Kant's Search for the Supreme Principle of Morality

At the core of Kant's ethics lies the claim that if there is a supreme principle of morality, then it is not a utilitarian or Aristotelian perfectionist principle, or even a principle resembling the Ten Commandments. The only viable candidate for the supreme principle of morality is the Categorical Imperative.

This book is the most detailed investigation of this claim. It constructs a new, *criterial* reading of Kant's derivation of one version of the Categorical Imperative: the Formula of Universal Law. This reading shows this derivation to be far more compelling than contemporary philosophers tend to believe. It also reveals a novel approach to deriving another version of the Categorical Imperative, the Formula of Humanity, a principle widely considered to be the most attractive Kantian candidate for the supreme principle of morality.

Lucidly written and dealing with a foundational topic in the history of ethics, this book will be important not just for Kant scholars but for a broad swath of students of philosophy.

Samuel J. Kerstein is Associate Professor of Philosophy at the University of Maryland, College Park.

Kant's Search for the Supreme Principle of Morality

SAMUEL J. KERSTEIN
University of Maryland, College Park

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2002

Printed in the United Kingdom at the University Press, Cambridge

Typeface ITC New Baskerville 10/12 pt. *System* LaTeX 2ε [TB]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication data

Kerstein, Samuel J., 1965–
Kant's search for the supreme principle of morality / Samuel J. Kerstein.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-81089-2
1. Kant, Immanuel, 1724–1804 – Ethics. I. Title.
B2799.E8 K45 2002
170–dc21 2001043918

ISBN 0 521 81089 2 hardback

For Lisa

Contents

Acknowledgments

This book would not have been completed without help and support from a variety of sources.

I would like to thank Terence Moore and Brian R. MacDonald of Cambridge University Press for their patience and expertise in guiding me through the publication process.

Material from four of my papers has been reworked into the book. Chapter 1 incorporates "Kant's (Not So Radical) Hedonism," in *Kant und die Berliner Aufklärung. Akten des IX. Internationalen Kant-Kongresses*, vol. 3, ed. V. Gerhardt, R.-P. Horstmann, and R. Schumacher (Berlin: Walter de Gruyter, 2001), pp. 245–253. Part of Chapter 3 stems from "Korsgaard's Kantian Arguments for the Value of Humanity," *Canadian Journal of Philosophy* 31 (March 2001): 23–52. Sections of Chapters 4 and 7 have been adapted from a paper I coauthored with Berys Gaut: "The Derivation without the Gap: Rethinking *Groundwork* I," *Kantian Review* 3 (1999): 18–40. Finally, parts of Chapters 5 and 6 were published in "The Kantian Moral Worth of Actions Contrary to Duty," *Zeitschrift für Philosophische Forschung* 53 (1999): 530–551. I acknowledge with appreciation the permission of the publishers to use material from these papers.

Most of the book was written during the academic year 1999–2000, which I spent as a Fellow at the National Humanities Center in Triangle Park, North Carolina. I would like to thank the National Endowment for the Humanities for supporting my stay there. The administrators and staff at the National Humanities Center could not have been more encouraging and helpful. In particular I would like to thank Karen Carroll, who edited an early version of my manuscript. (I would also like to thank Jane Strong for editing a later version.) Preliminary work on the manuscript was made possible by support from the University of Maryland, College Park, in the form of a General Research Board grant that relieved me from my teaching duties during the fall of 1996. I would like to thank the University of Maryland for this support, as well as for granting me leave to work at the National Humanities Center.

For their comments and criticisms of portions of this book, I would like to thank audiences at the British Kant Society Annual Meeting, the Central Division Meeting of the American Philosophical Association, the Midwest Study Group of the North American Kant Society, Duke University, the University of St. Andrews, and the University of North Carolina, Chapel Hill.

From early on I have been fortunate to have had outstanding teachers. I would like to thank Noël Carroll and Victor Gourevitch for their guidance, both philosophical and personal. I am grateful to Bonnie Kent who took the time to teach me not only how to work in the history of philosophy but to appreciate the importance of doing so.

I have learned a great deal about Kantian ethics from discussion and/or correspondence with many philosophers, including Paul Cohen, Michèle Crampe-Casnabet, Garrett Cullity, David Cummiskey, Raymond Geuss, Stéphane Haber, Thomas Hill Jr., Dieter Schönecker, Ralf Stöcker, and Allen Wood. I owe a special debt of gratitude to Berys Gaut. Some central ideas in the book stem from our collaborative work, and Berys has been generous in encouraging me to develop them at greater length. Readers for Cambridge University Press, as well as two others, offered comments that have, I think, enabled me to strengthen several of my arguments. During my stay at the National Humanities Center, I profited from (often ambulatory) dialogue with many colleagues, including Ruth Grant, Michelle Massé, Louise McReynolds, Bernard Reginster, Daniel Sherman, Eleonore Stump, Timothy Taylor, and Marjorie Woods. I was especially fortunate to have been able to discuss philosophy with Thomas Christiano, who not only provided intellectual inspiration, but patiently helped me to work out some key points in the book. My friends and colleagues at the University of Maryland, especially Judith Lichtenberg and Corey Washington, have aided me at several points, both intellectually and personally, in carrying out this project.

I am deeply grateful for the help and support I have received from Rüdiger Bittner, Thomas Pogge, and Michael Slote. From the beginning, these philosophers have played essential roles in the book's development. Each gave me valuable advice on my project as it unfolded, and offered trenchant and productive comments on the manuscript as a whole. My approach to Kantian ethics owes a great deal to each of them.

Finally, I would like to thank my in-laws John and Jane Strong, my parents Howard and JoAnn Kerstein, and especially my wife Lisa Strong, for their constant encouragement during the writing of this book.

Key to Abbreviations and Translations

Except for references to the *Critique of Pure Reason*, all references to Kant are to the Preussische Akademie der Wissenschaften edition of his works (Berlin: Walter de Gruyter [and predecessors], 1902). References to the *Critique of Pure Reason* are to the standard A and B pagination of the first and second editions. I list here the German title, academy edition (Ak.) volume number, and abbreviation for each of the works I cite. Under each entry, I specify the English edition I have consulted. The translations I employ sometimes vary from those of these English editions.

Anth *Anthropologie in pragmatischer Hinsicht* (Ak. 7)
Anthropology from a Pragmatic Point of View, tr. Victor L. Dowdell. Carbondale: Southern Illinois University Press, 1978.

GMS *Grundlegung zur Metaphysik der Sitten* (Ak. 4)
Groundwork of the Metaphysics of Morals, tr. Mary J. Gregor. In *Immanuel Kant: Practical Philosophy*, 42–108. Cambridge: Cambridge University Press, 1996.

KpV *Kritik der praktischen Vernunft* (Ak. 5)
Critique of Practical Reason, tr. Mary J. Gregor. In *Immanuel Kant: Practical Philosophy*, 138–271. Cambridge: Cambridge University Press, 1996.

KrV *Kritik der reinen Vernunft* (1st ed. (A) 1781; 2nd ed. (B) 1787; Ak. 3–4)
Critique of Pure Reason, tr. N. Kemp Smith. New York: St. Martin's Press, 1965.

KU *Kritik der Urteilskraft* (Ak. 5)
Critique of Judgment, tr. Werner S. Pluhar. Hackett: Indianapolis, 1987.

KUE *Erste Einleitung in der Kritik der Urteilskraft* (Ak. 20)
In *Critique of Judgment*, tr. Werner S. Pluhar. Hackett: Indianapolis, 1987.

LE	*Vorlesungen über Moralphilosophie*, "Moralphilosophie Collins" (Ak. 27) *Lectures on Ethics*, "Moral Philosophy: Collins's Lecture Notes," tr. Peter Heath, 37–222. Cambridge: Cambridge University Press, 1997.
MS	*Die Metaphysik der Sitten* (Ak. 6) *The Metaphysics of Morals*, tr. Mary J. Gregor. In *Immanuel Kant: Practical Philosophy*, 363–603. Cambridge: Cambridge University Press, 1996.
Rel	*Die Religion innerhalb der Grenzen der blossen Vernunft* (Ak. 6) *Religion within the Limits of Reason Alone*, tr. T. M. Greene and H. H. Hudson. New York: Harper & Row, 1960.

All of the English editions incorporate academy edition page numbering in their margins, except for the KrV and Rel. When I cite the Rel, I give the academy edition page number followed by that of the English edition.

Introduction: Derivation, Deduction, and the Supreme Principle of Morality

1.1 No Modest Claim

If there is a supreme principle of morality, then it is the Categorical Imperative. This claim, which lies at the core of Kant's ethics, is nothing if not ambitious. Establishing it would amount to proving that absolutely no principle other than the Categorical Imperative – no utilitarian principle, no perfectionist principle, no principle along the lines of the Ten Commandments – is a viable candidate for the supreme principle of morality. How does Kant (or might he) try to prove this? Does he (or might he) succeed? Questions of this sort are what this book is about. To answer them, we must understand what Kant means by claiming that if there is a supreme principle of morality, then it is the Categorical Imperative.

1.2 The Basic Concept of the Supreme Principle of Morality

To begin we need to know how Kant conceives of the supreme principle of morality. According to (what I call) his basic concept, this principle would possess four characteristics. It would be practical, absolutely necessary, binding on all rational agents, and would serve as the supreme norm for the moral evaluation of action. I call this concept of the supreme principle of morality basic because it emerges immediately in Kant's critical writings in ethics.[1] Already in the Preface to the *Groundwork of the Metaphysics of Morals* it is manifest that, in Kant's view, the supreme principle must have these features.

It belongs to Kant's basic concept of the supreme principle of morality that it constitute the supreme norm for the moral assessment of action. This means several things. The principle would distinguish between morally permissible actions, that is, ones that conform with the principle, and morally impermissible actions, that is, ones that conflict with the principle (see GMS 390). It would also specify which actions are morally required. As

Kant suggests in the *Groundwork* Preface, the supreme principle of morality would not only be the basis for appraising an action's moral requiredness, permissibility, or impermissibility, but also its moral goodness (GMS 390). Whether an action is morally good depends on how it relates to this principle. In particular, to be morally good an action must both conform with and be done "for the sake of" the principle. Finally, as the supreme norm for the moral assessment of action, the supreme principle of morality would be such that all genuine duties would ultimately be derived from it (see GMS 421).[2] The supreme principle would justify these duties' status as such.

Kant says that the supreme principle of morality "must hold not only for human beings but for all *rational beings as such*" (GMS 408; see also GMS 389, 425, 442; KpV 32, 36).[3] The supreme principle of morality would have an extremely wide scope: one that extended not only to all rational human beings but to any other rational beings who might exist – for example, God, angels, and intelligent extraterrestrials. In Kant's view, the supreme principle of morality would have to possess what I call "wide universal validity." It would have to be binding on *all* rational agents, at all times and in all places. This is the second feature that, according to Kant's basic concept, the supreme principle of morality would have to possess.

To say that the supreme principle of morality is binding on us (human agents) is to imply that we have an obligation to act in accordance with it. We ought to but, as a result of privileging inclinations over duty, might not follow its dictates. The same could also be said for any nonhuman rational agents who had characteristics, for example, natural cravings, on the basis of which they might act contrary to the supreme principle. The supreme principle's being binding on these agents would imply that they had an obligation to act in accordance with it. For all agents "affected by needs and sensible motives," the supreme principle of morality would count as an "*imperative*" (KpV 32). It would set out a command that we genuinely ought to obey, although we might not obey it (GMS 414). We can conceive of beings, however, on whom the supreme principle would be binding but regarding whom it would be incorrect to say that they had an obligation to obey it. According to Kant, one can be obligated to do something only if there is a possibility that he will fail to do it.[4] Yet some beings, for example, God, might be such that they cannot fail to obey the supreme principle of morality. It would thus make no sense to say that they had an obligation to obey it. For them, the supreme principle of morality would be a law but not an imperative (GMS 414, 439; KpV 32).

A third feature the supreme principle of morality would have to possess is that of being absolutely necessary (GMS 389). Kant's description of this feature answers the question of what it would mean for the supreme principle of morality to be binding on an agent. On every agent within its scope, for Kant every rational agent, the principle would hold without exception (GMS 408). For example, a human agent would always be obligated to act

in accordance with the supreme principle, no matter what he wants to do. For us, the supreme principle of morality would be an unconditional command. That we were obligated to perform the action it specified would not be conditional on our having any particular set of desires.

Finally, it is worth making explicit that for Kant the supreme principle of morality must be practical – it must be a rule on account of which agents can act. Kant implies this in the *Groundwork* Preface by specifying that morally good actions involve an agent's acting for the sake of the moral law, that is, the supreme principle of morality (GMS 390). In the *Critique of Practical Reason*, he defines practical principles, of which the supreme principle of morality would be one, as propositions that "contain a general determination of the will," thereby suggesting that this principle would be something on the basis of which an agent can set himself to do something (KpV 19–20).[5] One might conceive of the supreme principle of morality as a purely theoretical tool. For example, one might take it to be a rule that could be employed (perhaps by a team of experts) to categorize something an agent has done in terms of its rightness or wrongness, but which (perhaps due to its enormous complexity) could not be used by the agent himself in deciding what to do. This would be a very un-Kantian conception of the supreme principle of morality. For Kant the supreme principle must be able to figure directly in an agent's practical deliberations.

From the very outset of his first great work in ethics, Kant operates with a certain basic concept of the supreme principle of morality. It is evident from the Preface of the *Groundwork* that he thinks of this principle as practical, absolutely necessary, binding on all rational agents, and the supreme norm for the moral evaluation of action.

Three remarks are in order regarding Kant's basic concept of the supreme principle of morality. First, as we will see, there is more to Kant's concept of the supreme principle of morality than is captured in this basic concept. There are more features that, in Kant's view, the supreme principle would have to possess. It would, for example, have to be such that a proponent of its being the supreme principle of morality could coherently claim that obeying it "from duty" would have moral worth. The second point concerns the provenance of the four features that belong to (what I call) Kant's basic concept. Kant, I think, would claim that if we – that is, beings who possess "common rational moral cognition" – reflect a bit on what the supreme principle of morality would be like, we find that it would have to possess these four features.[6] Kant makes it clear that, according to him, commonsense morality is committed to the view that absolute necessity and wide universal validity must be features of the supreme principle of morality. Implicit in "the common idea of duty and of moral laws," says Kant, is that "a law, if it is to hold morally, that is, as a ground of an obligation, must carry with it absolute necessity; that, for example, the command 'thou shalt not lie' does not hold only for human beings, as if other rational beings did not have to

heed it, and so with all other moral laws properly so called" (GMS 389).[7] The third remark regarding Kant's basic concept of the supreme principle of morality concerns its role in this book. We will be probing arguments for the claim that if there is a supreme principle of morality, *corresponding to Kant's basic concept of such a principle*, then it is the Categorical Imperative. For purposes of this book, Kant's basic concept of the supreme principle of morality is assumed. As readers will quickly see, assuming this concept does not at all render it trivial or easy to establish that the Categorical Imperative is the only viable candidate for the supreme principle of morality.

1.3 Derivation and Deduction of the Categorical Imperative

To refine further our understanding of what Kant means by claiming that if there is a supreme principle of morality, then it is the Categorical Imperative, we need to place the claim into the context of the work in which it initially arises, the *Groundwork of the Metaphysics of Morals*. Kant divides the *Groundwork* into a Preface and three sections. In the Preface, he says: "[T]he present *Groundwork* is . . . nothing more than the search for and establishment of *the supreme principle of morality*" (GMS 392). In *Groundwork* I and II, Kant searches for the supreme principle of morality in the sense that he tries to discover what this principle would be, assuming there is such a principle. Kant presents the Categorical Imperative by name for the first time in *Groundwork* II: "[A]ct only on that maxim through which you can at the same time will that it become a universal law" (GMS 421, Kant's emphasis omitted). Right after he presents this principle, he says: "Now, if all imperatives of duty can be derived from this single imperative as from their principle, then, *even though we leave it undecided whether what is called duty is not as such an empty concept*, we shall at least be able to show what we think by it and what the concept wants to say" (GMS 421, emphasis added). Throughout *Groundwork* II, Kant reminds us that he is there offering no proof that the Categorical Imperative is absolutely necessary and universally binding, and thus no proof that genuine moral duties derive from it (see GMS 425, 431). At the end of *Groundwork* II, Kant tells us what, in his view, he has demonstrated to that point: "[W]hoever holds morality to be something and not a chimerical idea without any truth must also admit the principle of morality brought forward" (GMS 445). The "principle of morality brought forward" is, of course, the Categorical Imperative. So by the end of *Groundwork* II, Kant takes himself to have completed his search for the supreme principle of morality by showing that if there is a supreme principle of morality, then it is the Categorical Imperative.

