participation. A normative domain has been claimed by its progeny; it has been preempted. I will call this feature of the NP model ‘preemption’.

Less metaphorically, for many theories that invoke normative inheritance, normative sources cannot directly justify those things that are directly justified by their progeny. This seemingly odd feature can be found in various rather sophisticated teleological accounts of moral, legal, or political justification; I have never seen it described as part of a non-teleological moral justification. It may be tempting to say that in taking over the normative domain of what would otherwise be its normative source, normative progeny “kill off” their source. In the case of RU, it might be thought that justified rules “kill off” the principle of utility. But this picture may be misleading, since the source in RU (and in other instantiations of the model) retains its normativity, but is simply stymied in a certain important context. In RU the principle of utility sustains the normativity of justified rules; one way to see this is that a thoughtful justification of particular actions by a rule-utilitarian may even make reference to the utility-conduciveness of these rules (though it need not). As such, the principle of utility is not “dead”, normatively speaking. It is more helpful to think of the rules in RU--or more generally, normative progeny--as preemptively justifying their objects, preventing the justificatory force of their source from applying to these same objects.

i. Preemption and BP

In preempting the normativity of their source, normative progeny remove a potential competitor. Preemption rules out the possibility that the prescriptions of a normative source
might carry the day over those of its dependents. Normative Inheritance without preemption leaves open the possibility that the normative source of a dependent might “overpower” its progeny where their prescriptions happened to conflict. Preemption makes this an impossibility.

Preemption has potentially dramatic extensional consequences, and may be useful in solving BP. In certain contexts, the direct application of a normative source to (for instance) actions is undesirable. Hobbesians and Utilitarians are good examples of moral theorists who in many cases should try to protect actions from the normative influence of their ultimate source. Normative Inheritance by itself affords some such protection, at least extensionally. It exerts normative influence, if you will, upon its objects, but does nothing to protect its objects from other influences. Without preemption, the normative progeny of potentially dangerous principles like the principle of utility might be overwhelmed by their source, and morality’s extension would fail to be captured.

ii. Preemption and the ‘character’ version of PRR

I earlier suggested that Normative Inheritance might, by itself, help make sense of character change if the independence it bestowed upon its progeny were sufficiently robust. But it was left to the imagination just how a normative dependent could be “robustly independent” in such a way as to be immune from being infected by its normative source. Perhaps preemption can help make better sense of this suggestion. Since normative progeny can preempt the justificatory influence of their source completely in certain contexts, perhaps in these contexts it makes sense to think of such progeny having an “independent” character. As already noted, it isn’t clear how to get from independence of character to a character with
a determinate character. Still, preemption might be a good place to start, to ensure at the very least that a source’s normative character is not automatically received by its progeny.

iii. Preemption and the ‘independence version’ of PRR

The role of preemption in solving the ‘independence’ version of PRR can be guessed at, but it may well go too far in its attempt to quash all normative conflict with its source. I will shelve this concern for now, though, as it is best addressed by evaluating the role of preemption in various substantive contexts, which won’t be possible until the next chapter.

iv. Preemption and ad hocness

Note that preemption is not analytic to the concept of normative inheritance: it is not a direct conceptual consequence of normative inheritance that some normative progeny y of source x must preempt x. Still, together with other facts about normativity, preemption might be a natural consequence of normative inheritance rather than merely ad hoc. The thought would be this: normative force cannot somehow be multiplied (or, presumably, depleted) by merely changing its location. Just as there is a law of conservation of mass, there is a similar law that might be thought to apply to normativity. In order to avoid “double counting”\(^4\) we need to locate the normativity of a source in itself, in its normatively charged dependent, or split it between the two—but it cannot subsist fully in both. If, then, normativity is inherited in a certain moral context, in that context the recipient of normativity alone will be normative. Its normative source in that context would effectively be preempted.

Hobbes provides an example of how ‘preemption’ could naturally result from an

---

\(^4\) This is a phrase borrowed from Raz (1985).
attempt at normative inheritance. Hobbes argues that persons in a commonwealth cannot live peacefully by reason alone; in fact, reason alone would lead each individual to uncooperative, ultimately chaotic behavior. So some entity—Hobbes’s “Sovereign”—will be given the right to deliberate and judge for those who have given up their own rights to deliberate and judge. This transference of rights to the Sovereign is possible only because the rights of others have been given up in proportion. That gains in normativity in one place require a sacrifice of normativity somewhere else is intuitively plausible, if only because it is an extension of a basic understanding of bodies and matter. Once the transfer takes place, the rights of the sovereign seem to preempt those of his subjects who gave him his rights in the first place. This preemption, as I have described it, appears to be a natural consequence of normative inheritance.

v. Preemption and the preservation of consistency

There are other reasons to think that a normative dependent should have the ability, at least in certain contexts, of ‘preempting’ or blocking the normativity of its source. Preemption is perfectly suited to alleviate an apparent problem that naturally arises with normative inheritance. The alleged problem can be formulated simply, using the framework of rule-utilitarianism for expediency. Let us call S the source of the normativity of some rule, and D the dependent rule. Suppose that S is the principle of utility: the good ought always to be maximized. And suppose D is the rule that one ought always to keep one’s promises. Now, it may seem that D and S cannot both be correct. For if applied to individual acts, they will in some conceivable contexts (and almost certainly plenty of actual

---

5 An important qualification: It is not clear whether Hobbes means to claim a genuine transfer of rights here, or if he instead means to claim that it makes sense to act as though such a transfer has taken place.
ones as well) recommend incompatible courses of action: sometimes breaking a promise will maximize happiness. So this seems to mean that if S is correct, D cannot be—and vice versa. This would appear to amount to a refutation of the NP model, since this model requires that both S and D be correct—only if S is correct can it serve as the normative ground of D. And D must be correct if it can be said to have inherited the normativity of S. But here is where preemption could come in. The apparent inconsistency of S and D could be avoided is if S could not directly apply to actions—in other words, if S were preempted--so that there is nothing to conflict with D where the morality of particular actions is concerned.

**D. Features of the model jointly considered**

It may have become clear by now why I have titled this chapter “normative progeny,” and why the model of normativity I am exploring bears the same name. For, figuratively speaking, on this model normative sources can breathe a sort of normative life into their dependents, giving them a robust normativity of their own. Though also importantly dependent, normative dependents are in virtue of this fact in some way independently normative as well, often with a character that is different from that of their source. What is more, many dependents are, in a sense, unruly offspring, in that they block the normativity of their source, or “parents”, from applying directly to a particular normative domain. This puts them in a position to overrule their source, and will in certain contexts judge (with authority) their source to be wrong, all the while requiring help from their source for normative sustenance. (How ungrateful!)
II. A promissory argument

With the problem and model of a solution now fully on the table, I can issue a promissory argument on behalf of the NP model.

1. There are a number of things in morality that are clearly of derivative normativity.
2. Familiar ways of understanding the normativity of derivative or (synonymously) dependent normativity saddle derivative justifications, and the moral views that make use of them, with BP and PRR.
3. A full account of normativity must provide the resources to solve BP and PRR.
4. So, familiar accounts of normativity cannot be right. (from 2, 3 and 4)
5. The NP model provides the resources to solve PRR and BP.
6. So, the NP model meets a necessary condition for an acceptable account of normativity. (from 3 and 5)
7. There are prima facie no alternatives to NP.
8. If there are prima facie no alternatives to some model M, where this model meets a necessary condition for models of its kind, then M is prima facie legitimate (that is, it models a reality).
9. So, NP is prima facie legitimate (from 7, 8, and 9)

(6) is a significant (sub) conclusion, as is (9). Of course, (1)-(8) do not entail the legitimacy of the NP model, since for all I have said there might be some other model of
normativity that can solve BP and PRR. The legitimacy of the NP model would be entailed by my argument if the following strong premise were available:

\[ P' \) only if NP is defensible can BP and PRR be solved \]

However, I am not going to argue for this premise, since such an argument would need to take the form of an argument by elimination, the conclusions of which in this case could almost certain not be deductively established. Still, until or unless a compelling alternative to NP is proposed, NP is shown by this argument to be a promising model of normativity.

Of course, this argument will fail to make good on its promise if the NP model is shown to be incoherent or otherwise indefensible in its details. Since the argument I have offered leaves room for the possibility of alternative accounts of normativity—even if it eliminates received or familiar accounts as contenders--this is an important concern. I will need to offer substantial additional defense of NP that shows itself to be immune to the many objections that can be raised to it. If it is shown to be defensible under closer scrutiny, we have good reason to accept it as workable, at least within certain normative contexts.

III. Initial problems for the NP model

Some of the challenges the model faces are evident already, while others will only become apparent after attempts to apply it. At this early, “abstract” stage in the dialectic, I will confine myself to objections that are visible from a structural standpoint.
Even when considered abstractly, the NP model appears “exotic” in ways that threaten to make it, metaphysically, unacceptably mysterious. One motivation for the model is its promise to diminish or even eliminate the need for metaphysically suspicious intrinsically normative entities. But perhaps the NP model itself incurs a heavy metaphysical cost that is not offset by these gains. For it is not immediately obvious—to say the least!—how one might give a metaphysically lean account of an element of morality that, in virtue of its normative dependence, is normatively independent. It might to some appear equally difficult to understand how these same dependents could somehow block the normativity of their source from applying to the things they (the dependents) directly apply to, or how dependents could acquire a different normative character from their source. My use of the metaphor of normative progeny only heightens this sense of metaphysical excess. Certainly we can accept into our ontology actual human beings who give birth to children who require the sustenance of their parents for their survival, but who nevertheless have a different character from their parents, regularly interfere with their parent’s plans and even (when they’re older), judge, with authority, their parents to be wrong about certain things. But do we want to ascribe to normative dependents and their sources--very different sorts of creatures than human beings--these same kinds of properties? To do so might seem to require a gratuitous reification, tolerable only perhaps to Meinongians and believers in concrete possible worlds.

Another worry is the charge of incoherence. The problem was addressed already in section on ‘preemption’. I have suggested that the apparent incoherence introduced by normative inheritance could be solved by preventing normative sources from applying directly to its progeny’s immediate objects of justification. But this obviously would require
a vigorous defense of ‘preemption’, an otherwise puzzling, and so far only weakly undefended, feature of the NP model

The charge of incoherence is compelling. But there are other worries for the NP model, including worries about the defensibility of ‘character change.’ If we remain faithful to the received view of normative character acquisition, there is no obvious way to make sense of the sort of normative character change that would be required by numerous moral theories and justifications. I offered that perhaps normative inheritance could stanch the “bleeding through” of character from source to dependents, and that ‘preemption’ might be needed as well. The hope would be that normative progeny, because they are by hypothesis “independent” of their normative source, would or at least could have an independent character as well. But there is an immediate problem with this suggestion. Using the example of RU again, even if the justification of a particular act appeals only to justified rules, it is the act—not the rule—that is the recipient of the “independent” justification of the rule. The rule itself is still justified by the principle of utility, the latter of which has a non-moral character. The rule in question, then, would seem to take on the character of the principle of utility. But we want justified rules to be of a different character—such as those against casual promise-breaking—not just the acts that fall under them. Character change, at this point, remains objectionably mysterious, and looks like it may require the moral equivalent of alchemy.

A further worry is that it is not clear that there is a principled way of determining which normative dependents of a given source should be normative progeny, and which should remain “merely derivative” dependents. Most of the theories we will be considering privilege a certain sort of normative dependent over others. Rule-utilitarianism, for instance,
privileges rules over virtues or motives (or anything else). However, a rule-utilitarian does not deny that there are certain motives whose possession are highly productive of utility, and neither would a rule-utilitarian deny that there are certain virtues that are similarly utility-conducive. Moreover, the rule-utilitarian will have to admit that certain cars, restaurants, and brands of toothpaste are utility-conducive. But rule-utilitarianism evaluates the rightness of acts only by reference to their conformity with the best rule, and never by reference to motives or virtues, to say nothing of toothpaste. The explanation of this cannot be that only rules are productive of utility, since all sorts of other things are as well. There needs to be a principled way of sorting dependents into sensible normative categories.

Finally, there is a worry that the NP model cannot resist the implication that normative progeny must themselves have the ability to “beget” their own normative progeny, resulting in an unacceptable proliferation of normativity. That is, if normative progeny gain their features simply from their source-conduciveness or from being selected in the right way by their source, why couldn’t these same normative progeny—which, like their source, have a kind of normative independence--also confer upon a new “generation” of dependents the status of normative progeny? And why couldn’t this same status be conferred upon the generation after that? And so on. This possibility of rampant, Malthusian normative proliferation might amount to a reductio of the view being presented, and perhaps it is clear why. One reason is simply that if normativity progeny can be generated indefinitely this might seem to entail just too much independent normativity in the world to be plausible.

Also, on an instrumentalist model of NP, with each successive generation normative progeny

---

6 By speaking of “generations” of normative progeny, I hope it is clear that I do not mean to be making a historical claim; I am not claiming, for instance, that normative sources are temporally prior to their progeny. My concern, here as elsewhere, is with justification; I have in mind here justificatory rather than historical succession.
would likely be less productive of their (ultimate) source—that is, the original normative source that began this particular line of progeny. This is simply because normative progeny might be only imperfectly productive of their proximate source (that is, the normative source that, generationally, is immediately prior to it). According to some versions of rule-utilitarianism, for instance, following the best rule even with perfect vigilance will not always be maximally productive of utility. And the best motive for acting in accordance with these rules would likely result in less-than-perfect rule-following. And so on. As these imperfections are multiplied through the generations, it might seem that distantly related progeny may not only fail to promote their ultimate source, but might in fact regularly promote some good or consequence that directly opposes their ultimate source! If we balk at the evident inconsistency between a source recommending, under rare circumstances, a different action than its immediate progeny, we will surely find the possibility of an (ultimate) source and its (distant) normative progeny prescribing radically conflicting actions under a wide range of circumstances an unappealing result.

The problems just presented are formidable. As difficult as the problems may seem, it is important to see if they might be overcome because, since, as I’ve argued, the defensibility of morality—or at least many of the most appealing views of morality—may depend on it.

In the next chapter I will argue that the structure of a number of normative theories and views about justification embody, or are committed to, the NP model just introduced. Showing that a number of diverse normative theories have this common structure or commitment is one of the principal aims of my project, a result that is interesting, I think, in its own right. Exposing this common structure is also useful for my further aim of assessing
the defensibility of the NP model. Once these different concrete accounts are in place, the objections I have enumerated, and perhaps other objections as well, can be better addressed.
Chapter 4: The NP Model Embodied

I. An intermediate level of abstraction

BP and PRR are old problems—even if unidentified, philosophers have inevitably had to grapple with them. Attempted solutions to these problems that make use of something like the NP model also have a deep philosophical history. At a certain level of abstraction, that is, a remarkable number of moral views and moral theories share a justificatory structure. What level of abstraction am I speaking of? Simply put, the level of abstraction required to articulate this common structure, nothing more and nothing less.

Of course, there are various ways the same structure can be instantiated—this we have already seen. However, to think of those individual substantive theories that embody NP as species of the genus constituted by this normative model could be misleading. Interposed between the model in its full abstraction and the concrete views that embody it is another level of classification. Below I attempt to carve up these categories into useful units. To avoid confusion, the categories described at this intermediate level of classification I will call the ‘genus’ of individual substantive theories that fall under them. For want of a better term (but, for what it’s worth, in keeping with the recurring theme of biological metaphors) I will consider the NP model be to a member of a ‘family’ that constitutes a third tier of abstraction.
II. Taxonomy

All of the theories I will consider below embody, plausibly embody, or are committed to embodying the NP model of normativity. They share (or appear to share, or are committed to sharing) a certain structure, then, by belonging to the same family of theories. However, there are different ways one can be a member of this family, and the variations within the category are subtle but important. I will describe some of these variations in this section, before descending to the individual substantive theories that make use of them. Doing things in this order will help to make better sense of the substantive theories as we consider them, and provide the conceptual framework to recognize generic strategies that are more or less promising.

A. Teleological, qualified source-commitment

One approach has a commitment to the promotion, maximization, or compliance with its source, where the source is a good of some kind or, alternatively, a normative teleological principle. But this commitment is importantly circumscribed. On this approach, dependents indeed are justified in terms of their source-conduciveness, and in virtue of this connection they are granted their independent status (and maybe the other trappings that often attend such independence). But dependents are not selected as progeny because acting in accordance with these dependents will promote or comply with their source better than any other option. Rather, and somewhat surprisingly, dependents are selected because they promote or comply with their source better than any other moral feature of their same kind. So, for instance, the justification of some rule might be that if everyone follows it utility will be promoted better than if any other rule is followed. The justification is not that if
everyone follows the rule, utility will more likely be maximized than if any other option is taken (such as exhibiting the most utility-conducive virtue). As a result, such approaches do not have what could be called an overarching commitment to their normative source or sources.

B. Teleological, source-committed indispensability

On this approach, normative dependents are normative progeny because the only way to comply with or promote the source (where the source is a good of some kind or, alternatively, a teleological principle) is somehow to “go through”—that is, promote, or comply with, or inculcate, etc.—its normative dependent(s). On this approach, it allegedly never makes sense to try to bypass these normative dependents in favor of their source, because there is no other route to their source than through these dependents. So these dependents—in virtue of their indispensability—function in some way as though they were independently normative. Like things with intrinsic value—which serve as role models for dependents that aim to achieve independence as well—the value they are connected to cannot be had except by directly promoting (or following, or in some other way appropriately responding to) these dependents.