Let us call an argument aimed at proving that if there is a supreme principle of morality, then it is some particular principle, a "derivation" of this principle.[8] As we will see, Kant carries out a derivation of the Categorical Imperative not only in the *Groundwork* but in the *Critique of Practical Reason*

as well. He offers several arguments for the conclusion: if there is a supreme principle of morality, then it is the Categorical Imperative.

A successful derivation would prove this conditional conclusion. It would complete Kant's search for the supreme principle of morality (or, more precisely, his search for what would be this principle, if anything is). But, as we have seen, in the Preface Kant says that the *Groundwork* does more: it establishes the supreme principle of morality (GMS 392). In *Groundwork* III, Kant tries to close a possibility left open by *Groundwork* I–II: the possibility that duty is an empty concept, that is, that we actually have no (moral) duties. He aspires to prove that the Categorical Imperative is valid: absolutely necessary and binding on all rational agents (GMS 461).[9] Kant suggests in the *Groundwork* as well as later in the *Critique of Practical Reason* that proving this would amount to giving a "deduction" of the supreme principle of morality (see GMS 454, 463; KpV 47, 48). Kant's usage of the term "deduction" in the *Critique of Pure Reason* signals that to carry out a deduction of the Categorical Imperative would be to show that we have a right, that is, sufficient justification, for considering it to be valid (KrVA 84–85/B 116–117). By the end of *Groundwork* II, Kant takes himself to have shown that those of us who believe there to be a supreme principle must embrace the Categorical Imperative as this principle. Yet that we who believe that there is such a principle must embrace the Categorical Imperative does not entail that it is actually binding on us – that we actually have the duties this imperative specifies. Our belief in morality might be mistaken. A successful derivation of the Categorical Imperative would not eliminate the possibility that morality is a "chimerical idea."

The aim of producing an effective derivation of the Categorical Imperative seems less aspiring than that of giving a deduction of it. A derivation that worked would show us what the supreme principle of morality would be, if there was one, but, unlike a deduction, it would not show us that any given principle was actually binding on us. By giving a deduction of the Categorical Imperative, Kant would answer two different opponents. First, he would answer a moral skeptic, someone who holds that we are not obligated to do anything at all. For he would establish that we are obligated to act only on maxims that we can, at the same time, will to be universal laws. Second, if Kant provided a deduction of the Categorical Imperative, he would answer a "moral particularist," namely someone who believes in the reality of moral distinctions – for example, that there are right actions and wrong ones – but who denies that there are any moral *principles* binding on all rational agents or even all human agents.[10] For Kant would demonstrate that the Categorical Imperative is just such a principle. By giving a successful derivation of the Categorical Imperative, Kant would refute neither the moral skeptic nor the moral particularist. Both opponents would remain free to agree with Kant that if there were a supreme principle of morality, then it would have to be the Categorical Imperative, yet to deny that there is any such principle.[11]

It would be remiss not to mention that by the end of *Groundwork* II Kant takes himself to accomplish more than a derivation of the Categorical Imperative. In addition to demonstrating that if there is a supreme *principle* of morality, then it is the Categorical Imperative, he also thinks he proves a stronger claim: if morality *tout court* is not an illusion, then it has a supreme principle, namely the Categorical Imperative: "[W]hoever holds *morality* to be something and not a chimerical idea without any truth must also admit the principle of morality brought forward" (GMS 445, emphasis added). So, in effect, Kant implies that by the end of Section II, we have a response to moral particularism. Moral particularism entails moral skepticism, suggests Kant; morality not based on principle would be no morality at all.

I do not discuss this suggestion. Nor do I focus on Kant's deduction of the Categorical Imperative. Instead, I concentrate on Kant's derivation. The aim of generating a successful derivation of the supreme principle of morality is, I think, sufficiently ambitious to warrant our full attention. If Kant attains it, then he shows that as far as candidates for the supreme principle of morality are concerned, the Categorical Imperative is (and will be) the only game in town.

Even though our focus is on Kant's derivation, and not his deduction, of the Categorical Imperative, it is worth noting that Kant eventually seems to abandon the project of providing a deduction. In the *Critique of Practical Reason*, published three years after the *Groundwork*, he asserts:

> [T]he moral law is given, as it were, as a fact of pure reason of which we are a priori conscious and which is apodictically certain, though it be granted that no example of exact observance of it can be found in experience. Hence the objective reality of the moral law cannot be proved by any deduction, by any efforts of theoretical reason, speculative or empirically supported, so that, even if one were willing to renounce its apodictic certainty, it could not be confirmed by experience and thus proved a posteriori; and it is nevertheless firmly established of itself. (KpV 47; see also KpV 48 and 93)

This passage raises many complex issues, but for our purposes a brief treatment suffices. In *Groundwork* III, Kant implies that he is undertaking a deduction of the Categorical Imperative (GMS 461, 463). Yet in this second *Critique* passage, Kant suggests that the "objective reality" (i.e., validity) of the moral law is "firmly established of itself"; it does not need to be proved through philosophical argument. In stating that the moral law is given as a fact of pure reason of which we are a priori conscious and which is apodictically certain, Kant is apparently suggesting that the moral law necessarily presents itself to each rational agent as a valid practical requirement. To use Rüdiger Bittner's description, Kant seems to be implying that "one is cognizant of [the moral law] in such a way that in all practical considerations one knows of its validity and has to take this validity into account."[12] Since we are cognizant of the moral law in this way, Kant appears to hold,

there is no need for arguments to show us that we are genuinely bound by it. The project of deduction he undertakes in *Groundwork* III is, Kant now thinks, an unnecessary one. That it is unnecessary to prove the validity of the Categorical Imperative does not entail that it is impossible to do so. Yet Kant even goes so far as to make the further claim that this project *cannot* succeed: "[T]he objective reality of the moral law cannot be proved by any deduction."[13] Kant's grounds for this further claim need not concern us. However, that he makes it strengthens the impression that he eschews the *Groundwork* III attempt to prove the validity of the Categorical Imperative.

If, as it appears, Kant abandons this attempt, it does not, of course, follow that we ought to do so. Kant might have failed to appreciate the strength of his own arguments. But I do not try to make the case that he did.[14]

1.4 The (Alleged) Gap in the Derivation of the Formula of Universal Law

Readers familiar with Kant's derivation of the Categorical Imperative might wonder why it merits a book length treatment. After all, according to the received view, it falls conspicuously short. Kant sketches his derivation of this principle in both *Groundwork* I and II. Here are central (and famously difficult) passages in each:

> But what kind of law can that be, the representation of which must determine the will, even without regard for the effect expected from it, in order for the will to be called good absolutely and without limitation? Since I have deprived the will of every impulse that could arise for it from obeying some law, nothing is left but the conformity of actions to universal law as such, which alone is to serve the will as its principle, that is, *I ought never to act except in such a way that I could also will that my maxim should become a universal law.* Here mere conformity to law as such, without having as its basis some law determined for certain actions, is what serves the will as its principle, and must so serve it, if duty is not to be everywhere an empty delusion and a chimerical concept. (GMS 402)

> When I think of a *hypothetical* imperative in general I do not know beforehand what it will contain; I do not know this until I am given the condition. But when I think of a *categorical* imperative I know at once what it contains. For since the imperative contains, beyond the law, only the necessity that the maxim be in conformity with this law, while the law contains no condition to which it would be limited, nothing is left with which the maxim of action is to conform but the universality of a law as such; and this conformity alone is what the imperative properly represents as necessary.
>
> There is, therefore, only a single categorical imperative and it is this: *act only on that maxim through which you can at the same time will that it become a universal law.*
>
> Now, if all imperatives of duty can be derived from this single imperative as their principle, then, even though we leave it undecided whether what is called duty is not as such an empty concept, we shall at least be able to show what we think by it and what the concept wants to say. (GMS 420–421)

In both passages, Kant argues for a conditional claim. If duty is not an "empty" or "chimerical" concept, that is, if there are genuine moral obligations, then the Categorical Imperative is the principle of these obligations, the supreme principle of morality. In both passages, Kant is offering a derivation, or part of a derivation, of the Categorical Imperative.

If we are to believe the received view, both the *Groundwork* I and the *Groundwork* II derivation fail. They fail because they contain a crucial gap. In each, Kant embraces a principle that is, for practical purposes, virtually uninformative. Without argument, he then jumps to the Categorical Imperative as the only viable candidate for the supreme principle of morality.

Bruce Aune offers an influential expression of the received view. Aune argues that both versions of the derivation fail, but let us follow him in focusing on *Groundwork* I.[15] In the very sentence in which Kant sets out for the first time the principle we refer to as the Categorical Imperative, he says that "nothing is left but the conformity of actions to universal law as such, which alone is to serve the will as its principle" (GMS 402). According to Aune, Kant's saying this amounts to his embracing the principle L : "Conform your actions to universal law."[16] L, suggests Aune, "is a higher-order principle telling us to conform to certain lower-order laws."[17] L "formulates the basic moral requirement"; it commands that we conform our actions to these lower-order laws: principles that are necessarily binding on all of us.[18] But L does not tell us what these laws are. It fails to indicate, for example, that among them we would find "Do not commit suicide," rather than, say, "Minimize your suffering." Kant, Aune says, jumps directly from L to the Categorical Imperative, which Aune calls C1: "Act only on that maxim through which you can at the same time will that it should become a universal law."[19] In *Groundwork* I, Kant assumes that "we conform to universal law (and so satisfy L) just when we obey C1 and act only on maxims that we *can* will to be universal laws."[20]

Yet, notes Aune, this assumption is far from obvious, as it is easy to illustrate. Kant holds that in acting on a maxim of nonbeneficence – for example, "To maximize my happiness, I will refrain from helping others in need" – I would be disobeying C1 (GMS 423). Suppose Kant is right about this. According to the assumption in question, then, in acting on this maxim, I would not be conforming to universal law: to a principle that is necessarily binding on all of us. But it is unclear why I would not be. For all Kant has shown thus far, it could be that a principle necessarily binding on all of us is: "Always do what you believe will maximize your own happiness." In acting on my maxim of nonbeneficence, I could be conforming to this universal law. Kant, Aune suggests, embraces L as the basic requirement of moral action. Kant affirms that if there is such a thing as moral action, then it is action conforming to universal law. But then, without argument, Kant jumps to the conclusion that the only way for an action to conform to universal law is for it to conform to C1. The gap Aune finds in Kant's *Groundwork* I derivation is

between the (for practical purposes) uninformative principle L and C1, the Categorical Imperative.[21]

Aune is far from alone. Several other philosophers, even ones sympathetic to a Kantian approach in ethics, have claimed to find a gap of this sort.[22] In their view, in neither *Groundwork* I nor II does Kant succeed in defending a move he makes from a practically uninformative principle to the Categorical Imperative.

Allen Wood, for example, has recently interpreted the *Groundwork* I and II derivations in essentially the same way as Aune. According to Wood, in both derivations Kant tries to establish that "our maxims ought to conform to whatever universal laws there are."[23] But then Kant jumps without argument from this rather empty principle to the Formula of Universal Law. Kant illegitimately takes for granted that the only way to conform to whatever universal laws there are is to conform to the Formula of Universal Law.

Henry Allison discusses another characterization of the practically uninformative principle from which Kant (supposedly) moves directly to the Categorical Imperative. On this characterization, the principle is (what I call) the "principle of rightness universalism":

> RU: If a maxim or action is judged permissible for a rational agent in given circumstances, it must also be judged permissible for any other rational agent in relevantly similar circumstances.[24]

RU is rather vague: for one, it is not clear what are to count as "relevantly similar circumstances." However, this version of the traditional reading focuses on (what it sees as) Kant's move directly from RU to the Categorical Imperative. According to this version, Kant presents the Categorical Imperative in a parenthetical clause aimed at explicating the prescription that the will conform its actions to universal law as such, namely RU. Kant then implicitly identifies RU with the Categorical Imperative or, at the very least, claims that the former entails the latter.[25]

Obviously the two principles are not equivalent. Suppose someone acts on Kant's famous maxim of false promising: "When I believe myself in need of money, I shall borrow money and promise to repay it, even though I know that this will never happen" (GMS 422). If she acts on this maxim, then, for well-known reasons I need not here restate, she violates the Categorical Imperative.[26] But she does not necessarily violate RU. If she holds her acting on the false-promising maxim to be morally permissible, nothing need prevent her from judging that in circumstances relevantly similar to her own, someone else's acting on it would be morally permissible as well. And the notion that RU entails the Categorical Imperative has little, if any, more plausibility than the notion that the two principles are equivalent. Kant gives us no reason to think that someone who embraced RU would be rationally compelled also to endorse the Categorical Imperative. Once

again, it turns out that Kant's argument suffers from a glaring gap. Whether the practically uninformative principle is RU or L, Kant cannot legitimately move directly from it to the Categorical Imperative.

1.5 Terminological and Thematic Clarifications

This book explores responses to the common view, just elaborated, that Kant fails miserably at defending a foundational claim in this ethics, namely the claim that if there is a supreme principle of morality, then it is the Categorical Imperative.

Before sketching the book's structure, I need to make a few clarifications, some terminological, some thematic. I have used the term "the Categorical Imperative" to refer to the principle Kant states at *Groundwork* 421 (cited in 1.4) and variant expressions of this principle, such as the one he gives at *Groundwork* 402 (also cited in 1.4). Kant himself refers to this principle as the "categorical imperative," without capitalization (GMS 421). I have adopted the capitalization in order to emphasize that the term "categorical imperative" need not be used to refer to the particular principle Kant sets forth at *Groundwork* 421. In another, broader, Kantian usage, the term "categorical imperative" refers to any principle that is absolutely necessary and binding on all rational agents.[27] A categorical imperative in this sense is a "practical law" (GMS 420, 425, 428, 432; KpV 41). A burden of Kant's discussion in *Groundwork* I–II is to show that if there is a categorical imperative (that is also the supreme, practical norm for the moral assessment of action), then it is the Categorical Imperative. For the sake of clarity, I sometimes substitute the term "Formula of Universal Law" for the "Categorical Imperative."

In *Groundwork* II, Kant tells us that he has represented the supreme principle of morality in "three ways" (GMS 436). He has represented it in the Formula of Universal Law, as well as in two other formulas. These other two are often referred to in the Kant literature as the Formula of Humanity and the Formula of the Kingdom of Ends. The Formula of Humanity is this: "So act that you treat humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" (GMS 429, emphasis omitted). The Formula of the Kingdom of Ends seems to run as follows: "[A]ll maxims from one's own lawgiving are to harmonize with a possible kingdom of ends as with a kingdom of nature" (GMS 436).[28] According to Kant, these "three ways of representing the principle of morality are at bottom only so many formulas of the very same law, and any one of them of itself unites the other two in it" (GMS 436). So it seems that for Kant these three formulas are, in a practical sense, equivalent – for example, any action that is morally impermissible according to one is also morally impermissible according to each of the others.

In this book I discuss only the Formula of Universal Law and the Formula of Humanity, leaving aside the Formula of the Kingdom of Ends.[29] I focus

on the first two formulas because they are the most familiar and, I think, the most forceful Kantian candidates for the supreme principle of morality. Kant's claim that all three are formulas of the "very same law" appears to imply that the Formula of Universal Law and the Formula of Humanity generate the same results regarding the moral status of actions.[30] I do not believe that they do, but an account of why will have to wait until Chapter 8. Since I hold that the Formula of Universal Law (the Categorical Imperative) and the Formula of Humanity differ in their implications regarding the moral status of actions, I view them ultimately as competitors (albeit from the same stable) for status as the only viable candidate for the supreme principle of morality. This book considers derivations of two different Kantian candidates for the supreme principle of morality: the Formula of Universal Law and the Formula of Humanity.

1.6 Outline of the Book

Let me now explain briefly how the book unfolds and what it aims to show. According to a traditional and widely accepted reading, there is a conspicuous gap in Kant's *Groundwork* derivation of the Formula of Universal Law. The book is composed of two main parts. In the first, I criticize contemporary responses to the traditional interpretation; in the second, I construct a response of my own – a response that leads to a new approach to Kant's derivations of both the Formula of Universal Law and the Formula of Humanity.

If one accepts the traditional view that Kant's *Groundwork* derivation of the Formula of Universal Law plainly fails, it makes sense to look outside the *Groundwork* for a derivation of this principle. Henry Allison does just this. Appealing to the *Critique of Practical Reason*, Allison constructs an argument (available to Kant if not explicitly made by him) that, in Allison's view, establishes that if there is a supreme principle of morality, then it is the Formula of Universal Law. According to Allison, this argument succeeds whereas that of the *Groundwork* fails, since, unlike the latter, it relies on the assumption that rational agents have what Kant calls "transcendental freedom" – that is, "independence from everything empirical and so from nature generally" (KpV 97). I maintain in Chapter 2 that even if we accept Allison's use of the controversial notion of transcendental freedom, this derivation fails. In short, Allison claims that as transcendentally free, rational agents, we require a nonsensuously based justification of our maxims. Moreover, this justification must be the maxims' conformity to some practical law. But, concludes Allison, this law could only be the Formula of Universal Law. I argue that Allison does not successfully eliminate the possibility that conformity to some different law justifies our maxims.

Of course, the Formula of Universal Law is not the only principle Kant advocates. Among the others we find the Formula of Humanity, a principle

that many consider to be the most attractive Kantian candidate for the supreme principle of morality. Does Kant establish that if there is such a principle, then it is the Formula of Humanity? Chapter 3 focuses on this question. There are two key steps in this derivation, which Kant undertakes in *Groundwork* II. First, Kant claims that if there is a supreme principle of morality (and thus a categorical imperative), then there is an objective end: something that is unconditionally good. Second, he claims that this unconditionally good thing must be humanity. (If Kant proves these claims, he shows that if there is a supreme principle of morality, then humanity is unconditionally good. But if humanity is unconditionally good, Kant can go on to argue, then we are rationally compelled to do what the Formula of Humanity commands, that is, always to treat it as an end in itself.) Recently Christine Korsgaard has offered an influential reconstruction of Kant's defense of these two key steps, especially the second. I contend that despite Korsgaard's efforts, the defense of neither step is adequate. Kant falls far short of establishing that if there is a supreme principle of morality, then it is the Formula of Humanity.

Given the inadequacy of both Kant's *Groundwork* derivation of the Formula of Humanity and his second *Critique* derivation of the Formula of Universal Law (as reconstructed by Allison), the prospects for establishing that only a Kantian principle could be the supreme principle of morality seem very grim indeed. The second part of the book aims to show that we can make more progress toward establishing this than one might think.

Chapter 4 challenges the traditional reading of Kant's *Groundwork* derivation of the Formula of Universal Law, the reading according to which the derivation contains an unwarranted jump from a practically empty principle to this formula. The chapter introduces a new, *criterial reading* of the derivation, according to which it has three main steps. First, Kant develops criteria that any viable candidate for the supreme principle of morality must fulfill. These criteria include, but are not limited to, those that belong to his basic concept of this principle. Second, Kant tries to establish that no possible rival to the Formula of Universal Law fulfills all of these criteria. Finally, Kant attempts to demonstrate that the Formula of Universal Law remains as a viable candidate for a principle that fulfills all of them. With these three steps, Kant strives to prove that if there is a supreme principle of morality, then it is this formula. Defending a rejection of the traditional interpretation of this derivation in favor of the criterial reading obviously requires considerable textual analysis. Much of Chapter 4 focuses on difficult passages in the *Groundwork*, including the ones cited in 1.4. I aim to show that the text of Kant's derivation (in both *Groundwork* I and II) permits the criterial reading. At the end of Chapter 4, I offer a preliminary list of criteria, in addition to the ones contained in his basic concept, that Kant develops for the supreme principle of morality.

Chapter 5 focuses on this list of four criteria. How are we to interpret the criteria, and how does Kant defend them? The criterion that demands

most of our attention can be stated thus: the supreme principle of morality must be such that all and only actions conforming to this principle because the principle requires it – that is, all and only actions done from duty – have moral worth. An advocate of a particular principle as the only viable candidate for the supreme principle of morality must, according to Kant, be able (rationally speaking) to maintain that an agent's action has moral worth if and only if she does it from duty, that is, because this principle requires it. Chapter 5 probes both the meaning of this criterion and Kant's arguments for it.

It is one thing to understand this criterion and Kant's defense of it; it is quite another to embrace the criterion. Chapter 6 poses the question of whether we should do so. I argue that we should accept one part of the criterion (modified slightly) but reject another part. We should accept the idea that the supreme principle of morality must be such that *all* instances of willing to conform to it because the principle requires it have moral worth; but we should reject the notion that the supreme principle must be such that *only* instances of willing to conform to it because the principle requires it have moral worth. An advocate of a certain candidate for the supreme principle of morality, say the Formula of Universal Law, must acknowledge that an agent's action can have moral worth even if she does not do it because this principle requires it. Indeed, I argue that Kantian considerations rationally compel the advocate to acknowledge that actions *forbidden* by the Formula of Universal Law can have moral worth.

By the end of Chapter 6 we will have a complete list of Kant's criteria for the supreme principle of morality. In addition to the four that belong to Kant's basic concept of this principle, there are four others, modified in accord with the argument of the chapter. According to these, the supreme principle of morality must be such that: (v) every case of willing to conform to it because the principle requires it has moral worth; (vi) the moral worth of willing to conform to the principle because the principle requires it stems from its motive, not from its effects; (vii) an agent's representing the principle as a law, that is, as a universally and unconditionally binding principle, provides him with sufficient incentive to conform to it; and, finally, (viii) a plausible set of duties (relative to ordinary rational moral cognition) can be derived from the principle.

The first step of Kant's derivation is to establish criteria for the supreme principle of morality; the second is to show that no possible rival to the Formula of Universal Law fulfills all of them. Chapter 7 focuses on this second step. In the first instance, the criterial reading I defend is a reading of Kant's derivation of the Formula of Universal Law. It is, however, open to Kant to employ the same steps in deriving the Formula of Humanity. In any case, the chapter tries to show that with the help of some of these criteria – ones the plausibility of which I defend – Kant can eliminate key competitors to both of these principles. For example, relying on criteria v and vi, Kant is able to construct a kind of argument, which I call a "valuational argument," that

succeeds in eliminating many consequentialist candidates for the supreme principle of morality, including a utilitarian principle such as: "Always perform a right action: one that yields just as great a sum total of well-being as would any alternative action available to you." However, the valuational type of argument does not apply to nonconsequentialist principles, such as this detheologized imperative based on the Ten Commandments: "You ought to honor your father and mother; you ought not to kill; you ought not to commit adultery; you ought not to steal; you ought not to bear false witness; you ought not to covet anything that is your neighbor's." But, as Chapter 7 also tries to show, Kant is not without effective recourse against such principles.

To complete the second step of his derivation of the Formula of Universal Law, Kant must demonstrate that no possible rival to this principle fulfills all of the criteria he develops. He must eliminate not just a few familiar rivals but *all possible* principles other than the Formula of Universal Law as contenders for the supreme principle of morality. Yet, from the outset, it is hard to see how Kant could eliminate all possible contenders, if only because it is unclear how he could prove that he had even taken all of them into account. In my view, Kant does not prove this. I do not claim that Kant successfully dismisses all rivals to the Formula of Universal Law (or that he could successfully dismiss all rivals to the Formula of Humanity). I do, however, defend the view that he presents compelling arguments against some main rivals, including many consequentialist principles.

On the criterial reading, the third step of Kant's derivation of the Formula of Universal Law is to show that, unlike its rivals, this principle remains as a viable candidate for one that fulfills the whole set of criteria Kant has developed for the supreme principle of morality. (Showing that this formula *actually does* fulfill the whole set of criteria would involve giving a deduction of it. One of the criteria, one that belongs to Kant's basic concept, is that the supreme principle of morality be binding on all rational agents. No derivation could show even that the Formula of Universal Law is binding on all human rational agents, that is, that all of us are genuinely obligated to conform to it. A deduction, not a derivation, of the Formula of Universal Law would be needed for this.) In Chapter 8 I argue that the Formula of Universal Law stands as a viable candidate for fulfilling Kant's basic concept of the supreme principle, if we are willing to modify this concept slightly to accommodate my criticisms in Chapter 6 of how Kant views the relations between the supreme principle of morality and moral worth. The Formula of Universal Law is also not disqualified by three of Kant's further criteria. However, a serious problem arises regarding the fourth additional criterion, namely the one according to which the supreme principle of morality must be such that a plausible set of duties (relative to ordinary rational knowledge of morals) can be derived from the principle. The Formula of Universal Law is difficult to interpret; there is much debate about how, precisely, to apply it

in determining whether acting on a particular maxim is morally permissible. So the question remains: which duties stem from it? I do not offer anything approaching a thorough discussion of this question. But I try to show that on some leading interpretations of the Formula of Universal Law, this principle fails to generate moral prescriptions that square with common sense.

As I mentioned earlier, the criterial reading applies in the first instance to Kant's *Groundwork* derivation of the Formula of Universal Law. Yet there seems to be no reason why Kant could not take the same steps in a derivation of the Formula of Humanity that, according to this reading, he goes through in his derivation of the Formula of Universal Law. (If, as I hold, the two formulas are not equivalent, then a successful derivation of the latter would actually preclude a successful derivation of the former.) I argue in Chapter 8 that, like the Formula of Universal Law, the Formula of Humanity remains as a viable candidate for a principle that satisfies Kant's basic concept of the supreme principle of morality (if we modify this concept slightly), as well as three of the four further criteria Kant develops. But does the Formula of Humanity generate a plausible set of moral prescriptions? This question is difficult, since the Formula of Humanity itself poses interpretive challenges. Without pretending to give a full treatment of the issue, I argue that the Formula of Humanity holds more promise on this score than does the Formula of Universal Law, although it too has some troubling aspects.

This is where the book ends. It begins in Chapter 1 with a brief examination (too brief, I am afraid, to be entirely satisfactory) of some basic concepts in Kant's theory of agency. We have already invoked the notions of a maxim, the will, acting from inclination, and so forth. We need to clarify them in order to proceed without confusion.

As is already apparent, the book focuses mainly on arguments Kant makes in the *Groundwork* and the second *Critique*, since these are the works in which Kant is concerned with deriving the supreme principle of morality. Of course, I invoke discussions in Kant's other works in ethics, for example, the *Metaphysics of Morals*. However, the book does not in any way aim to give a comprehensive account of Kant's ethical doctrine.

In sum, the book sets out a new reading of Kant's *Groundwork* derivation of the Formula of Universal Law. It tries to show that this argument is philosophically far richer than the traditional interpretation suggests. No, Kant does not succeed in proving his strikingly ambitious claim that if there is a supreme principle of morality, then it is the Formula of Universal Law. But he does offer some strong reasons for rejecting rivals to this principle. What is more, Kant's derivation of the Formula of Universal Law opens the door to a heretofore unexplored way of defending the Formula of Humanity, a principle that many of us find especially attractive as a candidate for the supreme principle of morality.

1

Fundamental Concepts in Kant's Theory of Agency

1.1 Aims and Limits of the Discussion

Kant peppers each of his major works in practical philosophy with comments pertaining to what it means for us, rational agents, to act. Philosophers disagree on how best to interpret these comments, which are often difficult and sometimes obscure.[1] I offer some readings here that, I believe, cohere with Kant's texts, but they are surely not the only defensible readings. My aim in this chapter is to set out a plausible interpretation of (part of) Kant's theory of agency, an interpretation that will be useful as a reference point in discussions to come. Important issues regarding Kant's theory of agency, such as whether Kant does or should conceive of acting on a maxim on the model of Aristotle's practical syllogism, are not addressed here. A thorough reading of Kant's theory of agency, let alone a defense of it, would require a book in itself.

The chapter is divided into two main parts. The first focuses on a few key concepts in Kant's theory. In 1.2–3, I offer an account of Kant's notion of a maxim; then I turn very briefly to Kant's conceptions of the will (1.4) and of the will's "determining grounds" (1.5). The second main part of the chapter concerns Kant's account of actions *not* done from duty, that is, ones done on "material practical principles" (1.6–8). Understanding this account requires some painstaking textual analysis. I explain in section 1.6 why, in light of the main aims of this book, it is important to grasp Kant's account of actions not done from duty.

1.2 Maxims: A Basic Account

Let us begin, then, with the concept of a maxim. Kant tells us that a maxim is a subjective principle of acting (GMS 421, note).[2] By following Rüdiger Bittner and considering the sense in which a maxim is a *subjective* principle and that in which it is a principle of *acting*, we can develop a basic account

of maxims.[3] Having an example of a maxim at hand helps us to do so. Suppose that Mary has adopted the maxim M: From self-love, I will shorten my life when its longer duration threatens more troubles than it promises agreeableness.[4] A maxim is subjective in three respects. First, if there is a maxim, then there is a subject – that is, an agent – who holds it. A maxim is always some agent's rule. If neither Mary nor anyone else held M, then it would not be a maxim.[5] Second, an agent chooses his own maxims. Kant calls maxims "rules imposed upon oneself" (GMS 438). At any time he is free to discard the maxims he presently holds and to adopt new ones. Mary may have held M for the past thirty years, but it is up to her whether she will hold it even for the next thirty seconds. Third, an agent's maxim is a subjective principle in that it applies only to her own action (KpV 19). Mary's maxim expresses what she requires herself to do if continuing to live threatens more evil than satisfaction for her. It does not tell anyone else what he is required to do in these circumstances.

Maxims are not just subjective principles; they are subjective principles of acting. Agents act *on* (*nach*) maxims. This means that maxims play a role in the generation of their actions. An agent does not merely apply a maxim in hindsight to his action after it has occurred. If Mary has acted on M by taking poison, then M, or, more likely, a less precise representation of it, has contributed to the generation of her action. Of course, that someone has adopted a maxim – that is, given herself the requirement of acting in a certain way under certain circumstances – does not entail that she will act on it. The occasion for acting on it may simply never arise. Mary may never come to believe that her life's continuing threatens more troubles than agreeableness. Even if the occasion for acting on a maxim does arise, an agent is free not to act on it. She may just choose not to abide by the principle of action that she has given herself. Although faced with the prospect of a miserable old age, Mary might obey the Categorical Imperative and refrain from acting on M, that is, refrain from killing herself.[6]

Philosophers typically hold that for Kant, all acting is acting on a maxim.[7] It is not hard to defend this interpretation. According to Kant, all of an agent's actions are either morally permissible or morally impermissible.[8] The Categorical Imperative – "Act only on that maxim through which you can at the same time will that it become a universal law" (GMS 421, emphasis omitted) – gives us a procedure for determining whether an action performed *on a maxim* is morally permissible. A person's action is morally permissible only if she can will the *maxim on which she performs it* to become a universal law. If she cannot do so, then the action is morally impermissible. The principle does not give us a procedure for determining whether an action performed on no maxim is morally permissible. Kant, of course, takes the Categorical Imperative to be the *supreme* principle of morality. He suggests that it is the canon of the moral estimation of our action as a whole (GMS 424). If there were questions of moral permissibility to which the test

embodied in the Categorical Imperative could give no answer, then Kant's claim that this imperative is the supreme principle of morality would be hollow. With these considerations in mind, it is easy to show that, for Kant, all acting is acting on a maxim. Suppose that agents could perform actions without doing so on any maxim. The Categorical Imperative procedure would then yield no answer to the question of their moral permissibility, and the Categorical Imperative would thus not be the supreme principle of morality. Since Kant affirms it to be the supreme principle of morality, he must hold that agents perform each and every one of their actions on a maxim.

Kant's own examples of maxims illustrate that, at a minimum, they are rules that specify a type of action to be performed in a type of situation, for example, "When I believe myself to be in need of money, I shall borrow money and promise to repay it, even though I know that this will never happen" (GMS 422). When fully specified, however, it seems that a maxim also includes a description of the agent's end in doing what she does. In the *Groundwork* and in the *Metaphysics of Morals*, Kant suggests that all maxims contain a (description of) an end (GMS 436, MS 395).[9] The end implied in the maxim of false promising is presumably that of getting money. Moreover, some of the maxims Kant discusses contain descriptions of an incentive, for example, the maxim on which Mary's maxim is based: "*From self-love*, I make it my principle to shorten my life when its longer duration threatens more troubles than it promises agreeableness" (GMS 422, emphasis added). Here the agent's end, that is, the state of affairs he would aim to realize if he acted on the maxim, remains implicit, although it is obviously something like that of being free from that suffering which is not outweighed by happiness. The agent's incentive – that which would motivate him to act if he acted on the maxim – is explicit; it is "self-love."[10]

The notion that when fully spelled out, maxims contain descriptions of an agent's incentive for acting gains support from Kant's well-known claim in the *Religion within the Limits of Reason Alone* that the "freedom of the will [*Willkür*] is of a wholly unique nature in that an incentive can determine the will to an action *only so far as the individual has incorporated it into his maxim*" (Rel 23–24, English ed. 19). Later, in connection with Henry Allison's attempt to fill the (apparent) gap in the *Groundwork* derivation, we discuss this claim in detail. For now, note that, in Kant's view, we have freedom of the will. Moreover, if our will is determined to an action, some incentive constitutes a basis for this determination.[11] All of our actions are such that we have some incentive for performing them. (The typical sneeze or slip on a banana peel does not count as an action in the relevant sense.) Therefore, Kant's claim in the *Religion* implies that whenever we act, we do so on some maxim that, if fully specified, would include a description of our incentive for acting.[12] A fully expressed maxim would include not only a description of a kind of action to be performed in a kind of situation, but also a specification of the agent's end and of his incentive in performing it. A fully

expressed maxim would take the form of a rule that includes each of these elements. Of course, when we act, we might not have each of these elements in mind.[13]

1.3 Maxims and Other Rules of the Same Form

Before ending our discussion of maxims, we need to address one more issue, namely that of how to distinguish them from other rules of the same form. This issue is important. Suppose that someone in taking a karate lesson acts on the rule: "From self-love, every Monday at 3 P.M. I take live karate lessons in order to improve my endurance and flexibility." It seems reasonable to assume that, *at the same time*, she might also be acting on a different, more general rule: "From self-love, during my free time I exercise in order to stay in shape." If we took both rules to be maxims on which the agent acted, then Kant would face a serious problem. At least on one common reading, acting on the first rule would violate the Formula of Universal Law, whereas acting on the second would not. I take it to be obvious that acting on the rule of exercising during one's free time is in accordance with this formula. But consider the rule of taking karate lessons with a live instructor on Monday at 3:00 P.M. Not every agent *could* take live karate lessons Monday at 3:00 P.M. An agent cannot take a live lesson without a live instructor. But if all agents were taking live karate lessons Monday at 3:00, then there would be no instructor available to *give* lessons at this time. Given that not every agent could take live karate lessons every Monday at 3:00 P.M., it is not possible (as a rational being) to will that it become a universal law that every agent does so.[14] If both rules count as maxims, then it seems that our agent's action of taking a karate lesson is morally impermissible. For she is acting on a maxim such that she cannot, at the same time, will that it become a universal law. To avoid the difficulty suggested by this example, we must have a means of deciding which of the rules an agent acts on counts as the maxim of his action.