The supposed indispensability of dependents on this approach explains why I call it “source-committed”: the dependents of a normative source are granted normative independence only if the dependents in fact promote or comply with the source better than any other candidate dependent.

To mimic intrinsic value more completely, a normative dependent should also behave like intrinsically valuable things in this respect: it should always provide one with a reason to
promote it (or follow it, or otherwise respond to it as normative). To achieve this, though, it is not enough to be indispensable or necessary. For though something might be a necessary means to something of value, it may not always—or even reliably—lead to that thing of value. For instance, let us suppose that the only way to get to some remote island is by taking the lone bridge connecting the island to the mainland. But sometimes the bridge breaks apart or is otherwise out of commission. Under these circumstances, one does not have a reason to cross the bridge if one wants to get to the island—even though the bridge is indispensable for reaching the island. The more reliably some indispensable means leads to something of value, the more closely it mimics, in one respect, intrinsic value—the latter of which has as robust a sense of normative independence that can be conceived of.

C. Teleological, source-uncommitted dispensability

A third approach is like the first two in that dependents are normative progeny only if they have played or continue to play an instrumental role in bringing about something of value. A bit surprisingly, though, progeny are not necessary means to the end in question, and neither is there a commitment—even a qualified commitment—to maximally promoting or complying with the source. Plenty of other things can in practice or could in principle be substituted for the alleged progeny without the latter losing its special normative status. Additionally, their status would remain unchanged if plenty of other things—even things of the same kind—could more effectively respond to the normative source.
D. Non-Hypothetical Agent-election

A fourth approach breaks decisively with the first three. On this view, normative progeny possess their characteristic properties because they are the actual objects of a special kind of election: they are the objects of actual rational choice or some privileged attitude or mental state. Such an approach is clearly not teleological: progeny do not receive their status because are useful or indispensable means to their source, such as the end of choosing rationally. Instead, dependents are progeny simply because they are objects of the right kind of election, where the “elector” is a normative source.

Recall that to count as embodying the NP model, a source must itself have value that it confers upon the objects of its (rational, or otherwise privileged) choice or other mental state. Only in this case is the metaphor of normativity being inherited by a dependent from its source applicable. Thus, certain radically subjectivist views about morality do not count as versions of the agent-election approach. The views I have in mind have it that objects are valuable if they are valued in themselves, but there is no claim that their value is a result of the value of the person who values them. Elizabeth Anderson’s theory of value, though not “radically subjectivist” provides another example of a view that is easily mistaken for a agent-election version of NP. She holds that being the object of certain non-propositional attitudes (like love) makes an object valuable in some quasi-independent way. But she does not claim that the holder of these attitudes must herself be of value in order for the objects she elects to be independently valuable, or that the propositional states themselves must have value—much less that the value of either is somehow passed along to the objects she elects.¹

¹ Anderson (1993).
E. Hypothetical Agent-election

This is a variation of the previous approach. On the previous account, certain dependents are normative progeny in virtue of being the actual objects of rational choice, desire, etc. The normativity of such dependents thus might seem threatened with instability, as what is rationally chosen, or desired, or loved, etc., will seemingly vary with time and place. On the hypothetical variation, dependents have the status of normative progeny in virtue of the fact that they would, under certain idealized conditions, be the objects of choice, desire, or some preferred attitude of an agent or group of agents.

III. The NP model variously instantiated

Now I will argue that these versions of the NP model are variously employed (or in one case, apparently ought to be employed) by a number of important substantive ethical theories. I will group the substantive accounts according to the categories I’ve just described, calling attention to the features the possess that make them members of the NP ‘family’, and also the features that make them members of their respective ‘genera’.

A. Teleological, qualified source-commitment

i. Rawlsian Rule-Utilitarianism

Rule-utilitarianism has been described or used as an illustration a number of times already, and for good reason: it is probably the most obvious case of a moral theory that, in its classical and most frequent instantiations, is committed to the NP model. Just what version of RU, though, is the theory’s best representative? Though most well-known as an
act-utilitarian, Mill was probably the first philosopher to argue for something like rule-utilitarianism in a sustained way. However, since the interpretation of Mill on this point is controversial, let us turn to a more straightforward rule-utilitarian view, the view presented by the early John Rawls.

I choose Rawls not because his account is a paradigmatic version of RU—a better candidate for this might be Richard Brandt’s version—but because Rawls in his famous discussion “Two concepts of Rules" is acutely aware of and concerned to address BP and (at least some version of) PRR. Unlike the flat-footed arguments of many rule-utilitarians and other indirect consequentialists, Rawls offers an ingenious defense—or what he calls an “interpretation”—of the theory. He focuses on two kinds of cases—promise-keeping and punishment—though I will primarily discuss his example of punishment.

Rawls argues that (what most now call) the act-utilitarian account of punishment fails to capture the intuitively appropriate reasons for punishing wrongdoers. On the one hand, he notes, it does make sense, if asked why, in general, we punish wrongdoers, to cite the good consequences of having such a practice. On the other hand, he claims that if we are

---

2 See Urmson (1953) for a Rule-Utilitarian interpretation of Mill. George Berkeley, not primarily known as a moral philosopher, is also (surprisingly) sometimes credited with being the first rule-consequentialist. Brandt (1992) is also sometimes thought of as presenting one of the first classic versions of rule-utilitarianism.

3 For instance, Brandt (1992).

4 Rawls (1955).

5 Interestingly, James Rachels seems to rely on something like the view of normative dependents as normative progeny in his defense of retributive punishment (See Rachels [2002]). Rachels acknowledges that his defense of retributivism could be reasonably interpreted as utilitarian; but he nevertheless insists that his account is retributivist and desert-based. Some of his language echoes Rawls’ own: “To justify punishing someone, we may refer simply to what he or she has done—we may point out that they deserve it. But when we examine the arguments that support the general practice of treating people as they deserve, it turns out that those arguments all refer to ways in which people are better off under such a practice.”

6 Rawls (1955) 275, 276.
asked why we ought to punish some individual person, it seems that a retributivist justification should be available: so and so broke the law and should be punished in proportion to her offense. We would seem to misunderstand punishment, he thinks, if we were to say: we should punish so and so because this act of punishment will promote the overall good. But of course, only this last sort of justification is available to the act-utilitarian. This is an instance of what I have called PRR—act-utilitarianism gives us the wrong sorts of reasons for punishing. It apparently does not give the rules of punishment an appropriate independence from their source. Further, and not obviously just an automatic product of this lack of independence, act-utilitarianism seems capable of only a non-moral justification of punishment.

Rawls additionally addresses BP. He argues that the standard act-utilitarian defense of punishment seems to “allow too much”. Act-utilitarianism appears to justify “cruel and arbitrary” institutions of punishment; it might justify, for instance, the “punishment” of the innocent, or punishments that were disproportionate to the offense committed. This is a complaint about act-utilitarianism’s extension: what we might hope are hard and fast proscriptions against such unjust and arbitrary punishment and injury can too often be bypassed. Namely, they can (and should) be bypassed when doing so will be maximize utility.

According to Rawls, the best interpretation of utilitarianism shows promise in allowing us to avoid both of these problems. In its broad outline, his version of RU is

\[\text{Rawls, (1955) 277.}\]
familiar. Rawls thinks the right interpretation of utilitarianism⁸ applies the principle of utility directly to rules, but not to individual acts⁹. More precisely, we are first to evaluate the rightness of certain rules according to whether or not the rules, if widely promulgated and followed, would be better than any other rules. And (this is the second step) we are to evaluate the rightness of acts by reference to these rules alone. So, an action is right if and only if it conforms to the rules that have been “selected” by the principle of utility.

This interpretation of utilitarianism, according to Rawls, captures the strengths of familiar utilitarianism, and, in the parlance I have been using, also aims to defuse BP and PRR.

It attempts to solve the ‘independence version’ of PRR by employing its clever version of normative inheritance. Consider the case of punishment: it allows us to evaluate what Rawls calls the “practice” of promise-keeping in terms of its consequences. This kind of consequentialist justification of a practice such as punishment seems appropriate, and Rawls’ theory accommodates the intuition nicely. So if someone asks about punishment generally “why ought we to punish wrongdoers?” Rawls’ version of RU allows us to cite the benefits to society that come from having such an institution, since the practice of punishment is justified directly by the principle of utility. Put in other words, we can appropriately defend punishment by referring to its usefulness. Of course, showing morality’s dependence is the easy part; the hard part is showing how this dependence is compatible, and in fact makes possible, its independence. On this point he claims that if we are asked of an individual case “why ought so and so be punished?” we can, invoking his

⁸ By “right interpretation” Rawls likely means to refer to the interpretation that is most plausible as a normative theory, not the interpretation that is most faithful as an explication of actual utilitarian views.

⁹ Rawls (1955) 276.
conception of utilitarian rules, correctly respond that those who violate the law ought to be punished in proportion to their offense. For though the practice of punishment is justified in terms of utility, the practice of punishment also has normative force seemingly “of its own”, since people are punished or not punished directly by reference to the constitutive rules of the practice of punishment, rather than to their justifying principle. This gives the rules governing punishment and the practice they constitute at least an appearance of normative independence, since they function as though they were self-standing.

The extensional problems of act-utilitarianism, like the injustice allowed in cases of “punishment”, are likewise dealt with by Rawls’ rule-utilitarianism. Rawls’ theory attempts to block a direct appeal to the principle of utility that allows this sort of injustice, and will in many cases proscribe the performance of what intuitively seem to be immoral or unjust actions. This is simply because actions we judge to be immoral and unjust are very often actions that would be disallowed by the most utility-conducive rules.

Rawls’ interpretation of utilitarianism apparently aims to conform to the NP model along all three possible dimensions. As may be evident, on his account moral rules are the dependents of the principle of utility, their source, but also, and in virtue of their dependence, function as though independently normative of this relationship. So normativity seems to have been inherited by these rules from their source. Further, moral rules *preempt* the application of the principle of utility: justified rules block the principle of utility from applying directly to actions, and so block its normative force. Finally, and perhaps least obviously, the normative character of the principle of utility and the rules it grounds are supposed to be different. Pleasure and happiness may be great goods, but it is generally agreed that they are not *moral* goods. And so it is not clear that the principle of utility—
which enjoins the maximization of these goods—is a moral principle. Nevertheless, Rawls claims, the rule-utilitarianism he proposes is a moral theory, and the rules that are sanctioned by the principle of utility he claims are moral rules. If Rawls is right, then, moral rules have a normative character different from that of their normative source.

It is not yet clear to which of the five categories described at the outset of this chapter Rawls’ theory belongs. To judge this, we must look at Rawls’ defense of his picture of the relationship between rules, actions, and the principle of utility, not just the picture itself.

Part of Rawls’ defense of his picture is that it provides a statement of utilitarianism that explicates our considered moral judgments better than the more familiar act-utilitarian picture. As I understand Rawls’ claim, this is accomplished by providing a solution to the BP and PRR. Given Rawls’ understanding of moral epistemology (and perhaps moral justification as well) a theory’s ability to explicate our considered moral judgments is an important part of its defense. But it is not enough by way of defense of the NP model to show that some substantive normative theory that employs the model squares with our considered moral judgments better than some alternatives. A model of normativity should not only be capable of being employed to deliver moral verdicts that nicely align with such judgments, but should be independently defensible and coherent. In particular, to defend the NP model that undergirds his moral theory—and, as a result, to defend his moral theory itself—Rawls must show: 1) why moral rules have a different normative character from the principle of utility; 2) why the rules that promote utility better than any other rules are not

---

10 Rawls (1955) 274.

11 He defends an account of moral knowledge according to which moral truth (or something like it) is approached by modifying our beliefs about moral principles and the rightness of particular actions where they conflict. The goal is to reach a state of agreement between principles and actions. This is an understanding of moral epistemology that Rawls held onto even after he had abandoned hope in Rule Utilitarianism.
merely heuristic devices that can, under the right circumstances, be dismissed; 3) why, even if the best rules are independently normative, the normative force of the utility principle doesn’t simply join its independent progeny in exerting normative influence over individual actions.

Rawls’ defense of at least the normative inheritance and preemption (and perhaps also character change) in his theory depends upon a logical distinction between two different types of rules. Utilitarianism is typically interpreted as relying upon what Rawls calls a ‘summary’ conception of rules, while his favored interpretation of utilitarianism depends upon a ‘practice’ conception of rules. According to the summary conception, rules are “summaries of past decisions arrived at by the direct application of the utilitarian principle to particular cases”\(^{12}\). They are used because they help us to decide cases more quickly than we could through complex, individualized utility calculations. On this conception of rules, particular actions are ‘logically prior’ to the rules that summarize them. Actions exist whether or not there is a rule that summarizes them, and this makes it such that they may be described without reference to some such rule. This last fact—the ability to describe an action without reference to any set of rules—qualifies them as ‘logically prior’ to such rules\(^{13}\). Accordingly, one may justify one’s actions by direct appeal to the principle of utility, since such a justification can be intelligibly described.

\(^{12}\) Rawls (1955) 283.

\(^{13}\) The ability to specify an act without reference to any rules depends, according to Rawls, upon an ontological fact—namely, that the acts exist without the rules. But this ‘ontological priority’ is not, for Rawls, itself criterial for logical priority.
On Rawls’ preferred interpretation of the role of rules in utilitarianism, the logical order of rules and acts gets reversed: certain moral rules that constitute a moral practice\textsuperscript{14} are ‘logically prior’ to the acts that fall under them. An action that falls under a ‘practice’ could not be described as the sort of action that it is if the practice failed to exist\textsuperscript{15}. This is simply because, without the existence of the practice, the actions that fall under the practice would not exist. For instance, the rules in the game of baseball are logically prior to particular actions that fall under them because one could not properly be said to steal a base, balk, hit a homerun, etc., except in the context of a baseball game. If there were no such thing as the game of baseball with its set of rules, I could run from one bean bag to another on a diamond-shaped field of dirt, while one person hurls a small sphere wrapped in leather towards another player wielding a stick (etc.), but it wouldn’t be stealing a base. Similarly, thinks Rawls, certain rules of morality (like those concerning punishment and promise-keeping) are logically prior to the actions that fall under them: individual acts of punishment could not exist apart from the rules governing punishment, and so could not be described unless there were such rules. I could confine some person to a room after he performed some act, but I could not be properly described as punishing him unless there existed the rules governing punishment.

Rawls takes this logical point to have startlingly substantive implications. From this logical point Rawls thinks a kind of independent normativity of certain (namely, justified) moral rules follows. From this logical point he also thinks it follows that such rules insulate individual acts that fall under them from the normative force of the principle of utility. If one

\textsuperscript{14} By “practice” Rawls means “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.” (Rawls [1955] n1).

\textsuperscript{15} Rawls (1955) 286.
wants to engage in the practice of punishment, it is supposed to be obvious that the rules
governing punishment apply in an exclusionary way to a particular case of punishment,
preempting considerations of utility from bearing directly on the particular case. The rules of
punishment cannot be bypassed, then, even when doing so would maximize utility. The
“punishment” of an innocent person, for instance, would be immoral, even if doing so
maximized utility, because punishing the innocent is disallowed by the rules that constitute
the practice of punishment.

There is more to be said here, and more will need to be said when we evaluate Rawls’
argument in the next chapter. But for now the arguments are clear enough for the purpose of
taxonomy. Rawls’ theory can be categorized as teleological and source-uncommitted. The
teleological status of the theory is straightforward—a good is acknowledged (utility) and this
good ought to be promoted. But it is to be promoted only within limits, which means that
Rawls’ theory does not have an overarching commitment to promoting utility. Rawls does
not defend his brand of rule-utilitarianism on the grounds that following (or mostly
following, or promulgating) certain moral rules is required to the maximization of utility.
Some versions of rule-utilitarianism are defended in this way\textsuperscript{16}, but it is not Rawls’ way. In
this way he is “source-uncommitted”. Instead, even apart from his goal of explicating our
considered moral judgments, he gives a mixed defense of the normative status of moral rules
as normative progeny, where the normativity of the principle of utility determines which
rules are moral, but where considerations of ‘logical priority’ restricts the direct application
of the principle of utility to rules alone.

\textsuperscript{16} For example, see Brandt (1992).
B. Teleological, source-committed indispensability

i. Hare’s two level theory

I have suggested that classical act-utilitarianism’s attempt to establish a causal connection between the demands of utilitarianism and the following of conventional rules as well as the exhibition of accepted virtues comes up short in solving BP and PRR. Other utilitarians, such as R.M. Hare, make a better run at solving these problems than their intellectual ancestors, and they do this by providing better arguments for a regular causal connection between following conventional morality and satisfying the requirements of act-utilitarianism.