Unfortunately, Kant does not explain how to do this. The best way in my view is to specify that the maxim of an agent's action is the *fundamental* rule, of the form required of a maxim, on which he acts.[15] (Recall that, at least implicitly, a maxim must have the form of a subjective rule according to which, from a specified incentive, an action is to be taken in designated circumstances in order to realize some end.) More specifically, a practical rule Q of the requisite form has status as the fundamental rule of this form on which an agent performs an action when it fulfills either one of the following two conditions: Q is the only such practical rule on which he performs the action; or Q is not the only such rule on which the agent performs the action but is rather the most general rule of this form on which he does so. If Q fulfills this second condition, it governs the agent's selection of a more specific rule of the same form, that is, a rule ancillary to

Q, through which rule he implements Q (performs an action). The practical rule "From self-love, every Monday at 3:00 P.M. I take live karate lessons in order to improve my endurance and flexibility" is an example of one that might be ancillary to the maxim "From self-love, whenever I have free time, I exercise in order to stay in shape." An agent who adopted the latter might take up the former as a rule for *implementing* it. She would presumably do so because, as it happens, she has Monday afternoons free, wants to improve her endurance and flexibility, and judges that training in a martial art would be a good way of doing so. Given her circumstances, she would choose to act on her maxim by acting on this more specific rule. Of course, another agent who had adopted this maxim might choose a different rule through which to act on it.

In sum, a maxim is a subjective principle of acting. It is a subjective principle in that it is held by some agent, it can be freely adopted or discarded by her, and it applies only to her own actions. An agent's maxims are principles of acting in that they play a role in the generation of her actions. When fully expressed, a maxim includes a description of a kind of action to be performed in a kind of situation, as well as a specification of the agent's end and incentive in performing it. Not all rules of this form count as maxims, however. An agent's maxim is the fundamental rule of this form on which she acts. This reading of Kant's views regarding maxims is by no means thorough (or thoroughly defended), but it will, I hope, serve to fix ideas for discussions to come.

1.4 The Will

Another key concept in Kant's theory of agency is that of the will. Unfortunately, Kant's account of the will is a terminological mire. In the *Groundwork* and the second *Critique*, he typically uses *Wille* to refer to an agent's capacity to act on rules, for example, maxims or imperatives (see, e.g., KpV 32).[16] But he also uses *Wille* to refer in addition to an agent's capacity to give herself the rules on which she has the capacity to act, for example, to legislate for herself maxims or imperatives (e.g., GMS 431 and KpV 33). Later, in the *Metaphysics of Morals*, Kant typically employs *Wille* to refer *only* to the latter capacity (e.g., MS 213). We might call an agent's capacity to *act* on rules the "executive *Wille*" and his capacity to *give himself* these rules the "legislative *Wille*."[17] In the *Metaphysics of Morals* (and elsewhere) Kant employs another term, *Willkür*, that is sometimes translated as "will."[18] For our purposes, it will be safe to consider *Willkür* as the same capacity as executive *Wille*, that is, the capacity to act on rules.[19]

Fortunately, we need to focus only on Kant's notion of the executive *Wille*, to which I refer here simply as the will. According to Kant, to exercise the capacity of will – that is, to will – is to act. That is why Kant defines the (executive) *Wille* as the capacity to *act* on principles (GMS 412). Willing is

more than wishing or even deciding to do something. Someone might wish or decide to realize some object (e.g., to get away for a weekend at a bed and breakfast) yet change his mind and never actually make any effort to realize this object (e.g., never do any planning for the getaway). Willing involves making some effort to realize what one wills. In this sense, it is a kind of acting. In what follows, I alternate between speaking in terms of willing and in terms of acting. For our purposes, the two amount to the same thing: trying (on the basis of some rule[s]) to secure some objective.

1.5 Determining Grounds of the Will

The will is a capacity to act on rules. But what is a "determining ground" of the will? As determining grounds of the will, Kant mentions (at least) ends, inclinations, the expectation of pleasure, the principle of one's own happiness, and the moral law (see respectively MS 381; KpV 81, 22, 35, 72). I assume that each of the determining grounds (*Bestimmungsgründe*) of the will he mentions counts as such by standing in some particular relation to willing. But, to my knowledge, Kant never says explicitly just what this relation is. It seems to me plausible to interpret determining grounds of the will as motivating reasons or, more simply, motives for willing. They are what bring about willing. In Kant's view, however, each item on the list actually brings about an agent's willing only if she has taken account of it in her maxim, that is, made it part of a rule on which she acts. In other words, each of these items on its own might count as an incentive for an agent's acting, but the items actually motivate her to act only if she has incorporated them into some self-given rule.[20] For example, an agent might have an inclination to eat ice cream. But, according to Kant, this inclination determines her will (i.e., actually motivates her) only if she has taken account of it in some maxim – for example, one of allowing herself small pleasures to promote her happiness.[21]

One might wonder whether determining grounds of the will count not only as motivating but also as "justifying" reasons for acting. That depends on the sense of justifying reason one employs. Let us consider one particular kind of determining ground of the will, namely inclinations. Obviously, that someone has a particular inclination as a motive does not entail that, from an impartial perspective, her acting on this motive *is* justified. (Acting from the inclination to be the richest person in the county, a businessperson might hire someone to kill her competitor.) Determining grounds of the will are not justifying reasons in the sense of reasons that, from an impartial perspective, always do in fact justify an agent's action. Moreover, that an agent has a particular inclination as a motive does not even entail that, from her own perspective, her acting on this motive actually justifies her action. If a particular inclination serves as an agent's motive in acting, then she has incorporated this motive into one of his maxims. But she might

hold that her acting on this maxim is itself ultimately unjustified because it is *morally* unjustified. For example, if the agent has Kantian leanings, she might believe that her indulging her inclination to be the richest person in the county by acting on a maxim of ordering a hit on her competitors is contrary to Kantian duty and therefore ultimately unjustified.

However, Kantians have recently emphasized that, as a rational being, an agent must believe that acting on her maxim is *in some sense* good or rationally justifiable.[22] If she does not meet this "justification requirement" by holding that acting on the maxim is good morally, she must meet it by holding that acting on it is good prudentially. She would, for example, meet the requirement by virtue of believing that, given her end (e.g., to be the richest person), taking the means to it specified in the maxim (e.g., killing her competitor) is good in that it will likely be effective. In short, although a given determining ground of the will need not constitute a reason that actually justifies what an agent does, either from an impartial or from even her own perspective, she must hold that it is good, in some sense, for her to act on the maxim in which this determining ground has been incorporated.

1.6 Acting from Inclination: Three Interpretations and Their Importance

This brief examination of maxims, the will, and determining grounds of the will puts us in position to do some final stage setting for the main arguments of this book. In sections 1.6–8, we focus on Kant's account of actions that are *not* done from duty.[23]

Since these sections involve painstaking textual analysis, it is helpful before proceeding to have some idea of how they further the main aims of this book. In Chapter 4, I begin to defend a *criterial reading* of Kant's derivation of the Categorical Imperative. According to this reading, Kant develops criteria for the supreme principle of morality. He then tries to show that no rival to the Categorical Imperative for status as this principle can fulfill the full set of criteria. Finally, Kant suggests that the Categorical Imperative remains as a viable candidate for fulfilling the full set. So Kant's criteria for the supreme principle of morality are obviously crucial to my reading of his derivation. One criterion he develops is the following: the supreme principle of morality must be such that all and only actions done because the principle requires it – that is, all and only actions done from duty – have moral worth. It is not possible to comprehend this criterion, let alone to gauge its plausibility, without grasping what, according to Kant, it means to act from duty. But in order to grasp this we need to understand Kant's account of actions not done from duty. For example, only by understanding this account can we see that for Kant all actions done from duty are done from duty alone. For Kant there simply are no "overdetermined" actions, ones done (at the same time) from both duty and inclination (section 5.3). Since Kant's criterion

does not allow that an action can be done from both duty and inclination, it implies the view that absolutely no actions have moral worth other than those done *exclusively* from the incentive of conforming to moral principle. (In Chapter 6 I argue that this view is implausible. Kant should drop it from his criterion and maintain instead merely that all actions from duty have moral worth.)

In the spirit of Kant's practical philosophy, though not in its idiom, we might call actions not done from duty "nonmoral" actions. For Kant, of course, not all nonmoral actions are immoral. A nonmoral action can be morally permissible: even though it is not done from duty, it can be in accordance with it – for example, the action of a shopkeeper not overcharging an inexperienced customer (GMS 397). According to Kant, all nonmoral actions – that is, all actions not done from duty – are done from inclination (GMS 413, note).

Many philosophers believe that Kant defends a radically hedonistic account of non moral action. According to the traditional interpretation, Kant holds that whenever an agent acts nonmorally, she is motivated solely by the desire for pleasure.[24] Pointed criticisms of Kant have arisen from the notion that he embraces this account, with one philosopher going so far as to charge that Kant's account is not only false, but "utterly repugnant, derogatory, and degrading."[25] The most obvious objection to the account is that it fails to square with the phenomena. Agents seem to be motivated by more than a desire for pleasure, even when they are not acting from duty. Consider a serious pianist who in practicing a sonata is acting solely from her inclination to master the piece. Depending on the circumstances, many of us would find plausible her opinion that her motivation for practicing includes a desire to play the piece beautifully: a desire that she does not aim to satisfy for the sake of the pleasure its satisfaction promises. If the traditional reading is correct, then Kant defends a suspect account of nonmoral action.

Recently Andrews Reath has offered an innovative and influential argument against the traditional construal of Kant's account.[26] Philosophers have misinterpreted the relations Kant believes to hold between pleasure and inclinations, says Reath. Contrary to the traditional reading, Kant does not claim that in trying to satisfy an inclination, an agent is always motivated by the prospect of gaining pleasure for herself. He claims rather that pleasure plays a role in the development of inclinations.[27] An agent would not develop an inclination for an object, say, mastering a piano sonata, unless she expected that she would gain pleasure from realizing it. Once an agent has an inclination for an object, however, in pursuing it she need have no hedonic motivation at all. Once she has an inclination to master a sonata, the agent's motives in practicing it need not include her own pleasure.

I trust that the appeal of Reath's interpretation is evident. Unfortunately, the interpretation fails to cohere with Kant's doctrine, or so I contend. Examination of Kant's definitions of inclination, as well as some of his remarks on

material practical principles, suggests that he did indeed hold each action from inclination to have hedonic motivation. Nevertheless, for philosophers sympathetic to Kant but not to a radically hedonistic account of nonmoral action, all might not be lost. In my view, although Kant's assertions permit a reading on which an agent's own pleasure constitutes her only motive in acting nonmorally, they do not require it. They also permit the interpretation that, whenever an agent acts from inclination, she has her own pleasure as one, but not necessarily as her only, motive. I call this the "alternative interpretation."[28]

The alternative interpretation seems more attractive than the traditional one. According to the former, if, from inclination, an agent writes a short story or practices the piano, one of her motives must be her own pleasure. Yet at the same time she might have other motives: the desire to exercise her creativity or to play beautifully: desires the agent does not strive to satisfy for the sake of pleasure. On the alternative, Kant avoids the suspect reduction of all nonmoral motives to one. He can acknowledge some of the complexity of acting in ways other than from duty. As we will see, however, the traditional interpretation fits more naturally with some of Kant's claims in the second *Critique* than does the alternative.

1.7 Acting from Inclination in the *Groundwork* and in the *Metaphysics of Morals*

To construct an interpretation of Kant on nonmoral action, we must engage in close reading of some difficult passages. To begin, in an often overlooked footnote in the *Groundwork* Kant offers a dense definition of inclination:

> The dependence of the capacity of desire on sensations is called inclination, and inclination always indicates a *need*. The dependence of a contingently determinable will on principles of reason, however, is called an *interest*. An interest is present only in a dependent will, which is not of itself always in conformity with reason; in the divine will we cannot conceive of an interest. But even the human will can *take an interest* in something without therefore *acting from interest*. The former signifies the *practical* interest in the action; the latter, the *pathological* interest in the object of the action. The former indicates only the dependence of the will on principles of reason *in themselves*, while the latter indicates the dependence of the will on principles of reason for the sake of inclination, since reason gives only the practical rule by which the needs of inclination are to be aided. In the former case the action interests me, and in the latter the object of the action (so far as [*sofern*] it is agreeable to me) interests me. (GMS 413, note)

We need to go carefully in order to understand the note's main points.

As a first step, let us focus on Kant's notion of the capacity of desire (*Begehrungsvermögen*). Although it has largely been neglected, this notion is one of the most fundamental in Kant's theory of agency.[29] An agent's capacity of desire, says Kant, is her capacity to cause, through her representations

of objects, the reality of these objects (KpV 9, note; KU 177, note; MS 211). The term "representation" (*Vorstellung*) refers here to a *mental* representation, that is, an idea; "object" refers to a state of affairs or to an event. An agent who had an idea of an object and who brought about the object through this idea would count as having exercised her capacity of desire with respect to this object. For example, a person who had an idea of catching a butterfly and who, guided by this idea, caught one would count as having exercised her capacity of desire with respect to catching the butterfly.[30] It is crucial to recognize that by Kant's definition the capacity of desire is *not* a capacity to have or to acquire a desire. Rather, it is a capacity to try to realize a desired object. It is a capacity to *act on* a desire.

In the *Groundwork* footnote, Kant says that inclination is the dependence of the capacity of desire on sensations. When an agent acts from inclination, suggests Kant, his capacity to realize an object through his idea of it is dependent on sensations. Kant gives only an indirect answer to the question of how this capacity is dependent on sensations, an answer that emerges from his discussion of the concept of interest. Kant defines an interest as the dependence of a contingently determinable will – for example, the human will – on principles of reason. The human will is by definition dependent on principles of reason. For whenever an agent exercises her will, she does so on at least one such principle (GMS 412). Thus, whenever an agent acts, she has some interest. Kant distinguishes *practical* from *pathological* interest. He identifies a practical interest as an interest in an action itself. An agent, he says, takes a practical interest in an action when she acts from duty. A pathological interest is an interest in the object (i.e., end or aim) of an action, rather than in the action itself.[31]

Kant claims that when an agent acts from a pathological interest in the end of an action, the end interests him "so far [*sofern*] as it is agreeable" to him. In other words, to act from pathological interest is to act to realize an end that one is interested in realizing so far as he expects that its realization would give him pleasure.[32] Yet what does it mean to be interested in realizing an end, *so far as* one believes that its realization would give him pleasure? On my view, Kant's text permits two different readings of this notion: the first leads us to the traditional interpretation of acting from inclination; the second, to the alternative interpretation. According to the first, to be interested in realizing an end so far as one believes that its realization would give him pleasure amounts to being interested in the end *to the extent that* one expects to gain pleasure from its realization. The more pleasure one expects to gain from realizing the end, the more interested one is. Since, according to this reading, one's pathological interest in an end is directly proportional to the pleasure one expects from it, it is natural to assume that when one *acts from* pathological interest in the end, pleasure from it is one's only motive.

According to the second reading – the one that leads to the alternative interpretation – Kant holds that a *necessary condition* for the agent's interest

in the end is that he believe that its realization would give him pleasure.[33] Kant conceives of acting from pathological interest in an end as trying to realize the end on condition that one expect pleasure from doing so. For example, when, from pathological interest, someone attempts to master a piano sonata, her attempt is conditional on her expectation that mastering it would give her pleasure (see also GMS 442). As Kant makes clear in the note, acting from pathological interest amounts to acting "for the sake of inclination." In effect, Kant equates acting from pathological interest with acting from inclination. Therefore, according to our second reading, Kant conceives of an agent's acting from inclination as her trying to realize an end *only if* she expects the end's realization would give her pleasure.[34] Strictly speaking, that an agent performs certain actions only on the condition that she expects pleasure from doing so does *not* entail that she has hedonic motivation in performing them. After all, the pleasure the agent necessarily expects when she acts from inclination might be instrumental to, or serve merely as a sign for the attainment of, some further end she has. However, Kant does not seem to have these possibilities in view. He seems to embrace the notion that in acting from inclination an agent always has some hedonic motivation. In the second *Critique*, for example, Kant (as we will see) clearly suggests that when an agent acts on material practical principles (i.e., from inclination), his expectation of gaining pleasure constitutes a determining ground of his acting.

Kant's account of inclination in the *Groundwork* note weighs against Reath's interpretation. Reath asserts that Kant holds pleasure to play a role in the development of inclinations. This assertion seems true (see KU 207). But, as his remarks regarding Kant's famous example of the "philanthropist" (or "friend of humanity") will soon reveal, Reath also suggests that once an agent has developed a Kantian inclination, it is *not* the case that he acts from it only on condition that he expect pleasure from his action. This suggestion seems misguided. In the note, Kant strongly implies that an agent's expectation of experiencing pleasure plays a role *each time* he acts from inclination: a role as a motive for acting on the alternative interpretation; a role as the agent's only motive for acting on the traditional interpretation.

In his interpretation of Kant's account of inclination, Reath does not mention the *Groundwork* note. Nevertheless, he does appeal to the *Groundwork* to bolster his rejection of the traditional interpretation of acting from inclination. In particular, Reath appeals to the example of the philanthropist, a person who helps others not from duty but rather from inclination. According to Reath, Kant holds the following: "The object of [the philanthropist's] concern and the motive of his actions is their [others'] happiness."[35] The philanthropist, Reath unambiguously suggests, does not have the expectation of his own pleasure as a motive in helping others. On Reath's interpretation, Kant rejects the notion that an agent's expectation of his own pleasure constitutes a motive in all acting from inclination.

But a close look at Kant's remarks regarding the philanthropist will, I believe, show this interpretation to be flawed. Kant says: "To be beneficent where one can is a duty; and besides this, there are many souls so sympathetically constituted that, without any further motive of vanity or self-interest, they find an inner pleasure in spreading joy around them and can rejoice in the satisfaction of others, so far as it is their own work" (GMS 398). Here Kant speaks of sympathetically constituted persons, of whom the philanthropist is one, who find an "inner pleasure" (*inneres Vergnügen*) in spreading joy around them. According to Reath, Kant holds that, in their helping actions, such persons have improving the lot of others as their only motive. The "inner pleasure" they experience stems from their belief that they have actually managed to spread joy to others. The attractiveness of this interpretation is evident. It suggests that Kant understood the motives of sympathetically constituted persons much as many of us do. But I find this interpretation questionable. Kant does not state here that these persons fail to have their own pleasure as a motive. He does not say that without any motive of vanity or self-interest they try to help others. If Kant did assert this, then Reath's interpretation would obviously gain support. What Kant does say here is that without any *further* (*anderen*) motive of vanity or self-interest, the sympathetically constituted find pleasure in spreading joy to others. This statement leaves open the possibility that, on Kant's view, these persons do have a motive of self-interest: the pleasure they expect to gain from spreading joy to others. But they have no further motive of self-interest: they are not, for example, prompted to act by the expectation that those they help will render them some service in the future. Kant's discussion of sympathetically constituted persons, of whom the philanthropist is one, does not seem to justify Reath's rejection of the traditional (and presumably the alternative) reading of acting from inclination.

Reath bases his interpretation of inclination mostly on Kant's *Metaphysics of Morals* definition. But I argue that this definition, like the *Groundwork* one, fails to support his view. Instead, it lends credibility to the view that Kant must have embraced either the traditional or the alternative interpretation.

In his discussion of agency in the Introduction to the *Metaphysics of Morals*, Kant offers another dense and difficult definition of inclination:

> As for practical pleasure, that determination of the capacity of desire which must be *preceded by* this pleasure as cause is called *desire* [*Begierde*] in the narrow sense; habitual desire in this narrow sense is called *inclination* [*Neigung*]; and the connection of pleasure with the capacity of desire, provided that the understanding judges this connection to hold as a general rule (though only for the subject), is called *interest.* So if a pleasure necessarily precedes the determination of the capacity of desire, the practical pleasure must be called an interest of inclination. (MS 212)

The first aspect of this definition to notice is that Kant is employing the term "inclination" in a slightly different way than he does in the *Groundwork.*

Here Kant suggests a distinction between inclination and whim. We may say that a person has an inclination to begin her mornings with a cup of coffee. She habitually desires to begin her mornings this way. But suppose a person experiences a never-before-entertained desire to eat asparagus sauteed in raspberry jam. If we employ the sense of inclination contained in this definition, we may not say that she has an inclination for the dish.[36] We may, however, say that she has a "desire in the narrow sense" for it. Both inclinations and what I have called whims count as such desires. In his *Groundwork* definition of inclination as the dependence of the capacity of desire on sensation, Kant does not distinguish between inclination and whim. Since we are interested in Kant's general account of action not performed from duty, we can safely bracket this distinction. Important to us is what Kant says about desires in the "narrow sense," which we, following Kant's own *Groundwork* usage, call "inclinations."[37]

For our purposes, the central assertion in the *Metaphysics of Morals* passage is D: "That determination [*Bestimmung*] of the capacity of desire which must be preceded by pleasure as cause is called inclination." Reath argues that D amounts to the following: an inclination is a desire for an object such that before an agent can *come to have it*, she must *at some point* have determined that the realization of the object would give her pleasure.[38] So, for example, before I can count as having an inclination to play basketball, I must come to the view that playing would give me pleasure. Moreover, suggests Reath, D does *not* imply that once an agent has an inclination, whenever he tries to satisfy it, he must do so on the basis of his expectation that its satisfaction would give him pleasure. D does not imply that once I have an inclination to play basketball, every time I try to satisfy it I do so on the basis of my expectation that playing would give me pleasure.

Reath's interpretation is, I believe, based on a misunderstanding of Kant's notion of the capacity of desire. In D, claims Reath, Kant is merely pointing out a condition that must be fulfilled in order for an agent to come to have an inclination. Apparently, Reath takes the truth of this claim to be obvious. It would indeed seem obvious, if one made, as Reath apparently does, the following assumption: the capacity of desire is a capacity to *have* or to *develop* desires, including inclinations. Under this assumption, D seems to set out a necessary condition for the development of an inclination, namely that feelings of pleasure play a causal role in this development. Recall that D reads: "That determination of the capacity of desire which must be preceded by pleasure as cause is called inclination." The "determination" of this capacity would, under this assumption, presumably amount to the acquiring of a desire. D seems to specify that an inclination is a desire that an agent acquires in a certain way: by being prompted by feelings of pleasure (either experienced or expected) to do so. As we have noted, however, the assumption in question is false. Although in light of its name it is tempting to think otherwise, the capacity of desire is not a capacity to come to have a desire. Rather,

it is the capacity to realize an object through one's representation of it. What, then, is the "determination" of this capacity? To my knowledge, Kant never answers this question explicitly. Nevertheless, it is natural to suppose that determining the capacity of desire amounts to choosing to realize an object. It amounts to setting oneself to bring the object about. In effect, for an agent to determine her capacity of desire is for her to choose to realize the object of a desire.[39]

We can now see that, according to D, acting from inclination involves making a choice to realize an object, which choice is "preceded by pleasure as cause." D asserts: that choice to realize an object, which must be "preceded by pleasure as cause," is called inclination. But what would it mean for an agent's choice to realize an object to be "preceded by pleasure as cause"? We find an important clue for interpreting D in the *Critique of Practical Reason*. There Kant suggests *how* pleasure can determine an agent to choose to realize an object. It can do so only in the sense that her *expectation* of gaining pleasure from the object's realization determines her to choose to realize it (KpV 22). In light of this suggestion, it makes sense to think of an agent's choice to realize an object being "preceded by pleasure as cause" when the agent makes her choice because she expects to gain pleasure from the object's realization. For example, if someone's choice to master a piano sonata is preceded by pleasure as cause, then she chooses to master it because she expects pleasure from mastering it.

On this interpretation, Kant's *Metaphysics of Morals* account of inclination coheres well with his *Groundwork* account. Like its predecessor, it invokes the notion of an interest: "If a pleasure necessarily precedes the determination of the capacity of desire, the practical pleasure must be called an interest of inclination." Even when we act from inclination, we act on a "general rule" (e.g., a maxim). Inclinations do not bring about our action alone, but when incorporated into practical rules. Moreover, like Kant's *Groundwork* account, his *Metaphysics of Morals* account is amenable to two readings. Kant speaks of the determination of an agent's capacity of desire being preceded by "pleasure as cause." On our interpretation, he is indicating that for an agent to act from inclination is for her to do something because she expects that it will enable her to gain pleasure. His account permits both a reading on which her expectation of pleasure is her only motive and a reading on which it is a motive but not necessarily her only one.

1.8 Material Practical Principles: Acting from Inclination in the *Critique of Practical Reason*

No examination of Kant's account of nonmoral action would be complete without taking stock of his remarks on the topic in the Analytic of the *Critique of Practical Reason*. These remarks support a rejection of Reath's interpretation. In my view, they also permit both the traditional and the alternative

readings, though, as we will see, it is reasonable to contend that they fit better with the traditional reading.

Before analyzing Kant's account in the second *Critique*, it will be helpful to review some of the terminological background against which it takes shape. First, Kant says that practical principles are "propositions that contain a general determination of the will" (KpV 19). This remark is somewhat obscure. But I take Kant to be suggesting that practical principles "contain" a "determination" of the will in the sense that they are rules that some agent(s) have sufficient motive to act on. Second, a *material* practical principle is a rule such that an agent's having sufficient motive to act on it is conditional on her view that doing so will enable her to realize some object she desires (KpV 21). Take the rule: "During your free time, you ought to exercise." To say that it is a material practical principle is to say that an agent's having sufficient motive to act on it (i.e., to exercise,) is contingent on her belief that doing so will enable her to realize some object she desires (e.g., her staying in shape). Third, for Kant if an agent acts on a material practical principle, then she is not acting from duty.[40] Therefore, it seems, she must be acting from inclination: to act on a material practical principle is to act from inclination. As this book unfolds, we will have many occasions to refer to Kant's concept of a material practical principle.

With these points in mind, we can see that Kant's remarks in the Analytic of the second *Critique* clash with Reath's reading of Kant. For example, under Theorem II of "On the Principles of Pure Practical Reason," Kant states that all material principles "place the determining ground of the will in the pleasure or displeasure to be felt in the reality of some object" (KpV 22). As we just noted, if a rule is a practical principle, then someone has sufficient motive to act on it, and her having sufficient motive to act on it is conditional on her believing that acting on it will enable her to realize some object she desires. But, as Kant's statement suggests, this is not the end of the story. The agent's having sufficient motive to act on the rule is also conditional on her expectation that realizing the object she desires will enable her to gain pleasure or avoid displeasure. Therefore, whenever an agent acts on a material practical principle – that is, follows the principle's prescription for trying to realize an object – she has hedonic motivation. Or, what amounts to the same thing: whenever an agent acts from inclination (i.e., nonmorally), she has hedonic motivation.

In our discussion, the most serious question posed by Kant's remarks in the second *Critique* is not whether he held all nonmoral actions to have hedonic motivation but whether he held them ultimately to have hedonic motivation alone. The texts permit this reading but, in my view, do not require it. Take Kant's claim that all material principles place the determining ground of the will in the pleasure or displeasure to be received from an object (KpV 22). We might read him to be saying that all such principles place the one and only motive for willing in the agent's expectation of pleasure.

We are not, however, compelled to read him in this way. After all, Kant does not say that all material practical principles place the determining ground of the will *exclusively* in expected pleasure. What he does say permits a reading according to which he acknowledges that a particular material principle places the determining ground of the will not only in expected pleasure but in something else as well – for example, simply a desire to realize the object.

Here one might point out that, according to Kant, all material practical principles belong under the general principle of self-love or one's own happiness (KpV 22). Since Kant defines happiness as the uninterrupted experience of pleasure (KpV 22), does it not follow that on his view all material practical principles place the motive of action solely in the expectation of pleasure?[41] I think it is plausible to answer this question negatively. It is unclear what Kant means by "the principle of happiness," as well as by the claim that all material practical principles *belong under* this principle. But let us assume, as it seems reasonable to do, that the principle of happiness is a principle prescribing that in order to attain maximum experience of pleasure, an agent ought to perform those actions he believes will enable him to do so. We may then plausibly interpret Kant to hold that all material principles belong under the principle of happiness in the sense that an agent's acting on a material principle always has a feature in common with her acting on the principle of happiness: in both cases, she has the prospect of her own pleasure as a motive. In the former case, she has her own pleasure as one, though not necessarily her only, motive; in the latter, her sole motive is presumably her own pleasure.

Admittedly, with respect to Kant's discussion of "Theorem II," the traditional interpretation seems to have more textual plausibility than the alternative interpretation. In particular, it is more natural to interpret along the lines of the traditional interpretation Kant's statement that all material principles place the determining ground of the will in the pleasure or displeasure to be received from an object. Consider a similar English usage. A critic says: "In speaking thus, the author assumes that the motive for anyone's getting married lies in the desire for companionship." In my view, the critic might be describing the author as someone who takes the desire for companionship always to be one motive, but not necessarily the exclusive motive, for getting married. The critic's statement is consistent with this interpretation. But it is, I acknowledge, more natural to hold that the critic is describing the author as someone who takes the desire for companionship to be the only (real) motive for getting married.[42]

I have argued that Kant's texts permit either a radically hedonistic or a moderately hedonistic account of all actions that are not done from duty. It is more charitable to attribute the latter account to him, since, unlike the former, it does not have the highly questionable implication that, ultimately, all actions not done from duty are done solely from hedonic motives. But, as

we just noted, some of Kant's claims in the *Critique of Practical Reason* fit more naturally with the notion that he embraced a radically hedonistic account. I leave it to the reader to decide which of these two accounts to attribute to Kant (though I favor attributing to him the moderately hedonistic one). For our purposes, the important point to emerge from this discussion is the following. Regardless of whether he upholds a radically hedonistic or moderately hedonistic account of them, Kant maintains that all actions not done from duty – that is, all actions done on material practical principles – are hedonically conditioned. That means (contrary to Reath) that an agent's having sufficient motive to perform each of these actions is conditional on his expectation that doing so will have some hedonic payoff for himself. We might not agree with this position, but we need to recognize that it is Kant's.

2

Transcendental Freedom and the Derivation of the Formula of Universal Law

2.1 Derivation in the *Critique of Practical Reason*: Allison's Reconstruction

Having settled on interpretations of some key concepts in Kant's theory of agency, we are ready to focus on the main topic of this book: Kant's derivations of the Formula of Universal Law and the Formula of Humanity. On the traditional reading, Kant's *Groundwork* derivation of the Formula of Universal Law contains a conspicuous gap. Before I construct and defend a new reading of this derivation, one according to which it is much more forceful and philosophically rich than the traditional construal implies, I first consider how some contemporary philosophers have responded to the traditional reading.

Kant offers a derivation of the Formula of Universal Law not only in the *Groundwork*, but in the *Critique of Practical Reason* as well. At the end of section 8 in the Analytic of Pure Practical Reason, Kant concludes that the Formula of Universal Law "is the *sole* principle that can *possibly* be fit . . . for the principle of morality, whether in appraisals or in application to the human will in determining it" (KpV 41). Kant clearly devotes parts of sections 1–8 to defending this conclusion.[1] Therefore, in light of the apparent gap in Kant's *Groundwork* derivation of the Formula of Universal Law, it makes sense to look to the second *Critique* for a means to fill it.

According to Henry Allison, Kant's second *Critique* derivation relies explicitly on the premise that we have transcendental freedom. If we accept this premise, we can close the gap that has traditionally been found between the claim that the supreme principle of morality would have to require conformity to universal law and the claim that the supreme principle could only be Kant's Formula of Universal Law: "The problematic notion of transcendental freedom must be presupposed, if Kant is to arrive at a contentful, action guiding moral principle; but given such freedom, the derivation succeeds."[2] Allison defends this striking claim in a sophisticated fashion.