Hare argues for a two-level theory that at least superficially resembles Rawls’ rule utilitarianism. These two levels Hare calls the ‘intuitive level’ and the ‘critical level’. The critical level is simply classical act-utilitarianism, and is in fact the theory’s criterion of the right. At this level one is required to maximize utility. The intuitive level is the level of commonsense moral rules, and is the level at which one is recommended to make most of one’s moral decisions. Hare argues that, because of our epistemic limitations, we are very often not in a position to act at the critical level. We are ‘proles’, rather than ‘archangels’—we do not have the epistemic vision to see with clarity the outcome of our actions. This means we are unable to determine what actions are required in order to maximize utility. So instead, as ‘proles’, we ought to fall back upon the rules of commonsense morality. Doing so, Hare argues, will very often be better—from the standpoint of the maximization of utility—than self-consciously attempting to satisfy the act-utilitarian principle. Only rarely is one in such a strong epistemic position that it makes sense, from the standpoint of utility maximization, to attempt to directly satisfy the act-utilitarian principle.
Hare’s account might seem to attempt an embodiment of the NP model since the normative reach of the principle of utility is supposed to be extended to parts of morality (namely, certain commonsense moral rules) other than the principle of utility itself. If rules grounded by utility were indispensable to promoting utility—assuming certain facts about our epistemic limitations, of course—such rules would have inherited normativity from the principle of utility. This would seem to help solve the bypass threat for act-utilitarianism. Additionally, a good Hare-style utilitarian, while thinking as a ‘prole’, would offer as his reason for acting in accordance with moral rules a justification that made reference to the rightness of the rules themselves rather than to the fact that his action would maximize utility. This might seem to offer a solution to PRR, since ‘proles’ would justify their actions by reference to commonsense moral rules, rather than the principle of utility. And they would not appeal directly to the principle of utility at all, which might seem to be an instance of preemption. Finally, the fact that an appeal to the non-moral principle of utility is not allowed by a ‘prole’, but instead only by certain principles, might seem to show that these principles have a moral character—or at least to allow this possibility.

It should go without saying that Hare’s theory is teleological. It is ‘source-committed’ because it acknowledges that the only reason one should not self-consciously maximize utility is that this strategy is self-defeating. And finally, normative dependents are to be followed and treated as though independently normative because they are necessary or indispensable for comporting with their source.
ii. Gauthier's Contractarianism

I mentioned that there was at least a superficial similarity between Hare’s two level theory and Rawls’ rule-utilitarianism. The structural similarities between Gauthier’s Hobbesian contractarianism and Rawls’ rule-utilitarianism are also well-recognized. Both are “indirect” versions of consequentialism, and, in the parlance of this discussion, both attempt to embody the NP model.

Their structure, though broadly similar, is defended in importantly different ways. The two theories are also different in their aims. While Rawls sets out with the explicit intention to solve BP and the PRR, Gauthier only grudgingly, with a set of caveats, acknowledges BP as a goal, and has a similarly qualified concern for PRR. This is probably a reflection of different moral methodologies. While never committing himself to foundationalism, Gauthier is more comfortable than Rawls with jettisoning commonsense moral convictions where these convictions conflict with theory. Nevertheless, Gauthier’s theory does show some initial promise in solving BP and PRR—even if more as a general strategy than in its details.

Like Hobbes, Gauthier hopes to show that we have reason to be moral based upon non-moral considerations. Specifically, Gauthier hopes to show that it is rational to be moral—even if ‘morality’ here does not perfectly resemble commonsense morality. For Gauthier, an agent is rational if and only if she acts to achieve the maximal fulfillment of her preferences—or in other words, her expected utility. Gauthier does not assess the rationality of the preferences themselves, except to impose upon them certain minimal conditions of

---

coherence. Value for him is both subjective and relative, and the only non-relative norms for assessing preferences are provided by these conditions of coherence. Rationality, then, is independent of any agent’s particular desires or goals—which is something that will be gratifying if we share with Kant an aversion to such a contingent and variable ground of rationality—but it is not independent of the fact that we have desires or goals in general\textsuperscript{20}.

Gauthier thinks rationality requires the constraint of utility maximization. Given the rationality of maximizing our expected utility, Gauthier argues that in a number of bargaining situations (though not necessarily all\textsuperscript{21}) we ought to refrain from straightforwardly maximizing expected utility. Gauthier uses prisoners-dilemma scenarios\textsuperscript{22}—which, he thinks, reflect the structure of a great number of situations we find ourselves in—to illustrate his claim. The setup is this: you and your partner are self-interested, and you each individually aim to maximize your expected utility. You each have the choice to cooperate or not to cooperate. The ‘dominant’ strategy—the strategy that has the greatest expected utility no matter what your partner does—is not to cooperate. But, surprisingly, if you both employ the dominant strategy—that is, if you both don’t cooperate—you will both be worse off than if you had both cooperated! Now, you stand to gain the most if your partner cooperates and you do not, and you have the most to lose if you cooperate and your partner does not. So, you might think that the most rational strategy would be somehow to get your partner to cooperate and then yourself not cooperate. This is the “Foole’s” strategy, of course, and the apparent rationality of this strategy is an instance of the free-rider problem.

\textsuperscript{20} Gauthier (1991b) 23.
\textsuperscript{21} See, for instance, Gauthier (1991a) 330.
\textsuperscript{22} And other game-theory situations like it that, like prisoner’s dilemmas, exhibit ‘market failure’—that is, situations in which if each person chooses the straightforwardly maximizing option everyone could do better.
But this is not the strategy Gauthier recommends. Instead, he argues that it is rational—
again, where an agent is rational if and only if he acts to maximize expected utility—to
cooperate in a number of such dilemmas, constraining the direct pursuit of utility-
maximization.

To defend this claim, Gauthier draws on the contractualist elements of his theory. He argues
that it is rational to choose to dispose oneself to act in accordance with constraints that would
be agreed upon in an appropriate premoral contractual situation. These constraints would
prohibit straightforward maximizing, requiring instead constrained-maximization.

Unfortunately, Gauthier imposes a number of controversial idealizing conditions upon this
contractual situation (such as the full rationality of all participants, an inability to coerce one
another, mutual disinterest, and translucency\textsuperscript{23}). Given Gauthier’s characterization of
rationality in terms of expected utility-maximization, it is not obvious that he is justified in
appealing to a \textit{hypothetical} contractual situation to articulate what sorts of constraints are
rational. For what would maximize expected utility in the actual world might not be
consistent with the constraints chosen under these hypothetical and idealized situations.

Now, Gauthier is explicit in his aim to connect idealized agreement to how we ought
to act in the real world\textsuperscript{24}, and tries to make this connection by appealing to the stability of the
practices that would secure such agreement. He argues that the reflective capacity of
individuals--in particular, their ability to reflect upon the current practices and realize their
inferiority relative to those that would be chosen under an ideal situation—makes the current
practices unstable. In the interest of stability, then, which is presumably supposed to be

\textsuperscript{23} An individual is ‘translucent’ if their disposition to cooperate can be determined (by others) with greater
accuracy than would be achieved by blindly guessing, but without perfect accuracy. (Gauthier [1986])

\textsuperscript{24} Gauthier (1986) 174.
utility-conducive, rational agents have reason to act according to the constraints that would secure hypothetical agreement, and to encourage others to do the same. But it is not at all clear how acting according to such constraints will help actual individuals, given our limited ability to influence others to agree to these constraints. If we act according to constraints that would secure hypothetical agreement but others we interact with do not, we expose ourselves to being taken advantage of. Gauthier, then, is unable to show that acting according to constraints agreed upon in an idealized situation is rational according to his own notion of rationality. At best, he shows that we have an interest in nudging others towards acting according to these constraints, and then acting on them ourselves only to the (limited) extent that we have succeeded in influencing others.

Perhaps one of the reasons Gauthier appeals to hypothetical agreement is that the constraints that would be agreed upon under idealized situations will more closely resemble morality as we know it than are the constraints it would be rational to act in accordance with in actual circumstances. This would mean that Gauthier, despite at points eschewing explication of commonsense morality as an important aim of moral theory, has at least some wavering interest in trying to “save” it. In any case, I think we can discuss Gauthier’s

25 Sometimes Gauthier writes as though an individual acts rationally if and only if she acts according to constraints that would be agreed upon in this premoral situation: “…if one benefits more from a constraint on others than one loses by being constrained oneself, one may have reason to accept a practice requiring everyone, including oneself, to exhibit such a constraint. We may represent such a practice as capable of gaining unanimous agreement among rational persons who were choosing the terms on which they would interact with each other. And this agreement is the basis of morality.” In other places, he seems to concede that acting in accordance with such constraints might be irrational in the real world: “…it would be irrational for anyone to accept a long-term utility loss by refusing to comply with the existing moral order, simply because she comes to realize that such compliance affords her less than she could expect from pure rational agreement.” One way of reconciling these apparently conflicting views of rationality is to call what would be rationally agreed upon in an appropriate premoral situation moral constraints, and the constraints it is rational to agree upon in the actual world rational constraints. This would go along with Gauthier’s insistence that rational constraint is not by itself moral. This consistency would come at a cost, for he would no longer have shown us why it is rational to be moral. (Gauthier [1991a] 326).
argument without bringing in his problematic use of hypothetical agreement, and so without
going into any more of its details. In places, Gauthier himself seems not to rely upon
hypothetical agreement. His argument is stronger without it, and he may be able to get
much of what he wants simply by appealing to a plausible description of actual
circumstances, without running the risk of irrelevance imposed by his appeal to idealized
agreement.

If we abandon his problematic appeal to idealized hypothetical agreement, Gauthier
could still claim that in the actual world many of us would be rational to dispose ourselves to
be constrained maximizers. The rationality of choosing such a disposition is determined by a
number of factors. These include the degree to which we are translucent; the degree and
extent to which others are translucent, the percentage of people we interact with who are
moral (that is, cooperators); and the personal cost of deception. Surely there are other
factors as well. The numbers matter here, of course. But, for many people at least,
choosing a constrained-maximizing disposition might well be rational. With such a
disposition one might expect to benefit more than straight-forward utility-maximizers, since,
given reasonable assumptions about the degree to which one is translucent, we can expect to
gain the benefits of cooperation from which straightforward maximizers will be excluded.

To the Foole, then, Gauthier replies that because adopting the disposition to be a
constrained maximizer is rational, the rational agent acts rationally when he acts in a way that

---

26 For instance, in his (1984a) and (1984b). Here Gauthier is not trying to defend morality per se, but rather is
offering a solution to the paradox of deterrence. He nevertheless presents an argument with much the same
structure as we see in his moral arguments, only without an appeal to hypothetical agreement.
27 One such cost might be a sense of guilt, or perhaps a sense of alienation from others.
dispositions to these factors (and some others). He argues that though not everyone has a self-interested reason
to be moral (for instance, consider those who are both excellent at deception and can deceive without great
personal cost) many of us do have such a reason.
expresses this disposition. This is not because, as it might be tempting to suppose, in free-riding in any given situation one always runs the risk of being found out and so runs the risk of being excluded from utility-promoting cooperation in the future. According to this (mistaken) interpretation, the seeming paradox is dissolved by denying that a real possibility (instances of cost-free free-riding) is being described. Hobbes himself is frequently interpreted as offering an argument to this effect, but it is not Gauthier’s. And it is good for Gauthier that this is not his argument, for it would commit him to the highly implausible claim that for every situation in which one might be tempted to free-ride one could reasonably expect to be found out and excluded from future cooperation.

Gauthier has a better defense available to him. He claims that choosing a sincere disposition to be moral rules out the possibility of free-riding, and that binding oneself to such a disposition in fact maximizes one’s expected utility. This is an interesting empirical claim and worthy of discussion, but I suggest setting this claim aside and granting it for the sake of argument. If he is right that a sincere disposition to be moral precludes free-riding, then when sincerely moral person x is confronted with an opportunity where free-riding would maximize her expected utility, this will simply not be an option: it will not be considered, and it will not be taken. Moreover, it would not be rational. If the option is considered and taken by some person who appears to be moral, this just shows that she was not moral after all! Gauthier’s explanation of the apparent paradox, on my preferred interpretation, is that the sincere disposition to be moral precludes acting in ways that don’t express that disposition—so long as the relevant circumstances have not changed since
choosing that disposition\textsuperscript{29}. Gauthier writes: “Only the person truly disposed to honesty and justice may expect fully to realize their benefits....such a person is not \textit{able}, given her disposition, to take advantage of the ‘exceptions’.\textsuperscript{30} Given the \textit{inability} of a genuinely moral agent to act in a way contrary to her disposition, it makes no sense to evaluate the rationality of actions independently of their controlling disposition.

Consider what I take to be an analogous situation: John playing tennis and has approached the net. He has just hit a weak volley, and his opponent has his choice of shots: he might try to pass on John’s backhand side, his forehand side, or he might try to lob him. John knows from experience that his opponent favors passing in these situations, and that he usually goes to the forehand side. John realizes that he will maximize his chances of winning the point if he lunges to the forehand side in anticipation of his opponent’s shot. But it is also the case that once he has moved in that direction, he will not have time to change direction and hit the ball back if it does not go where he anticipates. He will be committed, in other words. So John, rationally—given his aim to win the point—commits himself to the forehand side by lunging in that direction. Uncharacteristically, though, his opponent hits a shot that is heading down the opposite sideline. Now, mid-lunge, we might ask: is John’s movement toward the forehand side still rational? But this question seems to make no sense apart from the rationality of John’s decision to the forehand side. If \textit{that} commitment was rational, then so is his mid-stride movement.

\textsuperscript{29} Gauthier includes this proviso in his discussion of the ‘paradox of deterrence’, where he does not appeal to hypothetical agreement. Where the rationality of dispositions is determined instead by hypothetical agreement, such a proviso is obviously unnecessary.

\textsuperscript{30} Gauthier (1986,) 182. Emphasis added.
Having a moral disposition, for Gauthier, seems to involve something like this kind of commitment—only the commitment is to acting morally. Thus, it makes no sense to evaluate an action that flows from a moral disposition independently of the rationality of the disposition itself. In evaluating the rationality of an action, no appeal to the egoistic principle of utility—the source of the normativity of the disposition from which the action flows—is allowed or even, in some sense, possible.

As mentioned, Gauthier does not worry much about whether his theory extensionally captures what Rawls would call our “considered moral judgments,” though one can see how Gauthier’s theory, if otherwise defensible, would be much better on this score than straightforward egoism. This is because Gauthier embodies the NP model. The moral disposition functions as though independently normative and, additionally, blocks the application of the egoistic principle of utility to particular cases. As a result, actions are to be done only if they express the moral disposition. If this is right, his theory makes use of normative inheritance and preemption to help solve BP.

His employment of NP might also give us at least some of the right reasons for acting morally: we are to keep our promise in a particular instance not because doing so will maximize our expected utility (sometimes, in a particular case, it will not!), but because keeping promises is rational. This is an intuitively better reason than the standard Hobbesian can give, whose reasons for acting morally are not only non-moral, but arguably immoral. One might worry that the non-moral character of the normativity of Gauthier’s normative source, and so the non-moral or immoral justification given for the moral disposition, would mean that the normativity of the disposition would have a non-moral character as well. But

---

31 The exception is when the circumstances in which one bound oneself to a disposition have relevantly changed.
Gauthier may think that the potential bleeding-through of normative character is stanched by the preemption built into his theory. The reasoning might be something like this: it is granted that the normative character of a thing is determined by its justification. But when justifying some particular action, one only need reference its conformity or failure to conform to a moral disposition; one need not reference the egoistic principle of utility that justifies the disposition. And so it might be thought that in the context of justifying particular actions the moral disposition is moral in character—or at least is not automatically non-moral in character—since it is not, in these contexts, justified by what would otherwise be the spoiling effects of the egoistic principle of utility.

I am now in a position to categorize the brand of NP model Gauthier employs in his theory. One might be tempted to suppose that Gauthier offers an agent-election account of normative progeny, the thought being that normative progeny have their characteristic properties because they are objects of rational hypothetical agreement. But Gauthier makes clear that this would be a confusion. He stresses that the framework of hypothetical agreement—which I prefer to jettison in any case—is merely a heuristic, and that the reason we ought to act in accordance with a disposition we would have chosen under hypothetical agreement is that doing so will maximize utility. He writes, “it is not the choice itself, but the maximizing character of the disposition in virtue of which it is choiceworthy, that is the key to our argument.”^32

Gauthier also, unlike Rawls, offers an unqualified source-committed teleological approach. This is again evident in his emphasizing that the only reason it is rational to dispose oneself to be a constrained maximizer is that in so disposing oneself one can expect

---

32 Gauthier (1986), 183.
to maximize utility. It is not just that this disposition is better than any other disposition for the purposes of promoting utility. Rather, Gauthier thinks that binding oneself to this disposition is likely to promote utility better than conforming to some rule, or acting from some motive, or straightforwardly maximizing, or doing anything else. The most rational dispositions possess their normativity just in virtue of the fact that they maximize expected utility, and they are indispensable to achieving this aim. Thus, Gauthier’s theory is an example of a teleological, source-committed indispensability account.