However, I argue that, in the end, embracing the notion of transcendental freedom helps us not at all to rescue the derivation of the Formula of Universal Law.

This chapter is not intended as a commentary to sections 1–8 of the Analytic of Pure Practical Reason but rather as a critical discussion of a particular derivation Allison finds there. Therefore, I discuss passages in Kant's text only when they are relevant to Allison's reconstruction. Later (7.2), I discuss (an aspect of) a different derivation argument that, I believe, Kant offers in the second *Critique*.

In outline, the argument Allison reconstructs contains four main steps. First, assuming that we, Kantian rational agents, take ourselves to be transcendentally free, we must hold that our inclinations and desires in themselves fail to constitute sufficient reasons for acting. Second, as Kantian rational agents, we require some nonsensuously based justification for our acting (or adopting a maxim on which to act). The idea seems to be that, as such agents, our adopting a maxim must have some justification. Since the mere presence of a desire cannot itself provide it, something else must contribute to doing so. Third, this other source of justification must be the maxim's conformity to unconditional, universal law. For us to have sufficient reason for adopting a particular maxim, our adopting it must be justified by its conformity to a universally and unconditionally valid practical principle. Fourth, only the Formula of Universal Law (or its equivalents) could be this principle. Therefore, we must hold the Formula of Universal Law to be binding on us.

Actually Allison's argument would do more than fill the gap in the derivation. Filling the gap would amount to showing that if there is a supreme principle of morality, then it is the Formula of Universal Law. However, suppose we accept the background assumptions of Allison's argument, that we are transcendentally free Kantian rational agents. If successful, his argument would then establish that the Formula of Universal Law is actually binding on us.

The chapter begins by exploring the basis of the argument, which is not merely the assumption of transcendental freedom, but a thick Kantian account of rational agency (section 2.2). It then considers the first two steps of the argument, especially the notion that, as rational beings, we require some nonsensuously based justification of our maxims (2.3). The bulk of the chapter (2.4–5) concerns steps 3 and 4 of the argument, namely the plausibility of Allison's claims that our maxims can be justified only by appeal to some universally and unconditionally valid practical principle, and that only the Formula of Universal Law could stand as such a principle.

2.2 A Thick Account of Kantian Rational Agency

Allison bases his argument in a thick account of Kantian rational agency, three features of which require our attention. The first is what Allison calls

the "justification requirement," which we have already mentioned (section 1.5). On Kant's view, says Allison, rational agents must hold that their maxims are subject to "criteria of reasonableness."[3] At the time they act, the agents must regard their maxims as "in some sense rationally justifiable," whether it be morally or prudentially.[4]

Second, the "central insight" of Kant's theory of rational agency lies in what Allison calls the "Incorporation Thesis," a thesis we mentioned earlier (1.5).[5] In Allison's words, this thesis implies that "inclinations or desires do not of themselves constitute a sufficient reason to act but do so only insofar as they are "taken up" or "incorporated" into a maxim by the agent."[6] So, in effect, rational agents never act on brute desires, but they sometimes do act on self-given rules in which they take account of their desires. To revisit one of Allison's examples, the mere presence of a strong desire to eat an ice cream cone cannot itself give me (a rational agent) sufficient reason to eat one. I can have sufficient reason to eat one only if I commit myself to a rule (a maxim), for example, one permitting me to indulge myself in this way under certain circumstances.[7] Actually, when fully expressed, the Incorporation Thesis says more, namely that *no* incentive, including the moral law, itself constitutes a sufficient reason to act unless it is incorporated into a maxim by the agent.[8]

In the context of the Incorporation Thesis, "reason" is a translation of "determining ground," which, I suggested (1.5), should be understood as motivating reason or, simply, motive. Allison seems to agree with this view.[9] So the Incorporation Thesis says that no incentive can itself constitute a sufficient motive for an agent to act. According to the thesis, a rational agent simply *cannot* act on inclinations alone. Whenever an inclination constitutes a motive of her action, she has incorporated it into some self-given rule.

Returning to the Incorporation Thesis itself, Allison argues that in Kant's account an agent must view his incorporating inclinations into self-given rules as an act of spontaneity on his part.[10] The agent must view his "taking up" a desire into his maxim (or, presumably, his refraining from doing so) as an act done independently of the causality of nature, that is, as an exercise of transcendental freedom.[11]

Transcendental freedom is "independence from everything empirical and so from nature generally" (KpV 97). According to Allison, a transcendentally free act would be one that was not causally necessitated by preceding events in time. However, he argues that there is more to transcendental freedom than that. If this were all there were to it, then an agent would have such freedom simply by virtue of possessing a capacity to make causally unnecessitated choices of means to ends that were themselves foisted upon her by nature. For example, suppose that the end of experiencing pleasure was one she, by nature, always necessarily pursued. She would count as transcendentally free if she could, without being causally necessitated to do so, choose between different means of pursuing pleasure.[12] To block such possibilities, in which the scope of transcendental freedom is restricted to

acts of choosing means to ends given by nature, Allison insists that transcendental freedom also involves what he calls "motivational independence." He defines this independence as "a capacity to recognize and be motivated by reasons to act that do not stem, even indirectly, from the agent's sensuous nature."[13] For Allison, an agent sees herself as transcendentally free only if she takes herself to be able to act for reasons that do not stem from her inclinations or desires at all.

In sum, at the basis of Allison's argument lies a threefold conception of rational agency. A rational agent must (1) view the maxims on which he acts as justified, (2) hold that he can act on an incentive only if he incorporates it into a maxim, and (since he holds 2) (3) regard his act of incorporation as an exercise of transcendental freedom.

Of course, one might question each of these claims. For example, one might find problematic the notion that accepting the second requires accepting the third. Suppose I agree that it belongs to the core of my rational agency that I cannot act on my inclinations directly but only through maxims in which I take account of them. Why must I look upon my act of incorporating an inclination into a maxim as an exercise of transcendental freedom instead of an event necessitated by natural causes?[14] At bottom, Allison responds that, if I looked upon my act of incorporation as an event necessitated by natural causes or always motivated by sensuously based reasons, then I would not be considering myself as a rational agent. In other words, there is a conceptual connection between seeing oneself as exercising the capacity to take up desires into self-given rules and seeing oneself as transcendentally free.[15] Allison himself acknowledges that this response is not likely to satisfy his naturalizing critics. In any case, I do not explore this issue here. My aim is to determine whether the argument succeeds in light of the assumption of transcendental freedom and a robust Kantian view of rational agency.

2.3 Desire and Justification of Action

Let us now consider Allison's argument. According to step 1, assuming that we (i.e., rational agents) view ourselves as transcendentally free, we must hold that our inclinations and desires in themselves fail to constitute sufficient reasons (in the sense of motives) for acting. Since we are assuming that the Incorporation Thesis is true, this step is unproblematic. For, according to this thesis, it is essential to being a rational agent that one's inclinations or desires do not of themselves constitute a sufficient motive to act. A rational agent can act on her desires only when she has incorporated them into some self-given rule.

As we have noted, Allison attributes to Kant a "justification requirement" on an agent's adopting a maxim on which to act (and thus on his acting). This is the requirement that the agent regard his maxim as, in some sense,

rationally justified. Step 2 of the argument applies this requirement to desire-based action. According to this step, given that we take ourselves to be transcendentally free rational agents, we must acknowledge that we require some nonsensuously based *justification* for our adopting a maxim on which to act (and thus for our acting). "[D]esire-based action requires a desire-independent warrant."[16] In other words, an appeal to desires alone does not constitute a *sufficient* justification for our maxims.

Step 2 does not follow trivially from step 1. Suppose an agent adopts a maxim in which she takes account of a strong inclination she has. Step 1 does not itself rule out the legitimacy of her justifying her adopting this maxim simply by appealing to the fact that she has this strong inclination. To return to the earlier example, an agent might have a strong inclination to eat ice cream. Granted, as step 1 indicates, her having this desire could constitute a *motivating* reason for her to act only if she committed herself to some rule allowing herself to indulge an inclination for ice cream under certain circumstances. However, the question remains as to whether she could *justify* her committing herself to such a rule simply by appealing to the notion that she has a strong desire for ice cream. In other words, step 1 leaves it an open question whether an agent would be able to justify her act of incorporation simply on a sensuous basis.

Allison himself seems to recognize this, for he offers an argument that is supposed to lead us from step 1 to step 2:

> [T]his conception of transcendental freedom has important implications for the justification of maxims. This is because, assuming motivational independence, the ground of the selection of a maxim can never be located in an impulse, instinct or anything "natural"; rather, it must always be sought in a higher-order maxim and, therefore, in an act of freedom. Consequently, even if one assumes the existence of a natural drive such as self-preservation, a transcendentally free agent is capable of selecting maxims that run directly counter to its dictates. And from this it follows that the mere presence of a drive or inclination does not of itself constitute a sufficient or justifying reason for acting on the basis of it.[17]

In the second sentence, Allison claims that, if an agent has motivational independence, then his justification of his choice of maxim can never be located in any impulse or inclination. It may be unclear, however, why one should accept this claim. Allison defines motivational independence as a capacity to recognize and be motivated by reasons to act that do not stem, even indirectly, from the agent's sensuous nature. Why does an agent's having *this* capacity entail that, rationally speaking, he cannot *justify* his choice of maxim simply with an appeal to sensuously based reasons?

Allison suggests a response in the latter half of the quoted passage: "[E]ven if one assumes the existence of a natural drive such as self-preservation, a transcendentally free agent is capable of selecting maxims that run directly counter to its dictates. And from this it follows that the mere presence

of a drive or inclination does not of itself constitute a sufficient or justifying reason for acting on the basis of it." Yet I find his reasoning puzzling. First, Allison's argument threatens to prove too much. Allison seems to reason that it is because we can act counter to our drives that these drives alone are not justifying reasons for our actions. But we can also act counter to the moral law. By the same reasoning, we should, it seems, conclude that the moral law is not a justifying reason for our action. Second, it is not obvious that a (transcendentally free) agent's being capable of choosing a maxim that goes directly against an inclination precludes him from justifying his action by an appeal to his having this inclination. Suppose someone is in excruciating physical pain and has an inclination for it to stop. Granted, as a transcendentally free agent, he is capable of choosing and acting on a maxim of enduring all pain as long as it lasts, a maxim that would preclude him from acting on his inclination by, let's say, taking morphine. But why does the agent's being capable of acting on such a maxim prevent him from acting on a different one, that is, a maxim of trying to relieve excruciating pain when it occurs, and, moreover, *justifying* his choice of this other maxim simply by appealing to the fact that he wants this excruciating pain to stop?

Even if the particular argument Allison suggests is problematic, there are familiar Kantian grounds for embracing step 2 of Allison's argument. Following Thomas Nagel, for example, we might argue that in undertaking to justify an action, an agent must remove herself from the purely internal perspective and consider her action from an "external" point of view.[18] From this point of view, her action is the action of "this person." To justify the action, she would have to justify the action of anyone in the same circumstances. Yet to justify what anyone would do in the same circumstances, an appeal to the fact that the agent had a certain desire would not be enough. She would need to appeal to some principle, for example, the principle that it is morally permissible for anyone to act on a certain kind of desire in certain circumstances. For example, suppose I do have an inclination to be rid of my excruciating pain, and I believe I can rid myself of it by taking morphine. Asking myself whether I am justified in acting on my inclination for my pain to stop requires me to distance myself from my particular inclination. It makes me see my inclination as that of a person in a certain situation. To justify my taking the morphine, I would have to justify any person's doing so in the same circumstances. Yet appealing to the fact that I now have this inclination to get out of pain would not accomplish this. To justify anyone's taking the morphine in the same circumstances, I would have to appeal to some principle, one such as: it is morally permissible for anyone in excruciating pain to take measures to stop it, if these measures do not cause anyone else comparable pain.

Of course, this argument, especially when sketched so broadly, does not leave an opponent of step 2 without recourse. He might, for example, demand an explanation of why, precisely, he must agree that to justify

his action, an agent must always take an "external" perspective toward it. However, pursuing this discussion would lead us far afield, and I hope I may be excused for not doing so here. I propose that we simply accept step 2 of Allison's argument. An appeal to desires alone cannot justify an agent's maxims. Principles are required not only for the motivation, but for the justification, of our acting.

2.4 Practical Law and Justification of Action

According to step 3, the justification of our maxims must lie in their conformity to some "universally and unconditionally valid practical principle" – some practical law.[19] Only our maxims' conformity to a practical law could justify them. Acknowledging that Kant does not explicitly defend this premise, Allison constructs a Kantian argument for it.[20]

The argument unfolds in two stages. In the first, Allison claims that a sufficient condition for a maxim's being justified is that it be adopted because it is required by some practical law. As Allison suggests, this claim seems relatively unproblematic: "[I]f my reason for *x*-ing is that it is dictated by such a law . . . then I have all justification I could conceivably need for *x*-ing."[21] In the second stage of the argument, Allison defends the view that a necessary condition for a maxim's being justified is that the maxim conform to some practical law. This far more controversial stage of the argument deserves a careful look.

Allison begins by observing that a transcendentally free rational agent, as we are assuming ourselves to be, cannot take his maxims to be justified unless he holds them to be permissible – not contramanded by whatever principle serves as their standard of justification. Allison then tries to show that such an agent must hold the following: a necessary condition of his maxims being permissible is that they conform to some practical law. In other words, only one kind of standard for the permissibility of maxims will do: a universally and unconditionally valid practical principle.

This latter move requires scrutiny. In defense of it, Allison points out that a standard by which we could determine the permissibility of *any* of our maxims would have to be "governing the pursuit of any end at all, including desire- or interest-based ends."[22] If a standard of permissibility applied only to certain ends, then the standard would not apply to some (possible) maxims. For every maxim contains, if only implicitly, a description of an end.[23] So, for example, a standard such as "If you want to maximize your life-span, you ought to do x, y, and z" would not do. For it would not be a basis on which to determine the permissibility of a maxim that does not (even implicitly) have maximizing one's life-span as its end. And we, as Kantian rational agents, can, of course, have such a maxim. Next, Allison argues that, since an adequate principle for determining the permissibility of our maxims would have to be one governing the pursuit of any end at all,

"it must not only apply to all transcendentally free rational agents, it must also apply to them regardless of what desires or interests they may happen to have."[24] And, concludes Allison, such a principle is precisely what Kant means by a practical law.

This argument is, I believe, unsatisfactory. As Allison suggests, Kant does conceive of a practical law as a principle that applies to all transcendentally free rational agents. However, Allison's argument falls short of showing that the standard for the permissibility of our maxims – that is, the maxims of *human* rational agents – must apply to all transcendentally free rational agents. Therefore, the argument does not establish that this standard must be a practical law.

An example helps to illustrate this point. Consider the following principle of happiness, PH: "Always do what you believe will maximize your own pleasure."[25] Let us suppose that PH is a categorical imperative in the following sense: it is a principle that commands us to act in a certain way regardless of what we do or might desire. PH prescribes that an agent do what he believes will maximize his pleasure even if he does not want to maximize it. Strictly speaking, PH is not a viable candidate for a practical law. There might be rational agents who, because of their constitution, cannot have pleasure. Speaking of the pleasure (or pain) experienced by these agents would be akin to speaking of the pleasure (or pain) experienced by the number 3: non-sense. PH could not be a practical law, for it could not serve as the principle by which to determine the permissibility of these agents' maxims.

Allison's argument, however, does not close the possibility that PH could serve as the principle by which to determine the permissibility of *our* maxims. As Allison points out, this principle would have to govern the pursuit of any end we could have, including desire- or interest-based ends. But PH could do so. According to PH, if an agent believes that adopting a particular maxim would enable him to maximize his pleasure, then his adopting it will be permissible; if he does not believe this, then his adopting it will be impermissible. This standard would be in effect no matter what the ends described in the agent's maxims might be, even if they include *minimizing* his own pleasure. Even though PH is not a practical law, it could serve as the principle by which the permissibility of our maxims is determined.

Allison might respond that we find in step 2 of his argument the key to seeing why PH could not serve as this principle. In step 2 he has established that we require a nonsensuously based justification of our maxims. For the same reasons we require this, we also require a nonsensuously based standard of our maxims' permissibility. Since PH is sensuously based, concludes the response, it could not be such a standard. Once again, "desire-based action requires a desire-independent warrant."[26]

This response is ineffective. What step 2 actually establishes is that an agent cannot justify his maxims simply by appealing to the notion that he has

some particular desire or impulse. To justify them, he must appeal to some principle. Likewise, let us grant, the standard of whether an agent's maxims themselves are permissible cannot simply be whether they further some particular desire or impulse he has. To determine his maxims' permissibility, he must appeal to some principle. The important point here is that PH is not some desire or impulse (or the mere notion that one has some desire or impulse). It is a principle. It gives some maxims a "desire-independent warrant" – a warrant no matter what an agent actually desires. PH does not prescribe that we maximize our pleasure on condition that we want to. It is, in the sense already described, a categorical imperative.

Nevertheless, Allison might insist that PH is sensuously based in that it invokes a concept of something sensuous, namely pleasure. I agree; if one wants to use the designation "sensuously based" in this way, then PH is sensuously based. However, Allison has not shown that no principle that invokes the concept of something sensuous could serve as the standard of the permissibility of our (rational human agents') maxims.

Let me put the general point I am making in a different way. Allison is correct in thinking that a principle cannot serve as the standard for the permissibility of each and every one of our (human rational agents') maxims unless it applies to us regardless of which desires and interests we may happen to have. A principle cannot serve as such a standard for our maxims unless it applies to us unconditionally. However, Allison holds (in my view wrongly) that if a principle applies to us unconditionally, it must therefore apply to all rational agents as well. Following Kant, he makes an unwarranted move from the requirement that a principle that serves as the standard for the permissibility of maxims be unconditional (binding on us regardless of what we might desire) to the further requirement that such a principle's scope extend to all rational beings.[27] Since PH fails to fulfill the first requirement, it could not be a practical law. For all Allison has said, however, it still might serve as the principle by reference to which we, transcendentally free human agents, are to determine the permissibility of our maxims.

My point here is *not* to defend PH as a candidate for the supreme principle of morality. I mention it merely as an illustration of a principle that Allison's argument here does not rule out as a possible justificatory basis for our maxims. Just as his defense of step 3 does not exclude PH as a justificatory basis for our (human beings') maxims, it also fails to bar a principle such as "Perfect your human rational and physical capacities," if we understand the principle as an unconditional command. Although this perfectionist principle is not a viable candidate for a practical law – it would obviously not provide a basis for nonhuman rational agents to judge the permissibility of their maxims – it might, nevertheless, serve as a basis for us to judge the permissibility of our maxims. In general, if I am correct, Allison has not established that the scope of whatever principle serves as the basis for the

justification of our maxims must extend to all rational beings. Therefore, he has not proved step 3 – that this principle needs to be a practical law.

2.5 Practical Law and the Formula of Universal Law

In step 3, Allison tries to establish that our maxims are justified only when we adopt them on the basis of their conformity to some practical law. According to step 4, this practical law must be the "moral law," specifically Kant's Formula of Universal Law. To establish this, Allison must obviously show that no principle besides the Formula of Universal Law is capable of justifying our maxims. Allison's argument for step 4 has two main parts, which we need to examine in turn.

The first part is relatively uncontroversial. For a practical law to serve an agent as a justification of her adopting a particular maxim, he suggests, it does not suffice that the maxim conform to the law.[28] Suppose, for a moment, that the Formula of Universal Law is the only practical law. An egoist might act on a maxim of keeping her promises, yet do so in order to secure a good reputation and, ultimately, simply to promote her own happiness.[29] She might contend that her maxim was justified simply by virtue of its conforming to the Formula of Universal Law. But this contention would not withstand scrutiny. For a maxim to be justified by reference to a practical law, it must noncontingently conform to the law. Yet, Allison seems to suggest, the only way a maxim can noncontingently conform to a practical law is if it is adopted at least in part because it conforms to the law. Although the egoist's maxim conforms to a practical law, it does so merely as a result of a coincidence of the law's dictates and her ultimate end. If in different circumstances the egoist believes that the best way to promote her happiness is to break her promises, then she will do that. According to Allison, a justified maxim is one that an agent has adopted ultimately at least in part because it conforms to a practical law.

But when does a maxim count as having been adopted ultimately at least in part because it conforms to a practical law? Allison, it seems, recognizes two main possibilities. First, if I treat my maxim's conformity to a practical law as a sufficient reason for my adopting it, then I have adopted the maxim at least in part because it conforms to a practical law. Second, if I treat my maxim's conformity to a practical law as "an ineliminable component in a jointly sufficient reason," then I have adopted it in part because it conforms to a practical law.[30] This second kind of scenario arises when my maxim is based on inclination. Allison appears to be suggesting the following kind of case. Suppose I adopt a maxim of exercising regularly and that if my maxim violated the law, then, for that reason, I would refrain from adopting it. In this case, I count as adopting my maxim ultimately (at least in part) because of its conformity to a practical law. I count as doing so even if I don't take the exercising maxim's conformity to practical law as in itself sufficient for my

adopting it, that is, even if I would not adopt it unless I had an inclination to stay in shape.

Now the question arises: why must a justified maxim be adopted by virtue of its conformity to the Formula of Universal Law, instead of some other principle? If Allison answers this question, it will be in the second part of his argument. He moves very quickly:

> [I]f I am required to adopt a maxim at least in part because of its conformity to universal law or, equivalently, an unconditional practical law, then, clearly, this maxim must be able to include itself as a "principle establishing universal law," which is just to say that the maxim must have what Kant terms "legislative form." In other words, the intent expressed in the maxim must be compatible with the putative universal law produced by the generalization of that maxim. Otherwise, its conformity to such a law could not possibly be the reason (or even a reason) for adopting the maxim in the first place.[31]

Allison claims that only if an agent's maxim has legislative form can she adopt it even in part because of its conformity to practical law. On Allison's reading, it seems, a maxim has legislative form just in case an agent can act on it and will that it become a universal law. In other words, a maxim has legislative form just in case it passes the test contained in Kant's Formula of Universal Law. A rational agent who assumes she is transcendentally free, says Allison, will find on reflection that only if her maxim passes the Formula of Universal Law test can she adopt it even in part because of its conformity to practical law. A necessary condition of her maxim's being justified is thus that the maxim conform to the Formula of Universal Law. Therefore, Allison seems to hold, the agent must conclude that the law on the basis of which she adopted any justified maxim would have to be this formula.

But why should we agree that only if an agent's maxim passes the Formula of Universal Law test can she adopt it even in part because of its conformity to practical law? In my view, Allison leaves this question unanswered. Let me illustrate this point with the help of a few examples. First, consider the principle PRU: "Act only on maxims such that you believe that your acting on them will maximize the general happiness." It seems that someone's reason for adopting a maxim could be that it conformed with PRU, even if his maxim failed the Formula of Universal Law test. Take a maxim such as, "In order to maximize the general happiness, I will devote myself entirely to the study of ethics." This maxim fails the Formula of Universal Law test, at least on a common reading. It is not possible for me, as a rational being, to act on it and at the same time will that it become a universal law. In willing the universalization of my maxim, I would be willing a state of affairs in which no one would develop the skills necessary for maximizing the general happiness, for example, skills in food production or medicine. I would, in effect, be willing it to become impossible for me to attain the end described

in my maxim. Despite the maxim's lack of legislative form, it seems that I could adopt it because it conforms to PRU.

In response, perhaps Allison would suggest that at this point in his argument it is already clear that an agent could not justify his adoption of a maxim by appealing to a principle such as PRU. According to his step 3, the justification of a transcendentally free agent's maxims must lie in their conformity to some practical law. But, Allison might argue, PRU is simply not a viable candidate for a practical law. Such a law must apply to all rational agents – something that PRU, like PH, fails to do, since presumably there might be some rational agents (e.g., angels) to whom the concept of happiness fails to apply.

Actually, it is not obvious that PRU fails to apply to all rational agents. Granted, we cannot presuppose that all rational agents can be happy. Perhaps angels cannot be. However, it would be mistaken to conclude simply from this observation that PRU could not apply to all rational agents. That angels cannot themselves be happy does not entail that they are incapable of acting on maxims that, they believe, would maximize the general welfare. They presumably hold that what they do has an impact on agents capable of various degrees of happiness – that is, agents like us. If PRU fails to apply to any rational agents, it fails to apply only to those who both cannot themselves be happy and who can in no way influence the happiness of others. Since it is possible that such isolated agents exist, it seems that, after all, PRU does not necessarily apply to all rational agents, and that, therefore, PRU is not a viable candidate for a practical law. Of course, if, as I have claimed, Allison fails to establish step 3, and we (transcendentally free human agents) could justify our maxims with reference to a principle that does not quite have the scope of a practical law, then this point is moot.

But even if Allison succeeds in showing that a maxim based on PRU could not both fail to have legislative form and be adopted on the basis of a practical law, other challenges are not difficult to generate. I will consider two. First consider the weak principle of universalization WU: "Act only on that maxim which, when generalized, could be a universal law." Notice that WU is not equivalent to the Formula of Universal Law. As Kant suggests in the *Groundwork*, a maxim of nonbeneficence could, when generalized, constitute a universal law (GMS 423). Since a world where no one acted beneficently is indeed a coherent possibility, acting on a maxim of nonbeneficence does not violate WU. On Kant's view, of course, acting on such a maxim runs afoul of the Formula of Universal Law. It does so, he thinks, because as a rational agent it is not possible to act on it and, at the same time, will that its generalization be a universal law. A maxim of nonbeneficence does not have "legislative form." However, Allison does not explain why an agent could not adopt a maxim of nonbeneficence on the basis of its conformity with WU. He does nothing to rule out the possibility that WU is the practical law on the basis of which we can justify our maxims. There remains a gap between

the notion that our maxims must be justified on the basis of some practical law and the notion that this law must be the Formula of Universal Law.

A rather bizarre principle can serve as a second illustration of this point, namely BP: "Act only on that maxim that you *cannot*, at the same time, will that it become a universal law." It is, of course, not hard to find maxims that would be in accordance with BP, yet that would not have legislative form. Once again, the maxim of nonbeneficence Kant discusses in *Groundwork* I fits the bill. If the maxim of nonbeneficence fails the Formula of Universal Law test (as Kant holds it does), then it passes the BP test. According to Allison, an agent's reason for adopting a maxim could not be its conformity to BP. No one, Allison claims, could adopt a maxim of nonbeneficence and have as his reason for doing so the maxim's conformity with BP. Yet why couldn't a transcendentally free agent's reason for adopting a maxim be that it conform to BP, rather than to the Formula of Universal Law? BP, we might specify, has the form of a practical law. It is an unconditional imperative, commanding that we act in a certain way regardless of what we desire to do. Moreover, like WU, it is not sensuously based, not even in the minimal sense of making use of sensuous concepts such as happiness or pleasure.

But surely we need not take BP seriously, one might here interject. For running maxims through the BP test has obviously counterintuitive implications. Just to name one, if we rely on BP as a standard, we find that we have no duty of beneficence, which goes strongly against ordinary moral thinking. I agree wholeheartedly that BP has counterintuitive implications (as does WU) and that, therefore, we need not take it seriously as a candidate for the supreme principle of morality. My point is simply that Allison's argument does not exclude this principle (or WU) as a basis on which transcendentally free rational agents justify their maxims. And since it does not, the argument is not a successful derivation.

One might see it as an advantage of Allison's approach that it does not rely on Kant's appeals to ordinary moral reasoning in *Groundwork* I, for example, on his discussion of which actions we take to have moral worth. These appeals generate considerable controversy – controversy that it might seem desirable to avoid. But shortcomings in Allison's argument suggest that an examination of ordinary moral reasoning must play a role in any effective derivation of the Formula of Universal Law. To put it bluntly, Allison's argument remains at too abstract a level to be successful. Even if we begin with a very robust notion of a Kantian agent and, contrary to my objection, grant him that such an agent (even if he is also human) must justify his maxims by an appeal to a practical law, we find that Allison fails to show that this practical law must be the Formula of Universal Law. There is a gap in this derivation: one perhaps just as wide as that in Kant's *Groundwork* I derivation (traditionally interpreted).

3

The Derivation of the Formula of Humanity

3.1 Outline of the Derivation

On the received view, Kant's derivation of the Formula of Universal Law fails, and, if I am correct, Allison's attempt to rescue it also falls short. But the Formula of Universal Law is not the only principle Kant defends. He also advocates the Formula of Humanity: "So act that you treat humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" (GMS 429, emphasis omitted). Perhaps the *Groundwork* derivation of this principle is a success. This chapter explores whether Kant shows that if there is a supreme principle of morality, then it is the Formula of Humanity.

From the outset, we should keep in mind that Kant employs "humanity" in a somewhat technical sense. The term does not refer to the class of human beings but rather to a set of capacities. In the *Metaphysics of Morals*, Kant tells us that "the capacity to set oneself an end – any end whatsoever – is what characterizes humanity (as distinguished from animality)" (MS 392). So at the very least, to have humanity involves having the capacity to set ends. Kant, it seems, uses "humanity" interchangeably with "rational nature" (e.g., GMS 439). In doing so, he suggests that to have humanity is to have certain *rational* capacities. Indeed, for Kant the capacity to set ends is a rational capacity. An agent sets herself an end through adopting a rule that specifies the end, as well as means to take to it in certain circumstances; and adopting such a rule is an exercise of reason (see MS 395). Unfortunately, Kant does not offer a list of precisely which rational capacities belong to humanity (rational nature). But we will, I believe, not go astray if we take the central ones to be the capacity to set oneself ends and the capacity to adopt and act on rules, including rules of prudence (hypothetical imperatives) and rules of morality (categorical imperatives).[1] This set of capacities is neither one that is (necessarily) possessed only by human beings, nor is it one that is possessed by all human beings. Nonhuman beings (e.g., extraterrestrials)

could have humanity and some human beings (e.g., ones in a persistent vegetative state) presumably lack it.

Let me now offer a brief sketch of Kant's derivation of the Formula of Humanity (GMS 428–429). On Kant's basic concept, the supreme principle of morality would have to be a categorical imperative, that is, a principle binding on us no matter what our particular inclinations might be. The derivation takes shape against the background of this fundamental tenet. First, Kant contends that if there is a supreme principle of morality (and thus a categorical imperative), then there is an objective end, something that is unconditionally good. Second, Kant claims that this unconditionally good thing would have to be humanity. In his view, therefore, if there is a supreme principle of morality, then humanity is unconditionally good. But, third, if humanity is unconditionally good, then we must always treat it not merely as a means but also as an end. Therefore, if there is a supreme principle of morality, then we ought to do just what the Formula of Humanity says. So the supreme principle of morality, if there is one, must be this formula, or at least something equivalent to it.

This chapter focuses on the derivation's first two steps. The chapter begins (section 3.2) by examining (and criticizing) Kant's attempt to show that assuming there to be a categorical imperative requires us to take something to be unconditionally good. Next (3.3) it explores Kant's discussion of the claim that this something must be humanity. Since, I suggest, this discussion is not an adequate basis for embracing the claim, we need to reconstruct Kant's argument for it. The most forceful reconstruction has been offered by Christine Korsgaard, and the bulk of this chapter (3.4–6) concerns it. In my view, Korsgaard's reconstruction does not succeed. I explain why in section 3.7.

Since I argue that the first two steps of the derivation fail, I do not here examine the third. I do not criticize the claim that if humanity is unconditionally good, then we must always treat it not merely as a means but also as an end. Yet I do not wish to suggest that this claim is unproblematic. In 3.8, I mention a possible difficulty concerning it.

As I understand it, Kant's derivation of the Formula of Humanity is not in any obvious way predicated on the success of his derivation of the Formula of Universal Law. However, I argue that the first step of the former derivation falls prey to a serious objection unless the success of the latter is assumed.

3.2 The Supreme Principle of Morality and Unconditional Value

In his initial step in the derivation of the Formula of Humanity, Kant claims that if there is a supreme principle of morality (and thus a categorical imperative), then there is something of absolute worth. In something of absolute worth alone "would lie the ground of a possible categorical imperative" (GMS 428), and "if all worth were conditional and therefore contingent,

then no supreme practical principle for reason could be found anywhere" (GMS 428). But why couldn't a principle be unconditionally binding on us if nothing was unconditionally good? How does Kant tie the notion of unconditional bindingness to that of absolute goodness?

This question is not easy to answer, but this much seems clear. According to Kant, an agent sets himself to do something – that is, he determines his will, on the basis of his idea that doing this thing will enable him to secure some end. In Kant's view, all acting has an end (see, e.g., KpV 34). This should come as no surprise since Kant holds that all acting is acting on a maxim and that, when fully described, a maxim will contain a description of an end (see section 1.2). Kant distinguishes between subjective and objective ends. Objective ends, if there are any, would hold for all rational beings. The idea of securing them would make available to all rational beings a sufficient ground (motive) for acting. But subjective ends do not give all rational beings grounds for securing them. These ends are such that their "mere relation to a specially constituted capacity of desire on the part of the subject gives them their worth" (GMS 428). Suppose a particular object is a subjective end. If an agent does not value this object, either in itself or as a means to something else, then it has no worth to him. And if the object has no worth to him, intimates Kant, then he does not have a ground to secure it. For him, it is not an end. Apparently, Kant has the following view: an agent has a sufficient ground to secure an object only if he values it – or at least is rationally compelled to value it. In the latter case, the agent is presumably able, through rational reflection, to come to value the object, thereby gaining a sufficient ground to secure it.