### iii. Razian reasons and authority

Another approach is inspired by an intriguing metaphor Rawls invokes in presenting his version of Rule Utilitarianism, one that appeals to the different roles appropriate to judges and legislators. Rawls employs this metaphor to illustrate the two-tiered structure of his moral theory. The thought is similar to the claim often expressed by politicians over supreme court decisions: it is not appropriate for judges to make legislative decisions in deciding a case; instead, judges should enforce the existing legislation. Similarly, thinks Rawls, actions that fall under particular practices are appropriately evaluated only in terms of their conformity to the rules that constitute practices. On the other hand, legislators are allowed and in fact ought to evaluate and change legislation, but leave its enforcement to judges. This parallels the appropriate evaluation of ethical practices (according to Rawls), which, unlike particular acts, are appropriately evaluated in terms of their utility-conduciveness.

As I have discussed, Rawls defends this two-tiered consequentialist system of justification by appealing to the logical relationship between certain practices and the actions

---

33 Rawls (1955).
that fall under them. Conrad Johnson\textsuperscript{34} defends a two-tiered system as well, but leaves Rawls’ own argument aside and instead appeals to a Razian concept of authority in order to explain why sometimes, like judges, we must follow certain rules, and at other times, like legislators, we are allowed to revise the rules themselves.

The intuition Johnson appeals to is simple and compelling: in many cases where we are tempted to bypass some moral rule in favor of promoting the best consequences, we ought not do so because we lack the proper “authority”. This thought nicely reflects the way many of us often intuitively react and judge the actions of those who bypass moral and legal rules. Consider the aggravating situation of being stuck in traffic as we watch a fellow motorist breeze by us on the shoulder, bypassing the gridlock and the rules of the road at the same time. Even if it seems likely the motorist will not harm or inconveniences anyone, a natural thought for the rest of us is: “who is \textit{she} to be doing \textit{that}?” We think that the motorist has overstepped the bounds of her authority, acting in a way that is only fitting for those who occupy a very different role (such as highway workers, police, or emergency medical personnel).

Johnson acknowledges that the success of his account depends on Raz’s own account of the relationship between authority and reasons, for which he offers no independent defense. So we must examine Raz’s account of, and argument for, the nature of this relationship.

Raz’s argument connecting authority and reasons for action is strategically similar to the arguments for a two-level theory offered by Hare and Gauthier. Though his argument explicitly concerns political authority, the argument can work equally well, at least according

\textsuperscript{34} Johnson (1988)
to Johnson, in the area of moral authority. Raz claims that the primary way one establishes legitimate authority over another “involves showing that the alleged subject is likely better to comply with reasons which apply to him (other than the alleged authoritative directives) if he accepts the directives of the alleged authority as authoritatively binding and tries to follow them, rather than by trying to follow the reasons which apply to him directly.”

So, on such a conception, an agent should rely upon authority due to the superior ability of the authority to reliably determine what reasons apply to the agent. If our agent had sufficiently good and reliable access to the reasons that apply to her—suppose she is God-like in her knowledge of such things—then he would have the authority, on this account, to bypass the directives of the “authority” in question. But if an agent’s knowledge on the whole regarding morality is inferior to some authority, she does better on the whole to comply with what the authority decides or commands rather than trying to figure out the best action for herself.

This understanding of authority as based on superiority of epistemic access to what reasons apply to us is not unfamiliar. It can be found, for instance, in the Biblical injunction, made to those of us who are less-than-omniscient, not to make the law, but rather to enforce it, where the law itself is presumably given by God. It is also familiar from the writings of direct consequentialists, who, in an attempt to accommodate certain commonsense intuitions about morality, stress our limited knowledge of causes and effects and our limited powers of moral calculation, and point instead to the good effects of simply falling into line and following established moral rules, or exhibiting conventional virtues (most of the time). Perhaps considerations of epistemic access are also behind the thought, already mentioned, that judges should not “legislate from the bench”. Familiar direct consequentialists who

---

36 See Augustine’s “Confessions”, where this opinion is echoed.
employ this epistemic strategy fail to solve BP and PRR, since they fail to give the appropriate sort of normative independence to the derivatively justified rules, virtues, motives, and so on that are claimed to be authoritative. On such accounts these dependents are merely heuristics. Raz’s theory, like Gauthier’s theory, is fascinating since it employs this same familiar consequentialist strategy, but claims to have escaped its pitfalls.

Raz thinks that directives of a legitimate authority, though derivative of other reasons that the directive is based upon—he calls such decisions “dependent reasons”—themselves provide subjects with reasons for action. Moreover, he thinks this reason is preemptive, in prohibiting a subject from directly appealing to the reasons that serve as its ground. How is this supposed to work? Of all the direct consequentialists considered, Raz is the only philosopher who specifically addresses this question. According to Raz, because an authority’s directives are based or dependent upon other reasons, it makes no sense to count the directives as an additional set of reasons to the ones they are derived from. Instead, Raz claims that the directives replace the reasons they are based upon. Through this replacement/preemption they show themselves at once to be independently normative in a sense, and also able to block a direct appeal to grounding reasons.

Raz’s argument isn’t explicitly addressed to BP or PRR, though Johnson recognizes its potential for capturing the strengths of both consequentialist and deontological theories for handling BP and some version of PRR. Using Raz’s account, one could insist that an authority’s directive is itself a reason for action, while also acknowledging that its

---

37 Though it seems awkward to call a directive itself a reason, rather than something that provides a reason.


normativity derives from other, deeper utilitarian considerations. The independence of these directives, combined with their preemptive force, might help solve the problem related to normative character in the ways I have already discussed. Finally, the independent normativity of authority’s directives combined with preemption apparently would rule out the propriety of bypassing these directives in favor of promoting their grounding reasons.

Raz acknowledges an overarching commitment to what he calls ‘grounding reasons’, which are normative sources of an authority’s directives on his theory, and argues that the best way to act in accordance with these sources is to “go through” a particular intermediary—namely, an authority. Raz has in mind a political authority, while Johnson thinks the argument can be made with respect to a moral authority. Raz’s theory, then—and Johnson’s adapted moral theory as well—are examples of source-committed indispensability accounts.

**iv. Plato’s “goodness in itself”**

In the Republic, Plato aims to show that justice “pays”: he thinks this is true of both states and individuals, but I will focus on his argument as it concerns individuals. Plato aims to show that it is “better in every way to be just than unjust” (357b). The argumentative agenda is largely set by a pointed challenge by Glaucon, one of Socrates’ interlocutors, to show that justice is not good only for the reputation it usually brings, but also “good in itself”. Plato wants to demonstrate that one always has a reason to be just, and if justice is good in itself then presumably one will always have such a reason (and moreover, he thinks,
an overriding reason\textsuperscript{41}. The idea is just that if justice is good in itself, then justice is good wherever it goes or in whatever circumstances that might hold—its goodness is “self-contained”, as it were. Assuming that goodness provides reasons, one would always, then, have a reason to be just. Further, if justice is good in itself then whatever good that justice has would appear to be attainable only by being just. No shortcuts—such as pretending to be just while practicing injustice—would allow one to possess the goodness of justice if justice has its goodness somehow internal to it.

We might ask what would be required to show that one always has a reason to be just. One option is Moorean: on this account, justice is good strictly in virtue of its intrinsic properties. If the Moorean interpretation is right, justice’s goodness would be self-contained, as it would be good wherever it went—none of its goodness would depend upon its relationship to variable and potentially unfavorable circumstances. Its goodness would supervene on its intrinsic properties and not some external thing (like the reputation of being just) that one could acquire, presumably, though clever acts of injustice.

But showing that justice is intrinsically good in this sense is more than is needed for the purpose of answering Glaucon’s challenge. This leads us to a second option. All that Plato would need to show in order to prove that one always has a reason to be just is that justice always leads to something good—no matter what the circumstances—and that the only way to get this further (external) good is through justice. Justice, then, would be functionally equivalent to something intrinsically good.

\footnote{Note that one will always have an overriding reason to be just only if justice “in itself” is so valuable as to outweigh any good that might come from the good consequences of being unjust, or any goodness that injustice might have in itself.}
Having set out two different ways Plato might answer Glaucon, we can ask which, if either, of these options is taken. It is fairly clear that Plato does not take the first option. That is, in arguing that justice is “good in itself” he does not mean justice is “good in isolation of every other thing” or “good as a result of its intrinsic properties alone”. My interpretation on this point is suggested, at least, at the outset of the Republic, where Socrates never questions the assumption of Glaucon and Adiemantis that in order to defend justice as good in itself it would be sufficient to show that the just person is happy. In fact, Socrates seems to take this assumption on board as central to his project, a fact that becomes evident at later stages in the Republic where the positive argument for justice gets presented. In accepting this assumption, Socrates apparently also accepts the possibility that justice would be good in itself even if it is good only because it unfailingly and indispensably leads to happiness. I suppose, from what I have said so far, it could turn out that justice is identical with happiness. For in this case justice would be required for happiness, simply because they are the same thing. But this would be a strange understanding of the relationship between two evidently separate things, and in any case the text does not support this interpretation. In fact, I argue that the more natural understanding of the relationship between justice and happiness as that of a relationship between separate things is consistent with Socrates’ own understanding of this relationship.

Consider the way Glaucon puts his challenge. Socrates is charged with showing that justice is, like “seeing, hearing, knowing” a thing that is “fruitful” by its own nature. The language here is telling. Something that is “fruitful” is good because of the “fruit” that it bears, not simply because of its own nature: the fruit is separate from that which bears the

---

42 367d Republic.
fruit. Socrates is not asked to show that justice is itself a “fruit”, as it were. The “fruit” in question, as assumed from the outset of the Republic, is instead happiness, and justice its sustaining source.

There is a further point to note about how this challenge is put. Socrates is charged with showing that justice is fruitful, and fruitful “because of its very self”. Justice counts as good in itself just as long as “because of its very self”, as opposed to its mere reputation, it leads to something else that is good, namely, the happiness of its possessor. Justice can be good because of what it leads to and count as good in itself, then, as long as its connection to some further good thing comes from itself rather than being mediated by its reputation.

This interpretation of Plato’s use of “good in itself” is supported not only by the way the central challenge of the Republic is posed and apparently accepted at the outset by Socrates, but also by the way Socrates attempts to answer the challenge. The final argument in Book IX, where he attempts to address this challenge is, according to Socrates, the “greatest and most decisive of overthrows”. But what Plato counts as a successful response to Glaucon’s challenge to shows that what Plato means by “good in itself” is very different from what Moore meant by the same expression. In this argument, Plato attempts to establish that the pleasures of the just person are “truer” and more “trustworthy” than those of others (these are the pleasures, presumably, of experiencing the form of the good, as well as the other pleasures suited to each part of the soul). Justice in the soul (rather than the mere reputation of being just) is claimed to allow for the experience of these valuable pleasures; the goodness of justice gets defended in terms of the good it (and it alone) allows, not in terms of the bare properties of justice.
Like the other philosophers considered in this section, then, Plato defends a normative dependent (justice) as an indispensable means to its source (justice), and a reliable one at that. The just person may not always experience the highest of the pleasures—the pleasures that come from seeing the form of the good—but the just person, and only the just person, will always experience the kinds of pleasures that are suited to each part of his soul.

In attempting to show that justice is an indispensable and reliable means to some further thing of value, his account is structurally like other “indispensability” accounts considered in this section. His account is unique, though, in making explicit that it must be the intrinsic rather than the extrinsic properties of his normative dependent—justice—that connect it to its source. Justice must be fruitful by its own nature. Stressing this at least gives him a rhetorical advantage over some others who would like to show that some dependent or other is “good in itself” because it is an indispensable and reliable means to some valuable end. Whether it gives him a legitimate, non-rhetorical advantage over other accounts of the same type is a further question that I will address in the next chapter.

On Plato’s picture of justice, there is no way that one could, through some action, attain the good that is justice’s normative source—again, certain appropriate pleasures—without justice in the soul. For Plato, an action is just if and only if it preserves or expresses justice in the soul. The normative dependent justice might seem to determine, by itself, the moral status of actions. If this is true of just actions, it is a feat that Plato accomplishes without invoking ‘preemption’. For Plato never denies that what makes justice good is its essential connection to happiness. But because happiness cannot in practice be pried apart from justice, acting in accordance with justice ipso facto is to act, in every instance, in a way that accords with happiness. Justice’s normative source does not threaten to undermine its
dependent in the same way that normative sources do for, say, the utilitarian and the Hobbesian. For all utilitarians, direct or indirect, it is granted *that an individual act* might maximize happiness—where happiness is the normative source—and fail to comply with a normatively dependent moral rule or moral virtue. For the Hobbesian, it is clearly possible that *an individual act* might maximize expected utility and yet fail to conform to the (dependent) moral disposition. But for Plato, it is not even possible that an act might produce happiness as he conceives it without at the same time being a just act. Justice, for Plato, is required in the soul if one is to experience the sort of true and reliable happiness that serves as the normative source of justice. As a result, Plato does not need to “protect” individual actions from the normative source of justice, since justice and its normative source are always, with every action, in agreement. No other philosopher we have considered yet attempts an identical strategy—no other philosopher considered, that is, claims that in order to promote the relevant normative source, *in every action* one must “respect” (by following or exhibiting or promoting) the relevant normative dependent as a matter of conceptual necessity.

Plato’s moral theory is all the stronger for avoiding ‘preemption’, which so far seems a mysterious and troubled notion. But with this benefit comes a burden. Other theories evidently appeal to preemption in order to make sense of normative character change—the hope, again, is that the normative character of the source does not bleed through to its dependents, freeing the dependent to have a different, and appropriate, normative character. This hope may be misplaced, of course; but without it or some other device Plato would seem stuck with the conclusion that the normative character of actions is determined by the normative character of Platonic happiness. There is nothing “between” the two that might
inhibit this. Fortunately, Plato perhaps less than most other teleologists is in obvious need of “character change” in his theory. The character of Platonic happiness is not inappropriately moralistic for the purpose of charging justice with normativity, and neither is it objectionably vulgar or obviously non-moral. Instead, justice’s source is spoken of as a sort of noble pleasure and the greatest good that one can attain. On the default understanding of how dependents receive their normativity, then, justice has a normative character that seems to suit its (moral) nature.

C. Teleological dispensability accounts

Some contemporary philosophers have attempted to draw a distinction between intrinsic value and what has been termed “final value”. It has also been suggested that intrinsic value can be separated from a thing having value “for its own sake” or having value “as an end”. Even more surprisingly, some have claimed that a thing can have “intrinsic value” \textit{in virtue} of being \textit{instrumentally} valuable, even where the instrument is not indispensable in promoting its apparently valuable end.

An example used to pump the intuition that this last contradictory claim is at least \textit{intelligible} is given by Shelly Kagan\textsuperscript{43}. The pen used by Abraham Lincoln to sign the Emancipation Proclamation, does not, probably, have much, if any, \textit{intrinsic} value\textsuperscript{44}. That is, it does not have its special value in virtue of the properties that are intrinsic to \textit{it}, the \textit{pen}, considered as a bare physical object. But one might think that, given the causal role it played

\textsuperscript{43} Kagan (1998).

\textsuperscript{44} I should note that Kagan does not obviously mean to claim the pen has value in virtue of properties intrinsic to it; rather, he seems to mean that the pen is worthy of value ‘for its own sake’, where this last phrase is given less interpretation than one might wish for (288).
in an important event in American history, the pen is nevertheless valuable “for its own sake”\textsuperscript{45}. The suggestion is that it is valuable for its own sake because of extrinsic features—namely, the relationship it bears to certain historical events and figures. Still, there is an air of paradox to these claims—what are they supposed to mean? Kagan writes “…in at least some cases, by virtue of the very fact that the object is (or was, or will be) worth having for the sake of something else, it is also worth having for its own sake as well.”\textsuperscript{46} Put in terms of the NP model I am developing, the pen would be the progeny of certain historically important or valuable events and people, or perhaps also the valuable state of affairs that it played a causal role in bringing about (namely, greater freedom for African Americans).

For now, at least, Kagan’s claims still sound paradoxical. Leaving worries of coherence aside until the next chapter, it is worth seeing what the upshot for morality would be if there existed things valuable for their own sake that depended for this status on their being valuable for the sake of something else. (Kagan himself is more concerned to defend the intelligibility of a certain logical space rather than defend a model of normativity as legitimate, or as a model that, if correct, would be useful in important ways. But I am of course interested in those questions). Now, we don’t know much about what is involved in a thing having “value for its own sake”. Kagan says frustratingly little on the subject, and instead notes that in a number of cases it seems appropriate to call things “valuable for their own sake” even when their value is also clearly dependent. But let us be optimistic for a

\textsuperscript{45} Kagan prefers the term “intrinsic value” for items like Abraham Lincoln’s pen, but because this term denotes an important ontological property I wish to distinguish from his revisionist conception of intrinsic value, I will not adopt his locution.

\textsuperscript{46} Kagan (1998).
moment and suppose that if something is valuable for its own sake because it is valuable for the sake of something else then it has *inherited normativity*.

Consider how a moral theory like classical act-utilitarianism could make use of Kagan’s suggestion, so interpreted. According to classical act-utilitarianism, happiness—and only happiness—has intrinsic value. Someone committed to this account of intrinsic value could appeal to a plausible story about the causal relationship between the exhibition of the moral virtues and happiness. Namely, on this account the exhibition of moral virtue leads very often to the maximization of happiness. It has done so regularly in the past, and it will likely continue to do so. Act-utilitarians have a famously difficult time showing that anything other than happiness has the sort of normativity that always provides reasons. But with the account of value I have been discussing in hand, it could be argued that virtue (its possession or maybe its exhibition) possesses this sort of normativity. If this is right, then one could be committed to the act-utilitarian account of what has *intrinsic* value (in the Moorean sense), and consistently claim that we have reason to be just or generous or charitable etc. even when the maximization of utility is not served (or in other words, one could consistently claim that we have a reason not to bypass being just or generous or charitable, etc.).

Depending on just how ‘value for its own sake’ gets cashed out, it might prove a useful concept for intensional as well as extensional problems that arise in moral justification. To determine just how useful Kagan’s suggestion might be, an analysis of what is meant, and what can properly be meant, by ‘valuable for its own sake’ must be given. This will come in the next chapter.
D. Non-hypothetical agent-election

Korsgaard presents a version of Kant according to which Kant makes use of the NP model in accounting for the value of things other than that of humanity, the latter of which Korsgaard takes to have intrinsic value. Whether Korsgaard is correct in so characterizing Kant may be disputed. However, I will leave interpretive worries aside and focus on the features—and later, the merits and limitations—of Kosrgaard’s account irrespective of its fidelity to Kant’s views.

Korsgaard’s approach differs markedly from the others considered so far, and it should soon be clear just how—but allow me to start with the similarities. Korsgaard is concerned to show how certain objects can be “finally valuable” or “final ends” (she uses both of these expressions) without being intrinsically good. That is, she would like to show how certain objects can have a special, robust sort of normativity even if the source of their value is external to them. In this aim, she is like Rawls, Gauthier, or Kagan, since each of these philosophers claimed that a dependent could receive a special and (in some sense) independent sort of normativity from an external source. But her argument for this conclusion is unique. Following Kant, she regards ‘humanity’, what she interprets as “the capacity to set an end”, to be the only thing that is intrinsically good47 (in the Moorean sense of having value just in virtue of its intrinsic properties). She argues that there nevertheless plenty of other things that have final value, and that things with final value get this value from their relationship to humanity. The argument for this claim, in brief, is as follows: When we set ends, we regard these ends as good. But to regard our ends as good commits us to regarding ourselves as capable of conferring value upon these ends—in other words, it

---

commits us to regarding ourselves as a source of value. But so far only our own ends are valuable, and without a further premise we would be mired in kind of subjectivism. So, again taking her cues from Kant, she points out that since every other rational being must regard themselves in the same way, we must regard them also as sources of value, as capable of conferring value upon ends they rationally choose. And so value is determined not just by us, but by others as well.48 This means, according to Korsgaard, that the objects of every rational person’s choice are “objectively valuable”. By this she means not that such objects are intrinsically valuable, but rather that that every rational being has a reason to promote or realize these objects or ends.49

Korsgaard’s argument to the conclusion that humanity is the source of normativity is, at every step, controversial50. I will not focus on the merits of her argument, however, and instead concentrate on what she takes to be the relationship between normative sources and the objects upon which they are alleged to confer value. In describing this source/dependent relationship, Korsgaard uses language that is similar in its imagery to the language I have used to describe the relationship between normative sources and their progeny. For example, she writes that value “flows from” humanity to its objects51. Humanity is the normative source of the objects of rational choice. Because these objects depend upon humanity for their normativity; Korsgaard calls their value conditional. Still, granting that there exist


50 See, for instance, G.A. Cohen’s response to Korsgaard (1996)

51 Korsgaard (1996) 269. Her use of this sort of imagery has provoked charges that her view is ontologically over-committed, though she insists this is not the case. In recent writings she has backed away from the metaphysically suggestive imagery of “transfer” (Korsgaard [1998]).
rational beings who set ends, every rational being necessarily has a reason to promote or foster the rational ends they set.

Ends seem “independent” of their source in virtue of this objectivity: they are valuable independently of their being valued by me, you, or any other rational being—as long as one rational being rationally values it. Moreover, these objects of rational choice have a nomological similarity to things that have intrinsic value. In calling them “ends”, “valuable as ends”, “final goods” and things of “final value” Korsgaard, fairly or not, brings to mind the sort of value we typically reserve for those things valuable in themselves in a more familiar way.

The particular genus of the NP model Korsgaard’s theory makes use of may be evident already. Since dependents get their value from being the objects of rational choice, her account is an instance of an agent-election account. Further, her account appeals to actual rather than hypothetical agreement, a fact that has significant substantive implications. With the exception of rational nature itself (which is ‘objectively good’ unconditionally, according to Korsgaard\textsuperscript{52}) something is not objectively good on her view unless it is actually rationally valued. Rather than viewing this as a liability of her account, Korsgaard seems to regard its circumstance-sensitivity as a strength, since it gives (what she considers to be) intuitively good results in cases that traditional notions of value stumble over. She asks: If a beautiful painting is locked up with no one to view it, is it valuable? It seems, she says, that a painting is only good on the condition that it is viewed and valued. Since the

\textsuperscript{52} Korsgaard (1983) 260.
conditions are not met in the example of the locked up painting, it is not finally valuable or objectively good\textsuperscript{53}.

Korsgaard is not obviously concerned with solving BP or PRR, though her theory has resources for making headway on both fronts. To illustrate, let us revisit the difficulty Kant has in defending the importance of our own happiness. He offers an instrumentalist defense in the *Groundwork*: “To assure one’s own happiness is a duty (at least indirectly); for, want of satisfaction with one’s condition, under pressure from many anxieties and amid unsatisfied needs, could easily become a great temptation to transgression of duty.”\textsuperscript{54} Unfortunately, very often happiness does not help one in fulfilling one’s duty. Even if a certain small amount of happiness is helpful in fulfilling one’s duties, one might hope that our responsibility to promote our own happiness extends beyond sustaining that minimum level of welfare—whatever that may be. We might also think that we have less dreary, moralistic reasons for promoting our happiness than those Kant provides. Intuitively, happiness matters in part because it is, in itself, valuable; but this is something that Kant, on traditional interpretations, cannot say.

And neither can Korsgaard—but she can come close, at least. On her account happiness is *finally and objectively good*—so long, that is, as it is the object of rational choice. If Kant is right that every rational being necessarily desires her happiness, then Korsgaard can help Kant preserve the intuitions regarding the extension of our reasons to promote our own happiness. Her account also looks like it might provide us with better reasons for promoting ends like happiness. Happiness and other objects of rational choice

\textsuperscript{53} Korsgaard (1983) 264, 265.

\textsuperscript{54} Kant’s *Groundwork* 4: 399. (trans. Korsgaard).
are not worth promoting just because they are instrumentally useful in promoting rational nature, or conformity to duty, but (also) because they are finally valuable.

Other examples could be given to show the potential usefulness of Korsgaard’s account of final value. Kant has a difficult time defending the humane treatment of non-rational animals. He argues that we should treat animals well because the mistreatment of animals might cause us to treat rational beings badly. He also claims that we are obligated to treat all living things, and not just animals, with sensitivity to the fact that they are alive, but he argues that this duty is grounded in a general duty not to do things that will harm our own characters. But, again, even if we can be convinced that harming animals will hurt our own characters or dispose us to harming rational beings, many will be dissatisfied with this justification. Korsgaard’s account gives us the resources for providing intuitively better reasons for treating animals humanely. Namely, animals are objectively and finally good (so long as they are rationally valued); and, if pressed as to why, we can say that this value is grounded in their status as objects of rational choice.  

E. Hypothetical agent-election

Thomas Scanlon offers a contractarian moral theory, but one that is different in several important respects from Gauthier’s Hobbesian-style contract theory. One important difference is that, unlike Gauthier, Scanlon—at least in his most recent work—aims only to give an account of interpersonal morality, or as he calls it, morality “narrowly construed”.  

55 Korsgaard offers a different defense of the humane treatment of animals than the one I have offered on her behalf here (Korsgaard [1996] 149-160).

56 That is, he aims only to articulate and defend the obligations of rational agents to one another, leaving aside any obligations rational agents might have to those things that lack rationality (like plants and animals), and also
On the subject of practical reason, Scanlon and Gauthier seem to split along the Kantian/Humean divide: where Gauthier grounds moral obligation on desire, Scanlon appeals instead to reason. Their audience seems to differ as well: Gauthier seems to be writing for the moral skeptic, while Scanlon has in mind someone who is already concerned to be moral. So, while Gauthier aims to justify morality on non-moral grounds, Scanlon does not hesitate to rely upon moral premises. He assumes that we have an interest in being moral—his goal is to articulate just what the content of morality is. Finally, where being the object of rational choice does not \textit{in itself} confer normativity upon an action on Gauthier’s account, but is instead normative \textit{only} because objects of rational choice promote an end of value, ‘choiceworthiness’ for Scanlon is not normative simply in virtue of its instrumental value. Instead, things are right and wrong in virtue of being rationally choice-worthy—or, more precisely, they are \textit{wrong} in virtue of being rationally ruled out as not worthy of choice.

He claims in “What We Owe to Each Other” to be offering an account of the property of moral wrongness\textsuperscript{57}. According to Scanlon, “an act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject as a basis for informed, unforced general agreement.”\textsuperscript{58} Now, Scanlon’s apparent commitment to the NP model is not immediately obvious from just this formula, but can be brought out by considering his best response to a familiar objection raised against his theory (and others like it).

\footnotesize{leaving aside morally questionable but consensual behavior between rational agents that does not affect any other rational agents.}

\textsuperscript{57} Scanlon (1998) 12 and n. 21.

\textsuperscript{58} Scanlon (1998) 153.
The charge is that his account of wrongness is redundant, a fifth wheel that is inert so far as justification is concerned. The most forceful argument to support this objection is as follows: if moral wrongness is understood as reasonable rejectability, then we ought to consider what kind of reasons could be given to show that something is reasonably rejectable. But, the objection goes, if Scanlon aims here to give an account of moral wrongness, then the reasons for rejecting some action or principle (or whatever) that meets this standard must already be moral reasons. For how could non-moral reasons possibly ground a Scanlonian reasonable rejection, the latter of which is supposed to be an account of moral wrongness? Michael Ridge gives the operative principle here the following slogan: “non-moral reasons-in/non-moral-reasons-out”.

So, continues the objection, Scanlon must grant that the reasons that ground reasonable rejectability are themselves moral reasons. But if that is the case, then he is caught in a dilemma: it seems that the Scanlonian contractual apparatus must be either circular or redundant. It is circular if it makes use of the concept of moral wrongness in order to explain moral wrongness. More precisely, if Scanlon claims that reasonable rejectability is constitutive of moral wrongness, and that some set of principles is reasonably rejected if and only if they are morally wrong, then he has helped himself in a circular fashion to a concept (that of moral wrongness) he is not entitled to. If, on the other hand, he claims that some set of principles is reasonably rejectable if and only if there are moral reasons for rejecting it (without mentioning moral wrongness specifically), then it is not clear why these moral reasons are not by themselves sufficient to explain moral wrongness.

59 See, for instance, Blackburn (1999), McGinn (1999), and Pettit (1993, 1999) each of whom raise a version of this objection.

60 Ridge (2001).
Scanlon could avoid this dilemma by returning to the previous stage of the dialectic. If the “character change” aspect of NP model is defensible, Scanlon need not automatically accept the “non-moral reasons in/non-moral reasons out” principle. The reasons that ground reasonably rejectability might be non-moral in character, but their normativity might nevertheless wind up moral at the level of the contractualist principle. If ‘character change’ were defensible, then moral wrongness could depend on non-moral reasons for its normativity. Just how this might happen would need to be explained at length, of course—in part because, as we have seen, the presumptively right understanding of normative character does not have room for character change. The point, then, is not that Scanlon can use the NP model to vindicate his theory in one fell swoop, but rather that the NP model might provide space to defend his theory that some of his opponents have overlooked and which could get him out of this particular argumentative corner. The NP model shows promise, then, in addressing a common and formidable objection to Scanlonian contractualism (and other versions of Kantian contractualism as well).
Chapter 5: Analysis

I. Overview and methodology

In this chapter I will evaluate the various attempts at employing the NP model that I discussed in the last chapter. Each of these attempts, I argue, fails to solve BP and PRR. Some of these attempts fail in a dramatic and misguided manner. Others approach a successful instantiation of the model, but come up short. For each case I discuss, I will be primarily concerned to see if the theory has the resources to defend the NP model of normativity, or, barring this, to see if a theory very much like it might have such resources. I will not discuss substantive details of each theory unless they appear to be critical to the success of theories of its kind in instantiating the model. For the hope is to find a successful general approach to NP rather than a particular substantive theory that makes the cut. To this end I will rule out theories based on the basic strategy they employ and general picture of normativity they present. An analysis of the different attempts considered here will come at the end of the chapter, and from their failures a list of conditions for success emerges. I am hopeful that these conditions can point the way to a more promising approach to instantiating the model.

II. NP accounts analyzed

A. Teleological, qualified source-commitment

i. Rawls’ Rule-Utilitarianism

According to Rawls, though the principle of utility is normative, its normative reach is circumscribed by the logical relationship between practices and the acts that fall under them. The argument for limiting the scope of the principle of utility he provides is, however,
unsound. More importantly, though, even if this argument were sound, the picture of normativity it attempts to defend would have extensional and intensional problems of the sort that characterize BP and PRR.

First let us consider the argument. Rawls attempts to limit the scope of the normativity of the principle of utility by appealing to a notion of logical priority. Recall that rules of morality are ‘logically prior’ to the acts that fall under them, in that (certain) moral actions (such as punishment and promise-keeping) cannot be described without reference to (certain) rules of morality (such as the rules governing punishment and promise-keeping). This means that the principle of utility cannot be directly applied to individual acts that fall under the rules governing punishment and promise-keeping. It seems, though, that Rawls’ confuses ‘logical priority’ with justificatory priority, or at least seems to mistakenly hold that there exists an entailment relationship between them. He holds that if some x is logically prior to some y then it is also prior to y with regards to justification. (For some x to be prior to y with regards to justification, as I am thinking of it, x would justify y and not vice versa).

But why think this? Just because some act cannot be specified independently of the practice it is a part of does not, without further premises, establish that an act cannot be justified independently of the practice it falls under. If justificatory and logical priority are disentangled, it becomes clear that the entailment claim requires an independent defense that Rawls does not provide.

The ‘logical priority’ of practices over the actions that fall under them is also supposed to yield the feature of preemption. The thought must be that if some x cannot be specified without reference to some y, then x cannot be justified directly by anything other
than y. So not only are acts that fall under the practice of punishment governed by the rules that constitute the practice, but they are governed only by these rules.

Rawls’ commitment to these entailment relationships between logical and justificatory priority gain some of their plausibility from his appeal to games like chess and baseball, which are supposed to parallel the structure of morality. These examples as well, though, confuse in an important way the relationship between the logical structure of a rule-governed practice and the scope of the normativity of its constitutive rules. He considers questions such as “why should I keep a promise in this situation?” to be of a piece with “why shouldn’t I be allowed four strikes instead of three in this situation?” The latter question seems absurd, and so, he reasons, is the first. An important relevant difference between baseball and morality, however, is that:

1. the point of games like baseball is almost always best served by complying with its rules, where this is not true to the same degree of morality, and;
2. the “stakes are lower” in the games like baseball than they are in morality.

I will defend these clams in turn.

(1) What I mean here can be brought out by thinking about the normative sources of morality and baseball, respectively. The normative source of justified moral rules is the principle of utility. What is the normative source of the rules of baseball? To answer this, we should think about what might be the point of the rules of baseball. The point or points of the rules of baseball seems to be something along the lines of making the game physically challenging, fair, and enjoyable for the players, and enjoyable for spectators. But allowing one player four strikes instead of three in a particular instance would not serve the point(s) of baseball, so construed. It would be unfair, obviously, to the fielding team. It would make the
players on the fielding team upset, presumably. And though it might momentarily cheer the batting team, they would not take as much enjoyment in a victory if the rules were altered in this way. Making other changes to the rules on a case-by-case basis would almost always have this same effect.

What about cases in which changing the rules of baseball does in fact serve their point? In these cases, changing the rules does make sense. If a neighborhood game of baseball is played where one team has a player that has never played baseball, it makes sense to allow her four strikes instead of three. It makes sense because it will make the game, on the whole, more enjoyable, fairer, and probably more enjoyable for the spectators (if there are any). So, it makes sense to apply the normative source of the rules of baseball directly to the actions that are logically “posteriori” to these rules. This suggests that the normative source of baseball directly justifies the actions of the game also in standard cases, and the appearance to the contrary is a just a result of the nearly perfect coincidence of following the rules of baseball and comporting with their normative source.

(2) An additional confusing feature of games like baseball is that the “stakes are low”, in that whether one complies with the rules of such games is, under standard conditions, not of great importance, at least relative to the importance of whether one follows the rules of morality. Our normative sensibilities register the normativity in baseball, under standard conditions, only dimly. As a result, the sensibility of breaking rules in light of their failure to promote their normative source is not as clear as it could be. If the stakes are raised sufficiently, however, it becomes clear that a direct appeal to normative sources is appropriate. If a batter will try to kill every member of the opposing team unless she gets a
hit, it makes sense to allow her four strikes (at least!). Appealing directly to the normative source of the rules of baseball is the thing to do.

The thought that Rawls’ logical point could not possibly have the substantive implications he thinks it does is hardly new. Despite the problems with Rawls’ argument, perhaps it will be thought that there is some way or other to rescue it and so defend more convincingly a version of rule utilitarianism that possesses a two-level justificatory structure.

For the sake of argument I will embrace this optimistic view, and look more abstractly at the picture of morality Rawls presents us with. I will begin by examining the alleged preemptive force of rules. As Rawls has it, the principle of utility cannot be directly applied to acts that fall under a moral practice. Though the point of this feature is to save certain extensional and intensional intuitions, it seems that it goes too far in these regards; though soothing certain of our intuitions, it chafes badly against others through an excess of rigidity.

The problem is perhaps clearest in the intensional case. Consider what appears to be a moral dilemma. As a news reporter I have promised some person not to reveal his identity, in exchange for certain newsworthy information he will give me. It turns out that revealing his identity would lead to large gains in net utility. Imagine that my source has told me that he is the leader of a group that is responsible for harassing members of a certain ethnic background, and that he plans on doing more of the same in the future. It may not be obvious to everyone just what, as a reporter, I should do in the case. Should I let some relevant legal authority know the identity of my source, so that he might be prevented from continuing his harmful ways? Or should I keep my promise to protect his identity? Whatever the answer, to construe the case as one in which there is no genuine moral tension
(and not just no epistemological tension\textsuperscript{1}) between breaking the promise and keeping the promise is seemingly to oversimplify it. Perhaps I should keep my promise. But then isn’t that \textit{in spite} of the reasons for breaking it? More generally, it seems that the utility that grounds certain moral rules in fact applies to cases that fall under these rules, such that even if they are often (or always) defeated, they are, like Ross’s (unhappily named) \textit{prima facie} duties, \textit{overwhelmed} rather than “killed off”.

One might be tempted to think that Rawls at least has the resources to show why the normative character of certain moral rules would not have the non-moral character of their source, the principle of utility. According to Rawls, when justifying some act one need only appeal to the relevant rule or rules that govern the act. Because these rules would appear to be rules of morality because of what they require—and because in justifying a particular act that falls under them no explicit reference need be made to the principle of utility—they might appear to have a \textit{moral} character, or at least have a character not given by the principle of utility. But, as I earlier, the normative character of any normative entity is, presumptively, given by whatever makes it normative. Even if the principle of utility perhaps does not need to be \textit{mentioned} when explaining why some act that falls under some rule is right or wrong, the principle of utility nevertheless justifies that rule—it makes it normative. It would be strange to think that a rule could have a non-moral character in one context—in the context of justifying a practice—and a moral character in another—the context of justifying acts that fall under its governance. Regardless of the context, the rule is the same, and so is the two-level structure of justification. What varies by context is just the justification that is appealed

\textsuperscript{1} What I mean is this: if two courses of action are in ‘epistemological tension’, then it is not \textit{clear} which course of action should be taken. This is consistent with there being no reason other than those provided by epistemological considerations to recommend one or the other course of action.
So the normative character of the principle of utility should “bleed through” to the normative character of the rules justified by it.

Despite these problems, Rawls’ theory is admirable in its ambition, and to my mind has at least one advantage over almost every other version of indirect consequentialism. Consider what Rawls does not try to do. He does not, unlike most indirect consequentialists, try to justify moral rules as normative just because following them (or on other versions, promulgating them, following them for the most part, etc.) will lead to better consequences than doing anything else. Rawls is open to the possibility that all moral agents with the aim of maximizing good consequences on the whole might serve this end best by attempting to straightforwardly maximize. This means that, for Rawls, the strict adherence to moral rules is not just a means to the end of maximizing good consequences. Where a normative dependent is merely a means for promoting its source, that dependent would not have the kind of normative independence, seemingly, to satisfy PRR. If asked “why ought I keep this promise?”, one bad answer is that keeping promises maximizes good consequences, and good consequences ought to be maximized. Instead, Rawls—supposing his view were otherwise defensible—has a more complicated, and intuitively better, response. He will grant that one ought to keep a particular promise because rules that constitute the best practice require keeping the promise. And if pressed on the question of what makes a practice better than another the answer is given in terms of the practice’s utility-conduciveness. So far this may sound like practices and the rules that constitute them may just be means to the end of utility maximization. But here is the crucial point: for Rawls, the constitutive rules of a practice do not, as a matter of logic, admit of a direct appeal to utility when justifying an act. Rawls can legitimately say that the very nature of moral rules precludes such a direct appeal.
This is an answer that, if defensible, gives rules a kind of independence that is intuitively satisfying; rival indirect accounts, through their reduction of rules to mere instruments, inevitably come up short.

Rawls’ rule-utilitarianism is admirable for another reason as well. His approach, if successful, would establish a *conceptual* connection between morality (moral rules) and particular actions, such that one of *conceptual necessity* has reason to act morally. (So long as a particular rule was justified, then the actions that fell under the rule were necessarily justified as well). This conceptual connection does away with the contingency that plagues most teleological theories, ensuring that one always has a reason to act morally. Perhaps this kind of connection is required for solving BP on a teleological account. Otherwise, the connection between morality and its source is contingent, and morality’s normativity would be vulnerable to shifting circumstances.

**B. Teleological, source-committed indispensability**

1. **Hare’s two level theory**

Recall that Hare argued that, because of our epistemic limitations, we are very often not in a position to act as self-conscious act-utilitarians, even though he thinks act-utilitarianism is the correct criterion of the right. This is because we are often simply unable to determine what actions are required if we are to maximize utility. So instead in these cases we ought to fall back upon the rules of commonsense morality. Doing so, Hare argues, will very often be wiser—from the standpoint of the maximization of utility—than directly attempting to satisfy the act-utilitarian principle.
Hare gets us closer to a legitimate NP model than does classical act-utilitarianism, but still comes up short. One reason for this is simply that, on Hare’s account, various evidently important parts of morality may sometimes be circumvented. This is because, as Hare acknowledges, sometimes the requirements of act-utilitarianism and the rules of commonsense morality conflict, and when this is the case—provided one is in the proper epistemic position to judge what act-utilitarianism requires—one ought to act in accordance with the demands of act-utilitarianism. The rules of commonsense or intuitive morality are thus seen to be merely very useful, or, if one insists, extremely useful, heuristics, instead of being rules that one has a reason to follow in every case. The point of following the rules is to maximize utility, and Hare thinks that the best way of maximizing utility is to follow these rules. Thus, the justification of any particular right act must be: that act is right because it is allowed by a rule that, if adhered to, will produce more utility than if one did anything else (such as attempting to directly maximize). Unlike Rawls, then, Hare has an overarching commitment to utility maximization, and the rules of morality are mere means to this end. Where Rawls can block the direct justificatory appeal to the principle of utility, Hare can only block the appeal as a matter of practical decision-making—the normative justification of any act comes from the principle of utility itself. This makes his theory open to BP, since if one’s epistemic status is improved such that one is able to have the moral vision to see what would maximize utility, then one could maximize utility and so ignore the rules of commonsense morality that classical utilitarianism famously transgresses.

Even if we grant Hare his empirical claims, his theory would, in any case, give us the wrong reasons for being moral for many of the same reasons that the other theories we have considered give us the wrong reasons for being moral. The deep, ‘critical’ level of
justification—which is all PRR is concerned with\(^2\)—infuses morality with its own non-moral character. And there is no other satisfying sense in which the rules of commonsense morality are independently normative—they ought only to be treated as such.

\[\textit{ii. Gauthier’s Binding Consequentialism}\]

Gauthier claims that our moral duties are grounded ultimately in the rationality of maximizing expected utility, where utility is a measure of the satisfaction of one’s own preferences. Gauthier argues that the rationality of maximizing expected utility makes it rational to adopt a disposition that bears some similarities to conventional moral virtue. We can expect to do better from the standpoint of utility maximization, he thinks, if we adopt such a disposition, and, importantly, bind ourselves to it. One surprising upshot, he argues, is that such a disposition is necessarily normative, in that it gives us a reason to perform the “moral” actions that fall under it in every circumstance. Unlike Hare, who acknowledges that if we are in the proper epistemic position to judge that an act will maximize utility (on the whole, over the long haul) even though it conflicts with conventional morality, we ought to maximize utility, Gauthier holds the bold view that even under such circumstances one ought to perform that action that falls under the quasi-moral disposition. This is even the case where one knows one has no risk of being found and suffering the repercussions of being so exposed.

The obvious objection here, mentioned in the previous chapter, is: fine, the most rational disposition to adopt might be moral in nature, but surely the most rational action will always be that action that maximizes one’s expected utility. And when the most rational action and the most rational disposition are in conflict in some individual case, surely the

\(^2\) See Ch. 1.
most rational thing to do is to do whatever will actually maximize expected utility. Gauthier claims that rationality requires maximizing one’s expected utility, and in these cases of conflict doing this would seem to require performing the most rational action. Gauthier’s response is that because sincere dispositions to be moral preclude acting in ways that don’t express that disposition—so long as the relevant circumstances have not changed since choosing that disposition— it makes no sense to evaluate the rationality of actions independently of their controlling disposition.

Gauthier’s argument is intriguing, but our intuitive resistance to it turns out to be well founded. The major problem with his argument is that he fails to give us a good reason for privileging dispositions when assessing the rationality of particular actions. One can conceive of circumstances under which privileging dispositions in this way would make sense. But these circumstances do not, in the actual world, obtain for most of us. It would only make sense to bind oneself to a disposition if one had reason to think that one’s ability to respond in morally appropriate ways to different circumstances based on moral information gathered at that time was sufficiently compromised. The intuitive idea is that one has a unique sort of access to the relevant moral information in the choosing of dispositions, but one has no such access, or sufficiently poor access, on a case-by-case basis, such that one never ought to act just based upon the moral information one has gathered in a particular case. A less extreme version of this view is shared by many utilitarians, who argue, like Mill and Hare, that conventional moral rules ought very often to be relied upon in choosing actions to perform. They argue, plausibly, that under sufficiently dim epistemic circumstances, one does better, from the standpoint of utility, to act according to

---

3 Gauthier includes this proviso in his discussion of the paradox of deterrence, where he does not appeal to hypothetical agreement. Where the rationality of dispositions is determined instead by hypothetical agreement, such a proviso is obviously unnecessary.
conventional wisdom. But, seemingly sensibly, they have exception clauses for overriding conventional wisdom when one has sufficient epistemic clarity. Gauthier, though, argues for the view that individual actions are wholly insulated from the direct normative force of their source.

Suppose that a captain of a ship is trying to navigate in the midst of a dense fog. The fog is so dense that she cannot see beyond the bow of the ship, and so she decides that she is going to navigate according to radar. The radar, however, cannot detect objects beyond a certain size, such as certain small (but dangerous) rocks. Our captain’s commitment to navigating by radar is sensibly conditional. She will navigate by radar unless the fog lifts enough and reveals some new piece of information relevant to how she ought to navigate (such as the presence of a small rock in her path). It would be foolish for a captain in this position to say to her shipmates “I must maintain the course, because I already decided to navigate only according to radar. Brace for impact!” A more sensible captain leaves open the possibility that she will have a moment of epistemic clarity where the fog will literally part that will justify overriding her commitment to navigating by radar.

Consider a slightly different situation: a captain knows that he is about to go blind. Luckily, he can program into the onboard computer a route for the ship based upon his knowledge of the hazards in the area. This captain acts wisely if, after going blind, he never deviates from the route he programmed when he still had his sight. And he would be foolish if he decided to try to steer by himself because he had a “hunch” that a rock might be in front of him. Whatever epistemic status his hunch might have, it is not as reliable as the information available to him prior to going blind.
For Gauthier’s argument to be convincing, moral agents would seemingly have to be like the blind captain, whose moral beliefs in individual cases will never be sufficiently reliable for guiding action as to justify overriding a commitment to acting according to one’s chosen dispositions. But this is a wildly implausible assumption. Very often there is morally relevant information that is available to a moral agent only in the thick of circumstances, much like the situation of our captain attempting to navigate through patchy fog. Though I might have a general commitment to telling the truth that is informed by an interest in maximizing my expected utility, I can see that in the circumstances at hand that telling the truth will be disastrous from the standpoint of maximizing expected utility. Clinging to the general rule governing truth-telling under such circumstances would be irrational. So he fails to show that normativity can be genuinely inherited, and that the source of normativity is preempted from directly applying to actions. What is more, these examples show that if he were successful in establishing these, his theory would be overly rigid. As in Rawls, ‘preemption’ makes his theory procrustean.

Whatever plausibility Gauthier’s argument has might come partly from a failure to appreciate an important difference between an agent’s ability to determine with great certainty what will maximize expected utility in particular cases, and an agent’s ability to recognize that one is in this unusual position. Suppose one knows that acting in accordance with a certain disposition \( d \) will maximize utility 80 percent of the time. One also knows that one’s own judgments about what is rational to do, when these judgments do not rely upon disposition \( d \), but instead rely upon individual appraisals of one’s circumstances, available actions, and likely outcomes, on the average maximize utility 50 percent of the time. Now, if one does not possess an ability to size up the reliability of one’s judgments about utility-
maximization in particular cases—more specifically, if one is unable to tell when one’s judgments about utility-maximization are more likely, if acted upon, to be more utility-conducive than acting in accordance with disposition d—then it would in fact make sense always to act in accordance with disposition d. One could grant, on this view, that agents are often in a “privileged epistemic position” in particular circumstances in the sense that they often extremely well-placed to judge that some action would maximize utility, where this action does not accord with disposition d. But because one would never be able to tell when and to what extent these judgments were likely to be true, it would still make sense to always act in accordance with the most rational disposition.

But epistemic privilege is not limited in this way; normal agents have the ability to size up in a rough way the reliability of their various moral judgments. A reflective agent might realize, for instance, that she is a bad or unreliable judge of whether she ought to be honest with a person that she has a great dislike for, knowing that her personal feelings threaten to prejudice her judgment on the matter. Or she might realize that she is in a good position to judge whether to keep a promise in a situation in which she has sufficient information about her circumstances that she could likely break the promise without being found out. Given this ability of reflective, responsible agents to appraise the reliability of their judgments concerning the rationality of particular actions, such agents could presumably judge with sufficient accuracy the relative reliability of acting in light of one’s case-by-case appraisals and acting according to the most rational disposition to make a certain amount of “straying” from these dispositions rational. Such an agent would be only provisionally bound to disposition d.
Even if I am wrong and Guathier could solve BP, other problems would remain, including prominently the character version of PRR. Gauthier claims that not only is the straightforward maximization of utility not rational, it also cannot be moral. A necessary condition for acting morally, he claims, is that one constrain the maximization of utility by acting only in accordance with the most rational disposition to adopt. Still, he acknowledges, morality “issues” from the concern to fulfill our individual preferences. The *point* of morality, one could fairly say, is, for Gauthier, to fulfill our individual preferences. But the most plausible account of normative character, I have argued, has it that a thing receives its character from that which justifies it or gives it its point. It may well be true that in deciding what action to perform one need not—and should not--make direct psychological reference to the egoistic principle of utility, and instead should act in accordance with the most rational disposition to have. Still, what gives this disposition reason-giving power—in other words, what *justifies* it—appears to be the egoistic principle of utility. The normative character of this principle, then, would seem to bleed through to the disposition it justifies. The issue is less complicated here than it is in Rawlsian rule-utilitarianism, where there is not an overarching commitment to maximizing utility. Gauthier does have such an overarching commitment, and so the moral disposition on his theory is rightly characterized as a merely indispensable means to an end—an end whose character is non-moral.

The normativity of fulfilling individual preferences is not immoral in character, but it is not moral either; instead, it appears straightforwardly prudential. Constraining its direct maximization as one does when one acts morally does not affect the point of morality, but just gives it a different *extension* than straightforward maximization. And even if the extension of Gauthier’s theory matched that of a rich sort of morality, this would not affect

---

its character. One constrains its maximization only because one can expect to do better from the standpoint of utility maximization if one so binds oneself. So we would need to know why “morality”, which gets its point from the normativity of individual preference satisfaction, does not share its prudential character. To this end, Gauthier does not offer an argument.

**iii. Razian preemptive, dependent reasons**

The response to Raz and Conrad will be in some ways similar to the response to Gauthier and Hare, since their arguments have a similar strategy. Raz is more explicit than Gauthier, however, about the claim that the reason one should not directly try to comply with normative sources is due to epistemic limitations of the agent. For Gauthier, the epistemic premises are buried as assumptions, while for Raz (and Hare) they are explicit and defended. Because of epistemic limitations of the agent, the reasons that directly apply to an agent are not, according to Conrad and Raz, to be self-consciously and directly acted upon. Instead, one does better to act in accordance with reasons that are given from an epistemically privileged position. For Raz, whose topic is the philosophy of law, these reasons are given by some political authority, who is in an epistemically privileged position to ascertain the reasons that apply to individuals. For Conrad, these reasons are given by a moral authority of some kind, perhaps in the form of generally agreed upon moral rules.

I will focus my analysis on Raz, since Conrad’s argument depends on the success of Raz’s defense of special kinds of ‘exclusionary reasons’. A surprising and at least initially puzzling claim made by Raz is that the authoritative directives issued by those with superior epistemic access to the reasons that apply to agents “replace” the reasons they depend upon.
By this Raz means that one has reason to comply with an authority’s directives, and no reason to directly comply with the reasons that ground these directives. This seems strange—after all, an authority’s directives only imperfectly guide us in responding to the grounding reasons their normativity depends upon. And this would seem to make these dependent “reasons” mere heuristics—even if indispensable ones. How could they replace that which they depend upon?

Maybe Raz simply means that, in practice, the directives of an authority should be treated as though independent, and as replacing their normative ground. On this interpretation, the real reasons for action are provided by the grounding reasons themselves in light of which authorities issue directives. Raz might be thought to have at best shown that we have a reason to take authoritative directives as independent, preemptive reasons. But this is not the same as showing that the authoritative directives are actually reasons. Put somewhat differently, one might object that Raz can at most make good on the claim that one’s ‘motivating reasons’ ought to be identical with the directives issued by a relevant authority, but the ‘normative reasons’ for acting are always provided by the reasons the authority herself is attempting to be responsive to.5

If this is all Raz means, or all that he is in any case entitled to claim, then his model of authority and reasons will fail to be of much help in solving the relevant problems for moral justification I have been pressing. Maybe there is more Raz can say in his defense, however. He may be relying upon a normative parallel to the law of conservation of mass; that is, he may think that, everything else being equal, the normative force of a reason cannot be diminished or increased by mere relocation—that is, by moving from a normative source to its dependents. Suggestive of some such background principle is his claim that to count both

5 Smith (1994) 95, 6.
a dependent reason (such as an authoritative directive) and a grounding reason (the reasons that ground an authoritative directive) would be to be guilty of “double-counting”. It is not as though, he writes, there exist *twice as many reasons for acting*, with cumulative weight, once an authority issues a directive based upon other reasons. And this seems intuitively right, and an unmysterious account of how ‘preemption’ might work. Since he thinks we have to choose one or the other, and because there are practical reasons for choosing the authority’s directives as what guides one’s actions, the authority’s directives “preempt” their normative ground.

But Raz cannot establish the preemption thesis, for the same reason that Gauthier cannot show that the normativity of the moral disposition alone directly applies to actions. Authoritative directives would only ‘preempt’ their justifying ground if it always made sense to act in accordance with an authoritative directive. But there is no good reason to think that binding oneself to Raz’s privileged dependents will in fact cause one to be more likely to promote or comply with their normative source than failing to so bind oneself. It may be right that one would do better with respect to complying with “ultimate” reasons if one were to bind oneself to an authority’s directives than if one were to pay no heed at all to an authority’s directives. But these are not the only ways one can act in response to authoritative directives, and to present them as such would be to present a false dilemma. One could, instead, provisionally bind oneself to an authority’s directives, where the proviso for non-compliance is satisfied when one is in an epistemic position to judge that one can better comply with the normative sources if one does not follow the authority’s directives. Like Gauthier, Raz’s ability to establish normative inheritance and preemption depends upon implausible assumptions about features of the actual world. Raz’s argument requires, in

---

particular, that moral agents are never epistemically well-placed enough to have reason to override directives that only imperfectly reflect the reasons they are based upon. Raz—and perhaps Conrad as well—would almost surely try to disassociate themselves from such a claim, and would likely grant that a particular agent is sometimes in a better position to judge what one ought to do than their normative “superiors”. But without this assumption about radical epistemic limitation on the part of the agent, it becomes clearly false that, in every case, one ought to act in accordance with another’s reasoning that is only generally better than mine. BP, then, remains unsolved—it does still make sense to bypass Razian normative dependents in favor of directly promoting their source.

In the same way that Gauthier’s and Rawls’ theory suffer rather than benefit from ‘preemption’ at the level of intension, so would Raz’s theory, if otherwise defensible, be morally simplistic because of ‘preemption’. If an authority issues a directive that I know does not accord with the deep reasons that apply to a particular case, then I am stubborn and immoral to ignore these deeper reasons. For if Raz is right that an authority’s directives are preemptive and themselves normative, then his theory cannot accommodate a proper understanding of the conflicting reasons that can exist in certain moral dilemmas. An authority might issue a directive that people ought not to steal. Suppose this directive is based upon the importance of preserving the well-being of persons, because the authority knows that if people steal the stability of society and so the well-being of persons will generally not be served. But if I am in a situation in which stealing something will clearly lead to a net improvement in well being (maybe my family is starving and a loaf of bread at a profitable market could be taken without much risk that I would be found out) then the moral reasons bearing on the act of stealing would be misdescribed by Raz. Even if it turns out that
I should not steal the bread, do I not still have a reason for stealing it that is just overridden in this case? Raz would say that the source of this supposed reason has been 'preempted', and so would have to deny what seems to be a genuine moral conflict in the case. Preemption, then, creates an intensional problem for Raz, even if he can avoid BP.

Raz, like Gauthier and Hare, seems to acknowledge an overarching commitment to satisfying the reasons that ground the normativity of an authority’s directives. Unfortunately for Raz, the directives of a legitimate authority can coherently be understood only as means to the end of complying with their grounding reasons. It is not obvious that Raz is open to the ‘character’ version of PRR. But it is clear that his directives do not have the independence required to help with the independence version of PRR. This is true even if Raz succeeds in showing what I have argued is implausible, namely that such directives are to be treated (always) as independently normative. Conrad, who hopes to enlist Raz’s account of legal reasons in the realm of morality, consequently fares no better than Raz himself.

iv. Plato’s “goodness in itself”

Plato’s defense of justice avoids appealing to the problematic properties of preemption and character change. These are virtues of his defense, since a plausible justification of these properties has not been forthcoming. Additionally, the structure of his defense is well-designed to solve BP, and in a very straightforward way. For Plato, performing an unjust act will never, as a matter of principle rather than as a matter of contingent fact, lead to Platonic happiness. The primary task, then, is to determine not whether justice is “good in itself” in the right way to solve BP—which appears to be the
challenge posed by Glaucon at the outset of *The Republic*—but rather to see how Platonic justice deals with the two versions of PRR.

Recall what it is required for justice to be good in itself for Plato: justice itself, and not the reputation for justice, must be a reliable and indispensable route to a sublime sort of happiness—the “fruit” of justice. One might put this another way: the intrinsic properties of justice, rather than its extrinsic properties, must make it a reliable and indispensable route to some further good thing.

Does Plato’s emphasis on the intrinsic properties of justice leading to a good result make it a better approach to PRR than any of the other ‘teleological, indispensability accounts’? Well, it gives it an advantage over, for instance, Hare’s two-level utilitarianism. Plato can say—if his theory is otherwise defensible—that being just is indispensable to obtaining happiness, since only the intrinsic properties of justice lead to happiness. There’s no possibility of a “go-around”. But this is just to say that Plato can solve BP by making justice an indispensable means to a certain end. We have already seen from our discussion of Gauthier and Raz that even this kind of indispensability is not enough for a sufficiently robust normative independence. Though functioning in some ways like an independently normative entity, it should be clear by now that a solution to the extensional problem is not a solution to the intensional problem.

C. Kagan-style teleological dispensability

This type of account does not, on further analysis, have much to recommend it. Certain intuitively plausible examples like the apparently derived but intrinsic value of Abraham Lincoln’s pen can be explained away. In the abstract, that something could acquire
value ‘for its own sake’ or (the term he prefers) ‘intrinsic value’ by being a means to
something else of value meets with strong intuitive resistance. In fact, the claim that
something could have “intrinsic value” because of its instrumental value turns out to be
incoherent, given what must be meant by the term.

According to Kagan, dependents are intrinsically normative because they are the final
items in a justificatory chain; they are not valuable because of anything further, but instead
are valuable “for their own sake”. This is the sort of value that one is inclined to think of as
attaching only to things that are intrinsically normative in Moore’s classical sense—namely,
the sort of value a thing has strictly in virtue of its intrinsic properties. If ‘intrinsic value’ is
given this Moorean interpretation, then it is contradictory to claim that something could be
valuable for its own sake in virtue of its instrumental value. For it should be clear that if the
source of some thing’s normativity is external to itself, then this thing is in fact valuable for
the sake of some other thing or set of things (than itself). Namely, it is valuable for the sake
of its normative source(s). Abe Lincoln’s pen, considered as a bare object stripped of, among
other thing, its history, is valuable only for the sake of its causal connection to various
valuable external events, persons, and states of affairs.

Why is Kagan tempted, then, to make the paradoxical-sounding claim that an
instrumentally valuable thing like Lincoln’s pen is valuable for its own sake in virtue of its
instrumental value? I think the answer is that ‘the value of Abraham Lincoln’s pen’ is used
to refer not to the bare object of a particular pen, but to the pen and the various important
events and persons the pen is causally connected to. In other words “Abraham Lincoln’s
pen” is being used to refer to a state of affairs in which the pen played a causal role. What is
valuable, then, is this state of affairs, and its value perhaps supervenes on its intrinsic
More generally, when Kagan claims that a non-intrinsically valuable thing is intrinsically valuable in virtue of its non-intrinsic value, the valuable thing in question is some valuable state of affairs, which is not always recognized because of the equivocal meaning of the relevant referring terms. If Kagan were only to be referring to the valuable state of affairs of which some object x is a part when he calls x ‘valuable for its own sake’, then he would not be contradicting himself, but his thesis would then be rather uninteresting. He would be entitled only to the claim that some states of affairs have intrinsic value in virtue of their internal properties, hardly a remarkable thesis. Instead, he makes the fallacious move from the innocuous claim that some state of affairs can be valuable in virtue of properties internal to it, to the surprising (and false) claim that some thing can be intrinsically valuable in virtue of properties external to it.

It is also not clear how Kagan has a principled way of stopping a problematic proliferation of progeny. He distances himself from the thesis that everything of instrumental value also has intrinsic value. But he does not indicate what feature a thing must possess in order for it to make the jump from being instrumentally valuable to intrinsically valuable. Not only might there be too many things that are intrinsically valuable in virtue of being instrumentally valuable to promoting ends of familiar intrinsic value, but since a whole new “generation” of things with intrinsic value would exist if Kagan is right, this new generation could bestow intrinsic value on those things that are useful in promoting them. And so on. A glut of things with intrinsic value looms, then.

---

7 Kagan (1998) and Rabinowicz and Rønnow-Rasmussen (1999) anticipate a similar objection. (However, they never suggest that the confusion comes from a word usually denoting a “bare object” instead being used to denote a state of affairs. This feature of my argument is important, since it helps to explain away the appearance of non-intrinsic, final value in a way that their account does not).
D. Non-hypothetical agent-election

Some philosophers claim that an entailment relationship exists between something being normative for its own sake and it being elected in some way (chosen, desired, loved, willed, etc.). On these accounts, the elector herself is a source of normativity, and her election of various objects entails that they are normative for their own sake.

One is reasonably suspicious of such claims at first blush, since it is *prima facie* consistent for any actual being to value something and for that thing nevertheless to have no value. The *equation* of something’s being elected for its own sake and something’s being valuable for its own sake might appear to be a species of the fact/value fallacy. So it is at least initially puzzling how the connection between actual agent-election and value might be forged.

One way we might interpret at least certain agent-election accounts is as only trying to establish the modest claim that some things that have no independent value are nevertheless *valued* for their own sake. For the purposes of fine-grained moral taxonomy, such a distinction might be useful. But for the purposes of this discussion the distinction is only interesting if something’s being valued for its own sake provided us with reasons for treating it in ways that we otherwise would not. But what arguments could be offered for thinking this? And could they solve the problems for moral theory I have presented?

Korsgaard, Allen Wood, and others claim to have an argument that connects rational valuing with “final” or “objective” value, where the former explains the latter. Even without such an argument, there is some plausibility to claims that attempt to connect value and valuing (or choosing, loving, etc.) in this way. Sometimes people in a close personal or romantic relationship will claim “you have a reason to value such-and-such because I value
such-and-such”. My best friend might desire to learn how to play tennis, but I find playing tennis with him terribly boring and frustrating and I would prefer not to play with him. It seems that his desire to learn tennis nevertheless gives me a reason to help him achieve this desire, and, derivatively, a reason to practice with him.

Still, desiring or valuing or choosing some thing surely is not in every case sufficient to give me a reason to likewise value, desire, or choose it. It matters what is being valued, desired, or chosen. What if the object of value is something horribly immoral? In this case, I don’t seem to have any reason to value, desire, or choose the object of value. Mere valuing or desiring or choosing something don’t seem to be the right sorts of things to infuse their objects with independent value, since each of these psychological states can be directed at objects without value, or even towards objects with disvalue. To solve BP and PRR, an agent-elected account would need to be more selective. It would need to show that only certain things are elected—namely, things that are moral. But people desire, choose, and value all sorts of things that are immoral.

One possibility is that a kind of constraint or filter needs to be put on agent-election so that the agent elects only those things that are genuinely valuable. One might suggest that only those things that are rationally valued, chosen, desired etc. can become independently normative. This is what Korsgaard and Wood, for instance, claim. An account of rationality would need to be developed and defended, of course. Maybe the thought would be that something is rationally elected only if it is actually elected, and it is something that would be elected in an idealized contractual setting, or (a non-Kantian possibility) by some hypothetical entity with a privileged moral vantage point. The possibilities are numerous--too numerous to discuss individually. Fortunately, I do not think a discussion of these
numerous possibilities are necessary for my purposes, since any such account will inevitably stumble over BP and PRR.

Because such “filtered” accounts require that an agent actually elect something in order for it to be independently normative, there is the possibility that some things that intuitively have independent normativity, but that need “external funding” to realize this independence, would fail to be elected. Even though rational agents would perhaps not be capable of electing immoral objects, they might nevertheless fail to elect something that was of moral value. In the context of actual election, rationality (or any other such constraint) can only set a limit on what is elected, but it does not ensure the proper breadth. Korsgaard writes that if a beautiful painting is locked up with no one to view it, it is not valuable. Maybe this is right in the case of paintings. But what if what is similarly beyond the purview of rational agents, and so not elected, is some unknown species of plant or animal, or a part of the planet or universe not yet explored? Whatever particular things not graced by the valuation of rational agents would not be valuable. The edges of some rainforest, where rational agents have ventured and valued, would be valuable, but not the center, which is too thick to be penetrated by humans (or other rational valuers). Parts of the ocean might be valuable, but spottily, depending upon which parts rational agents were aware of, and whether these parts were rationally valued.

If they are otherwise defensible, the problem I have identified for non-hypothetical agent-election accounts is perhaps not a version of BP, but only because of a technicality. Normative dependents on this schema are not in danger of being bypassed in favor of promoting their source. Instead, they can be problematically ignored, which, like BP, is an extensional problem for the view.
I also cannot see how a non-hypothetical agent-election account could solve PRR. A solution to PRR requires that normative dependents be in some sense independently normative. The value of the natural environment, for instance, is difficult to accommodate on a non-hypothetical agent election account because such an account does not seem to lend the right sort of “independence” to its elected dependents. For instance, Korsgaard’s final ends depend for their value on being actually valued; but, she stresses, they are valuable (in the sense of providing reasons for promoting or fostering them) independent of any other rational being’s positive appraisal of them. The value of such dependents is independent of something, but not the right thing. If I am convinced that the natural environment is valuable independently of its relationship to valuers, then even if I concede that its value is not “merely derivative” on a Korsgaardian sort of account I will not be satisfied. Even if I could be shown that rational agents necessarily value the environment—thus solving extensional worries about its value—one might hope that the value of the environment doesn’t come strictly from “outside” of the environment itself, from the agents who value them.

E. Hypothetical agent-election

Scanlon’s contractarianism

Recall the objection to Scanlon and other contract theorists I dealt with in the last Chapter: Scanlon aims to give an account of moral wrongness. So he needs to be able to explain why an action with feature F is morally wrong. He claims that some action with feature F is wrong if and only if it is reasonably rejectable; but if an action with feature F is reasonably rejectable that act had better not be reasonably rejectable for non-moral reasons. If it were rejectable for non-moral reasons, then that act would not be morally wrong. And
so the act with feature F must be reasonably rejectable for moral reasons. But if Scanlon is committed to this claim, then he has helped himself to a notion of moral wrongness in a question-begging manner. Moral wrongness, after all, is what reasonable rejectability is supposed to explain.

I suggested that Scanlon reject the claim an action with feature F must be reasonably rejectable for moral reasons if reasonable rejectability is to explain the moral wrongness of F. In other words, I suggested that Ridge’s slogan “non-moral reasons in/non-moral reasons out” be abandoned, in favor of working out a way to make sense of ‘character change’.

Until now, I have defended what I have called the default understanding of normative character acquisition. On this understanding, a normative dependent inherits the normative character of its source. It may seem strange, then, that I think that Ridge’s slogan—which apparently aligns with this default understanding—does not apply here. It does not apply, I think, because the reasons that make something reasonably rejectable are not normative reasons, but are rather explanatory reasons. They explain the intrinsic properties of the action with feature F, but reasonable rejectability does all of the justificatory work. A distinction must be made between the criteria that must be satisfied to make an action wrong, and the things that satisfy this criteria. That an action has feature F might explain why it satisfies the criteria for moral wrongness; that the action satisfies this criteria is what makes it wrong. For instance, if I ask why some action is wrong, I might be told “because it is an act of torture!” But a Scanlon would rightly say that this response is only useful as an explanation of the act under consideration; sure, the fact that it is an act of torture means that it has intrinsic properties that will (likely) make it reasonably rejectable, and so wrong. So it wouldn’t be wrong to say that the act is wrong because it is an act of torture, but the
“because” here would be explanatory rather than justificatory. Or so Scanlon could (and maybe should) claim.

If this is right, it needs to be explained why it seems grotesquely pedantic and beside the point to mention reasonable rejectability when explaining why something is wrong. Scanlon can say is that often by a mere description of a thing we can recognize that it is obviously reasonably rejectable. To say as much would be like telling someone that on a subzero winter morning one ought to wear a jacket. It goes without saying that one ought to wear a jacket on a subzero morning. But what justifies, in a normative sense, jacket-wearing on a subzero morning are considerations of prudence.

If I am right, one can consistently hold the default understanding of normative character acquisition and reject the “non-moral reasons in/non-moral reasons out” principle as irrelevant here. Because the reasons that make F reasonably rejectable are merely explanatory, they couldn’t infect actions with feature F with their non-moral character, or any other character for that matter.

This is good news for Scanlon, but it is bad news for the NP model, and specifically ‘character change’. It turned out that the normative character of the reasons that make something reasonably rejectable did not “change” once they reached the level of contractualism since they are not normative in the first place. So we are left with a disappointing result for our purposes: what looked like a promising example of normative character change turns out to be fully consistent with the default understanding of normative character acquisition, the latter of which rules out the possibility of character change.
III. Interpreting the results

Analyzing these different attempts at embodying the NP model has yielded a pattern of problems that any successful attempt must avoid. More positively, various conditions of adequacy have emerged, imposing constraints that any successful version of the model must satisfy. As options are ruled out, the remaining logical space begins to suggest what sort of occupants it might still accommodate.

A. Plato’s Principle

“Normative dependence” and “normative independence” are evidently opposites of one another, and their coinstantiation as normative inheritance would require it threatens to be contradictory. This much is obvious. But what, exactly, would constitute a contradiction in the realm of normative inheritance, and what are the logical possibilities for its avoidance? Plato’s discussion of opposites in the Republic is useful here. He notes that “opposite” desires can issue from the same individual—a desire both to drink, say, and not to drink—only if these desires arise at different times, from different parts of the individual, or with respect to different things. Plato is making use of the following principle: the same thing cannot be in opposite states at the same time, with respect to the same thing, with the same part of itself.

The principle may seem self-evident to the point of banality, but it is surprisingly useful: it exhausts the ways in which apparently contradictory claims could in fact be consistent. Suppose I am asked how I greeted my friend Bill when I got off the airplane, and I claim that I both embraced him and pushed him away. If I am telling the truth, I must have

---

8 Plato (1992) 12 and fn. 21; also 391.
either performed these different actions with different parts of my body\textsuperscript{9}, “pushed” and “embraced” him at the same time in a way that wasn’t really incompatible with its apparent opposite action, pushed him away and embraced him at different times, or greeted (and responded rather differently!) to more than one friend named Bill.

Plato’s principle makes clear the most abstract logical options for a successful version of normative inheritance: it must not make some normative entity \( x \) normatively dependent and independent of the same thing \( y \), at the same time, and with the same part of itself, on pain of contradiction.

Put positively, a normative dependent \( x \) and its source \( y \) must relate in one of the following five ways:

1. normative dependent \( x \) is not one thing, but in fact two different things—one is dependent on \( y \) and the other is independent.
2. \( x \)’s “opposite” states of dependence and independence with respect to \( y \) are not actually opposite.
3. \( x \) is in opposite states of dependence and independence with respect to \( y \) at different times.
4. \( x \) is in opposite states of dependence and independence with respect to \( y \) with different parts of itself.
5. \( x \) is in opposite states of dependence and independence with respect to different \( y \)’s.

\textsuperscript{9} Supposedly, a man who had undergone a commissurotomy—a procedure that severs the main connection between the hemispheres of the brain—once embraced his wife with one hand and pushed her away with another. (Nagel [1979]154 and Parfit [1984] 246).
This exhausts the logical space for possible versions of normative inheritance. We haven’t seen any examples of (1), but several instances of (2). Kagan, Korsgaard, and Plato use the language of intrinsic or independent value, inadvertently or not, in a way that confuses the status of the relevant dependents. Each of these philosophers does more than merely obfuscate the issues with suggestive rhetoric, but anointing dependents with properties that receive names like ‘final value’, ‘intrinsic value’ and ‘goodness in itself’ is misleading. Such terms suggests that these dependents are independent of their source in a certain robust way, when the description of these same dependents reveals that they are not.\textsuperscript{10}

We have seen no instances of (3), and for good reason, as it is pretty clearly a non-starter. I mentioned earlier that normative inheritance need not be instantiated along a temporal axis, even if the metaphor of inheritance suggests a historical process. But we may go further, and claim that normative inheritance cannot avoid the clash of normative dependence and independence by situating them at different temporal locations. The status of legitimate moral rules illustrates this nicely. As many consequentialists and non-consequentialists alike acknowledge, if it were not for the desirable consequences of at least certain rules, there would be no point in following them. Consider a moral rule that prohibits contact with birds because of the real risk of becoming infected with and spreading a dangerous flu virus. This rule is normative only during that time where infection is a genuine risk, or perhaps only during that time we have reason to believe it a genuine risk. After this

\textsuperscript{10} C.I. Lewis and William K. Frankena have employed the term ‘inherent value’ in a way that might similarly, and misleadingly, trade on the apparent overlap—at least in common language—of their favored term and ‘intrinsic value. The worry is that they might—intentionally or not—exploit the similarity of the familiar meaning of ‘inherent’ and ‘intrinsic’ to suggest that something can be both ‘intrinsically valuable’ and, in my sense, dependently valuable. As it turns out, for Lewis and Frankena something is ‘inherently normative’ if its intrinsic properties produce in an observer the intrinsically valuable appreciation of beauty (Lewis [1946] 391 and Frankena [1973] 82). Once spelled out, inherent value is clearly not equivalent to Moorean or even quasi-Moorean intrinsic value.

\textsuperscript{11} Cf. Chapter 3.
period of time the rule would no longer be normative. One can imagine irrationally believing
that the rule was normative for long after it had ceased having normative force, for it is easy
enough to develop beliefs (and desires) that long outlast their rational impetus. But this
would be a description of a psychological rather than a normative phenomenon, and in the
absence of a genuine threat (or reason to believe in such a threat) to continue to treat our
“bird-avoidance” rule as normative would be merely to make a fetish of it.

There have been no examples of (4), but it seems likely that any such examples would
be unsuccessful. To see why, consider an example. Aristotle’s virtues as they are discussed
in the Nichomachean Ethics are alleged to be both good in themselves and good for the sake
of something else (their contribution to eudaimonia). But his virtues would be good in
themselves even if they did not contribute to eudaimonia; they are not good in themselves,
then, in virtue of their contribution. Of course, they are instrumentally good in virtue of this
contribution. The goodness in itself of the virtues, then, is one “part” of the virtues, and it
this goodness is independent of the goodness of eudaimonia. The instrumental goodness of
the virtues—another “part”—is dependent on the goodness of eudaimonia.

This kind of account, however, is clearly not a version of NP, as the independent
normativity of the virtues is left unexplained by the normativity of eudaimonia. As such, it
does not solve the problems that motivate NP in the first place: it does not provide unity,
avoid positing mysterious intrinsically normative entities, or satisfy the intuition that many
things—including very prominently the moral virtues—seem to be good in themselves only
because they are good for the sake of something else.
(5) Rawls, Korsgaard, Raz, Gauthier, and Plato all have it that their respective dependents are independent of and dependent on different things.

For Rawls, legitimate rules are normative independent of whether a given act that fell under the rule maximized utility, but dependent on being more utility conducive than another other practice-constituting rules that apply to the same actions.

Korsgaard’s ‘final ends’ are normative independent of whether any rational being, besides the rational elector herself, values them for their own sake. But they are dependent for their normativity on being valued for their own sake by at least one rational being.

According to Raz, the directives of a legitimate authority are normative independent of whether in a given case following the directives will allow citizens to respond appropriately to the reasons that would, in the absence of his directives, apply directly to them. But these directives depend for their normativity on being issued by someone who is in a better position than his citizens to ascertain, generally speaking, which reasons apply to them.

For Gauthier, the moral disposition is normative independent of whether acting in accordance with the disposition in a given case will maximize expected utility. But its normativity is dependent on binding oneself to it being the best way, given certain facts about human nature, to maximize expected utility.

Finally, Platonic justice is normative independent of all possible changes in one’s circumstances. If one becomes enormously powerful and would have the power to be unjust with impunity, for instance, one still ought to be just. The independence of justice is possible
because of its normative dependence. For justice’s normativity is dependent on it being a reliable and indispensable means to happiness.

There is not one way of satisfying Plato’s principle that is clearly preferable to all others (though two of them appear to be ruled out). Satisfying one of the five ways, though, constitutes a necessary condition for the success of normative inheritance.

**B. Preemption and rigidity**

From certain teleological approaches to NP we have seen that preemption amounts to a kind of over-correction for a problem that direct-consequentialist approaches face. The admirable goal of preemption is to help solve BP and PRR by making impossible a direct justification of some particular thing (an action, say) by its normative source. But as we have seen, this feature results in extensional and intensional problems of its own. Though it is evidently desirable for dependents to possess a certain independence of normativity, to circumscribe the reason-giving power of sources so that they don’t apply in particular cases yields extensional results that are often contrary to commonsense morality. Further, and just as importantly, preemption takes away reasons that, intuitively, ought to exert normative force (even if not decisive normative force). A more nuanced understanding of the various reasons that bear on particular cases allows for a tension that preemption takes away. Preemption gives “backbone” to a theory, but at the price of over-rigidity and justificatory simplification.
C. Character change

I argued earlier that the normative character of a thing appears to be determined by the normative character of that which justifies it, or that which gives it its point. Nothing we have considered suggests that this presumptively true view should be overturned; quite the opposite, in fact. And if this view is true, character change appears impossible. For character change requires that a normative dependent have a normative character that is different from that of its source—that is, different from that which justifies it. Rawls’ version of rule utilitarianism perhaps came closest to making sense of character change without disturbing this default view. In the context of justifying particular actions, at least, legitimate moral rules are supposed to apply to actions as independently normative principles. Even here, though, there is a worry that, since in the context of justifying practices these rules are justified by their conformity with the principle of utility, the non-moral character of the principle of utility ultimately infects the character of legitimate rules even when the context changes. But even if this worry is unfounded, and the normative character of the principle of utility is truly blocked from affecting the character of moral rules, then it seems that there is nothing to give these rules their normative character. Their normative character, given the presumptively true view of normative character acquisition, would be indeterminate.

Scanlon’s contractarianism appeared at first like it might describe a plausible version of character change. But it turned out that Scanlon’s account actually conformed to the default understanding of normative character acquisition. The appearance to the contrary resulted from a confusion between explanatory and justificatory reasons; once it was sorted out which reasons were explanatory and which justificatory, there was no reason to posit character change to explain the character of moral wrongness.
An argument would need to be made to overturn the presumptively true account of normative character acquisition. But none of the approaches discussed in this chapter provide an argument for doing so. Since no such argument is obviously available, the notion of character change remains extremely problematic. A successful version of NP would apparently do well to avoid it.

D. Means and ends

Not every teleological account necessarily relegates its dependents to mere means. Rawls has a complicated—even if ultimately indefensible—teleological justificatory picture according to which moral rules are not merely instruments to promoting the end of utility maximization. But all of the “source-committed” teleological accounts we considered turned out to be more (or less) sophisticated ways of dressing up mere instruments to look like ends. Even where these accounts were successful in capturing morality’s extension, they failed to infuse the objects of their justification with a suitable independence.

That instrumental value could confer intrinsic or some other form of independent value meets with strong intuitive resistance, and it turns out for good reason. It goes against what remains one of the few truisms of moral theorizing. Harry Frankfurt voices the intuition nicely (his expression “terminal value” is equivalent in meaning to “value for its own sake” or “intrinsic value”):

“A means acquires no terminal value from being useful. The relationship in which it stands to its end can endow a means only with instrumental value. Of course, what
has instrumental value may have terminal value as well. But it cannot have the latter by virtue of the fact that it has the former...”12

Frankfurt puts the authority of historical consensus behind his opinion, as he claims to be voicing a view that dates back at least to Aristotle. He does not talk about sources and dependents in the way I do, though his point could be recast in my terminology: mere source-conduciveness does not seem to be the sort of thing that could confer independent value upon the source’s dependents.

The inability of means to transubstantiate, as it were, into ends is not a function of (what is typically) their causal role. If a component of some whole was not valuable in itself, but only valuable as a component of the whole, it would still lack normative independence—even if it was a necessary component of the whole. For instance, suppose (as Aristotle claims) the property of having a long life was a necessary component of the good life, where the good life is itself valuable. Suppose also that on its own a long life would have no positive or negative value. In this case, a long life does not in virtue of its constitutive role with respect to the good life have value in itself. Its value is still merely instrumental, even though it is not a “cause” of its source in the temporal sense.

E. Selectivity/promiscuity

Korsgaard’s account highlights a condition of adequacy for any successful version of NP. Korsgaardian rational election handed out normative inheritance too liberally. She argued that anything that was rationally valued had “final” and so objective value. But rationality apparently does not impose strict enough constraints to prevent, say, one person’s

---

coffee maker collection, or another’s stamp collection, from acquiring objective value and so imposing objective duties—duties, that is, that are normative for all rational beings. A successful account must be sufficiently selective in what it counts as independently normative.

Kagan’s view of intrinsic normativity made vivid how a parallel problem of “promiscuity” could arise in a teleological context. If things were thought to acquire independent value by being a reliable means to some independently valuable end, one can imagine how independent value could objectionably proliferate. For instance, if pleasure was taken to have independent value, then money, various pleasure-inducing (and sustaining) drugs, and objectionably dishonest and unfaithful dispositions could all acquire, through their utility-conduciveness, independent value. A constraint that imposes sufficiently strict limits on what can inherit normativity is required, then, in a teleological context as well.

IV. A final Clarification

‘Preemption’ and ‘character change’ have been judged to be aspects of the NP model that are best trimmed off. That leaves us with the core of the model that we began with, normative inheritance.

Philosophers seem to desire that the recipients of normative inheritance resemble as closely as possible intrinsically normative or intrinsically valuable entities in the strong Moorean sense. However, normative inheritors can attempt to approximate intrinsic normativity only within the strict framework provided by Plato’s principle. Within this framework, certain basic strategies have been revealed as non-starters.
Perhaps it will be thought that Plato’s framework provides logical space for Moorean intrinsic normativity earned through normative inheritance, but this optimism would seem to rest on a confusion. The thought might be that some x has intrinsic normativity in virtue of the intrinsic properties of x. This is a classic way of describing intrinsic normativity. Why can’t we speak of x’s intrinsic properties as the normative source of x’s normativity, and x’s normativity as a dependent of these intrinsic properties?

Unfortunately, the relationship between a thing’s intrinsic properties and the normativity of these properties is not that of a source to its dependent; rather, we say of such things that their normativity *supervenes* on their (non-normative) properties.¹³ This is not just an alternative way of describing a source/dependent relationship. Rather, we have in either element that composes a supervenience relationship—the naturalistic base and the normativity that supervenes upon this base—no normative source. We have nothing that is, itself, normative independently of its normative or naturalistic counterpart. We simply have one, individually insufficient, component of an independently normative entity. Without the source/dependent relationship, intrinsically normative entities cannot satisfy the guiding demand for dependent yet independent normativity, a demand that gives rise to the problems of BP and PRR I have grappled with throughout this essay.

As I recently mentioned, various philosophical descriptions of normative justification mimic or even borrow in interesting ways the *language* of intrinsic value and intrinsic normativity—as a result, they also mimic the language one would use to describe supervenience. But none of these attempts have filled in that logical space between simple supervenience and self-defeating justification.

¹³ See Rabinowicz and Rønnow-Rasmussen (1999) for a nice discussion of (among other things) the difference between simple supervenience and normative justification.
V. Conclusion

In this dissertation I hope to have identified a structural problem for moral justification, a problem that, though remarkably widespread and hugely troublesome, has resisted a properly general characterization. A common approach to this problem—attempting to make use of the model of justification I have been describing—has been similarly reluctant to reveal both its popularity and its structural features. I am hopeful that an understanding of the structure of this problem and its most popular (and plausible) solution makes the problem less intractable. More broadly and probably more ambitiously, I hope that in making explicit certain deep structural features of morality and its defense I have shed some fresh light on the point, ambitions, prospects, and limits of moral justification.
Bibliography


Lewis, Clarence I. 1946: *An Analysis of Knowledge and Valuation*, La Salle: Open Court.