Against the background of this view, we can reconstruct the basis of Kant's claim that if there is a categorical imperative, then there must be an objective end – something absolutely valuable. A categorical imperative would be necessarily binding on all rational agents. But a principle could not be necessarily binding on all rational agents unless each of them necessarily had a sufficient ground (motive) at his or her disposal for obeying it.[2] Take us, human rational agents. To say that a principle is binding on us is to say that we ought to (i.e., have an obligation to) conform to it. Kant, of course, holds that, if an agent ought to do something, then she must be able to do it (e.g., KrV A 807/B 835; KpV 125, 159). But if an agent did not have a sufficient ground available to her for conforming to a rule, then she might not be able to conform to it. Thus, if not all rational agents necessarily have a ground for obeying a principle, then it cannot be a categorical imperative. For Kant, as we noted, to have a ground for doing something an agent must hold (or be rationally compelled to hold) the action or its effects to be valuable. Therefore, Kant seems to conclude, if there is a categorical imperative, then there must be something that everyone holds (or must hold) to be valuable: an objective end. There must be something that everyone, in every context, is rationally committed to valuing: something that is absolutely valuable or, equivalently, unconditionally good.

A brief discussion of how Kant conceives of unconditional goodness will put us in position to see that this argument is problematic. For something to be unconditionally good, it must (obviously) be good under every possible condition, in every possible context. Moreover, if something is unconditionally good, then it is good from the perspective of an impartial rational spectator, or so Kant makes clear at the beginning of *Groundwork* I. He dismisses the notion that happiness is unconditionally good thus: "[A]n impartial rational spectator can take no delight in seeing the uninterrupted prosperity of a being graced with no feature of a pure and good will, so that a good will seems to constitute the indispensable condition even of worthiness to be happy" (GMS 393). In contemporary terms, we might say that for Kant unconditional value is agent-neutral value. But now the question arises: to hold a principle to be a categorical imperative, must an agent really maintain that there is something unconditionally good in Kant's sense? Suppose that Fred holds the following to be a categorical imperative: PW, that is, "Maximize your power over rational beings." In Fred's view, PW is binding on all agents, no matter what they might desire. Given that Fred holds PW to be a categorical imperative, he must hold that each agent is rationally committed to the view that obeying PW is always good for her (or necessarily enables her to promote something good for her). Fred might conclude the following: he always has available a sufficient ground for obeying PW by virtue of being rationally compelled to take his having maximum power over rational beings to be good for him, whereas another person, *b*, always has available a sufficient ground for obeying PW by virtue of being rationally compelled to take *b*'s having maximum power over rational beings to be good for *b*, while yet another, *c*, always has such a ground by virtue of being thus compelled to take *c*'s having such power to be good for *c*, and so forth. There does not seem to be anything self-contradictory in Fred's conception of PW as a categorical imperative. Yet Fred does not hold his own power (or anything else) to be unconditionally valuable, at least on Kant's conception of unconditional value as agent-neutral value. He has no particular commitment regarding whether an impartial spectator would take his having maximum power to be good; for all he knows, such a spectator would be totally indifferent to this possibility. Although a categorical imperative requires a ground, it does not seem as if this ground must be something unconditionally good.

Perhaps this objection would not worry Kant. At the point in the *Groundwork* at which Kant derives the Formula of Humanity, he has already completed his derivation of the Formula of Universal Law. He has, he believes, proved that, if there is a supreme principle of morality, then it is the Formula of Universal Law (or something equivalent to it). So, from his perspective, any principle conformity to which would require violating the Formula of Universal Law is thereby eliminated as a candidate for the supreme principle of morality. (If conforming to a principle requires violating the Formula of Universal Law, then the principle is obviously not equivalent to the Formula

of Universal Law.) In response to the objection, Kant might say that, even if he granted that the ground of a categorical imperative need not be something unconditionally good, it would be the burden of his opponent to show that a principle that had a different ground could be compatible with the Formula of Universal Law. Perhaps one could successfully assume this burden, but I have yet to discover how. In acting on PW as Fred would do, one obviously might be violating the Formula of Universal Law, for example, by making a false promise. Given the context of the derivation of the Formula of Humanity, it makes sense that Kant would not address the objection we have raised.

However, we need to take the objection seriously. The response that, I have suggested, he might offer turns on the assumption that the derivation of the Formula of Universal Law has been successful. But this is obviously not an assumption that we are in a position to make. Does Kant have the resources to meet the objection without relying on such a robust assumption?

To reply to this objection, one might appeal to some of Kant's remarks in the second *Critique*. Kant distinguishes between well-being (*das Wohl*) and good (*das Gute*), and ill-being and evil. Well-being and ill-being refer to a person's "state of feeling" (KpV 60). A scoundrel who provokes an innocent person and gets a thrashing for it experiences an ill. But, says Kant, "everyone would approve of it and take it as good in itself even if nothing further resulted from it" (KpV 61). He goes on to say: "What we are to call good must be an object of the capacity of desire in the judgment of every rational human being, and evil an object of aversion in the eyes of everyone; hence for this appraisal reason is needed, in addition to sense" (KpV 60–61). For us legitimately to hold an object to be good (as opposed to conducive to our own well-being), it must be something that each rational agent judges, or at least should judge, to be worth bringing about – to be desirable. If this is correct, then Fred's conception of the "ground" of PW is flawed. According to Fred, each agent is rationally compelled to take her own power to be good for her, and thereby always has available to her a sufficient motive to conform to PW. However, there is obviously no reason to assume that each agent does in fact judge *other* agents' having maximum power over rational agents to be desirable. Nor does it seem that each agent *should* judge this to be desirable. After all, another agent's maximizing her power might prevent him from maximizing his own power, that is, from securing what he must (in his view) take to be valuable. From this we can see that Fred cannot, rationally speaking, take his maximum power over rational beings to be good. Since he cannot, he is obviously not rationally compelled always to take it to be good. Therefore, Fred does not necessarily have a motive available to him for conforming to PW. (Although Fred might have an inclination to maximize his power now, there is no guarantee that he will want to do so at other times.) Despite initial appearances, Fred cannot really hold PW to be a categorical imperative.

Another passage in the second *Critique* suggests a related response to the objection. According to Kant, "a law, as objective, must contain the *very same determining ground* of the will in all cases and for all rational beings" (KpV 25). Kant appeals to this notion in order to show that a principle of happiness cannot be a practical law. But he might just as well appeal to it to show that each agent's notion that his own power is necessarily good could not serve as a ground for PW to be a practical law. For Kant here suggests that each agent must have *the very same* motivating reason available to him for conforming to a rule, if this rule is to be a practical law. And, on Fred's own conception, not every agent has the very same motivating reason available to him for maximizing his own power. Fred has a ground to maximize his power in that he is rationally compelled to take *his* power to be good; another agent has a ground to maximize her power in that she is rationally compelled to take *her* power to be good, and so forth. These are obviously not the very same grounds. (Fred's ground for maximizing his power does not lie, for example, in his being rationally compelled to take someone else's power to be good.) If grounds such as these are all the ones agents have available for conforming to PW, they do not suffice to ground it as a practical law.

These two replies are convincing – if one accepts their premises. But I do not think we can credit Kant with proving the premises to be true. The first argument rests on the notion that for us legitimately to hold an object to be good (as opposed to conducive to our own well-being), the object must be something that each rational agent judges, or at least should judge, to be desirable. However, consider the stage in Kant's dialectic at which he has not successfully derived any of his candidates for the supreme principle of morality. At this stage, which is after all our stage, I do not find in Kant a demonstration of the notion in question. (Kant himself makes the claim that "what we are to call good must be an object of the capacity of desire in the judgment of every rational human being" *long after* he has completed his second *Critique* derivation of the Formula of Universal Law.)[3] What, precisely, would be irrational in an agent's calling an object good (as opposed to conducive to his well-being), even if, in the agent's view, not everyone was rationally compelled to judge the object desirable? The agent might reasonably believe that the object – for example, her having maximum power over rational beings – would actually diminish her well-being. To put the point in contemporary terms: is it always irrational for an agent to take an object to be good in an agent-relative sense, that is, good from her standpoint (though not in terms of her happiness), but not good in an agent-neutral sense, that is, desirable from an impartial perspective?

On the face of it, this sort of view does not always seem to be irrational. Suppose that you sacrifice your professional ambitions and your ties to your lover for the sake of your child, who has a painful, though not debilitating, disease. You take a high-paying job you despise in a new city. Difficult as it is for you to acknowledge, it turns out that you are even more miserable than

you would have been had you not made the sacrifices and your child remained sick. But your extra income allows you to obtain effective treatment for him. Your child thrives. As it stands, though, if you had donated to relief organizations the money it took to treat your child, a dozen other children would have been saved from intense suffering and death. Now the state of affairs that results from your action is obviously not good in terms of your well-being. Moreover, it might not be good from an impartial standpoint, for example, if the impartial standpoint is a utilitarian one. Nevertheless, it seems that you might, without irrationality, hold the state of affairs to be good *from your standpoint.*

According to Amartya Sen, there would not necessarily be anything irrational in your holding this view. He defends the coherence of "evaluator relativity," according to which, ultimately, the goodness of a particular state of affairs depends intrinsically on the position of the evaluator in relation to the state.[4] For Sen, there need be nothing self-contradictory in saying that the state of affairs we have been discussing is good from the position you hold but not good from an impartial standpoint. Of course, much discussion would be needed to defend the notion of evaluator relativity. But my aim here is modest. I simply want to point out the following. The first Kantian reply relies on the view that for us legitimately to hold an object to be good (as opposed to conducive to our own well-being), the object must be something that each rational agent judges (or ought to judge) to be desirable. However, this view is controversial and, so far as I can tell, not one that Kant bolsters with arguments.

The second reply rests on the premise that each agent must have *the very same* motivating reason available to him for conforming to a rule, if this rule is to be a practical law. But this is not obvious. Let us grant for now that no rule could be unconditionally binding (and thus a candidate for the supreme principle of morality) unless everyone necessarily had a sufficient motive available to him for abiding by it. It does not follow from this that everyone would have to have the *very same* motive for abiding by it.

One might construe Kant to be suggesting an argument for this view in the second *Critique* (KpV 28): suppose that everyone has a motive, but not the very same motive, for conforming to a principle. In this case, conformity to the principle would not produce harmony. Consider what would occur if each agent was motivated by the notion that his own happiness was good to conform to the principle that one ought to maximize one's own happiness. Everyone's conforming to this principle would surely result in disharmony, if only because some agent's promoting his happiness by securing an object would preclude others from promoting theirs by securing this object. For another example, look what would happen if each agent had a motive to conform to PW, and that motive was the notion that maximizing his own power was good. Each agent's conforming to PW would bring about disharmony, a vast competition to gain the upper hand. However, the argument

continues, a law, whether it be practical or natural, must produce harmony. Hence for a practical principle to count as a law, conformity to it must promote harmony. Yet unless everyone has the very same motive for conforming to a principle, conformity to it might not promote harmony. Therefore, to be a practical law, a principle must be such that everyone has the very same motive for conforming to it.

This argument (whether or not it is really Kant's) suffers from several difficulties. First, it is questionable whether natural laws necessarily make everything harmonious. On Kant's view, such laws presumably govern the occurrence of earthquakes, droughts, and floods. In so doing, are they promoting harmony? From the perspective of those struggling to survive these events, the answer would seem to be negative. In response, one might claim that from a god's-eye view, in governing these events natural laws are promoting (or at least maintaining) harmony, specifically the harmony constituted by all events being governed by such laws, instead, say, of being random. But this answer to the first difficulty yields a second one. For it seems that the scenario in which everyone acted on the principle of maximizing one's own happiness (or power) would also produce harmony from a god's-eye view. Although, to an individual battling to advance his happiness (or power) amid others battling to advance theirs, the world might not seem a harmonious place, it would appear as such to someone who stepped back from the fray to consider that each agent was acting on the same practical principle, namely one commanding him always to promote his own happiness (or power), not on whatever principle he happened to stumble upon.

Let us agree that unless everyone has the very same motive for conforming to a practical principle, universal conformity to it might not, from the perspective of those doing the conforming, yield harmony. But why should we accept the notion that to count as a practical law, a principle would have to yield harmony in this sense? Why isn't the kind of harmony that would be manifest from a god's-eye view enough? One might observe that it would be unfortunate for us if universal obedience to a practical law failed to yield harmony from the perspective of those conforming to it. That it would be unfortunate, however, does not entail that it would be impossible.

Finally, let us grant for the sake of argument that natural laws yield a kind of harmony that everyone's obeying the principle of happiness or of power would fail to yield. For Kant natural laws determine what is; practical laws determine what ought to be (GMS 387–388, KpV 19–20). Given that natural laws differ in kind from practical ones, why should we assume that the latter (when universally obeyed) must share the harmony-promoting characteristic of the former?[5] Kant fails to give us grounds for accepting his claim that no practical law could promote discord. He fails to prove the relevance of the question of whether a practical principle, when acted on universally, would lead to disorder to the question of whether such a principle has (or could have) status as a practical law.

As far as I can tell, Kant does not offer a convincing argument for the claim that a practical law (in the sense of a categorical imperative) must be such that everyone has the very same motive for conforming to it.[6] Since he does not, we are not in position to reject a principle such as PW as a possible categorical imperative on the basis that it does not "contain the *very same determining ground* of the will in all cases and for all rational beings."

To sum up this section, Kant claims that if there is a categorical imperative, then there is something unconditionally good. His argument for this claim seems to go as follows. Suppose some principle is a categorical imperative. By definition, this principle would be unconditionally binding on all rational agents. But now suppose that there is nothing unconditionally good. Some rational agents might find themselves with insufficient motive to conform to the principle, and thus might be unable to do so. If some agents were unable to conform to the principle, then it would not be binding on them; for an agent does not have a duty to do something that she cannot do. Contradicting our initial supposition, the principle would not be a categorical imperative. Kant's argument turns on the notion that only if an agent holds there to be something unconditionally good can she maintain that every agent always has available to her a sufficient motive to conform to a given principle. I have suggested that Kant fails to establish this notion. He does not successfully block the possibility that an agent can deny there to be anything unconditionally good (i.e., good in every possible context, according to an impartial rational spectator), yet at the same time coherently maintain that every agent always has at her disposal a sufficient motive to conform to a given principle.

3.3 The Unconditional Value of Humanity: Kant's Argument

In the first step of his argument, Kant tries to show that if there is a categorical imperative, then there must be some object (or objects) that all rational agents must hold to be unconditionally good. In the second step, Kant claims that this something is humanity. If we hold that there is a categorical imperative, then we must conclude that humanity is unconditionally good. Kant packs his defense of this claim into one dense and difficult paragraph.[7]

On its face, the defense seems inadequate. Kant needs to demonstrate that only humanity could be the "ground" of a categorical imperative. What Kant does is dismiss three candidates for unconditional goodness: objects of inclinations; inclinations themselves; and beings "the existence of which rests not on our will but on nature" (GMS 428), for example, animals. Then, without further ado, he announces that humanity could be unconditionally good.

His dismissal of rival candidates for unconditional goodness seems precipitous. For example, regarding inclinations, Kant merely says: "But the inclinations themselves, as sources of needs, are so far from having an absolute

worth, so as to make one wish to have them, that it must instead be the universal wish of every rational being to be altogether free from them" (GMS 428). Most of us, I venture, are not tempted to the view that our desires themselves (as opposed, perhaps, to their objects) have an absolute value. It seems strange, however, to dismiss this view on the grounds that all of us wish to be altogether free from our desires.

Even if the remarks Kant here makes do eliminate these candidates for absolute goodness, the question arises as to whether he is entitled to conclude that it is humanity he is looking for. After all, might not Kant have overlooked some other candidate for absolute goodness? What about the state of affairs of all rational agents being happy? How does Kant dismiss this possibility? This candidate is not itself an inclination. It need not be considered an object of an inclination. (We can easily envisage a world in which no one desires everyone – including his enemies – to be happy.) And everyone's happiness is not obviously something the existence of which would rest on nature rather than on our will.

In sum, Kant's argument that only humanity could be the absolutely valuable thing required if there is to be a categorical imperative (supreme principle of morality) appears to suffer from two shortcomings. First, his arguments against other candidates seem too quick; second, he offers no grounds for the conclusion that he has considered all other candidates. On the face of it, Kant's argument that if we take there to be a categorical imperative, then we must take humanity to be unconditionally good does not seem very promising.

3.4 Korsgaard's Reconstruction: Preliminaries

In suggesting an opposing view, Christine Korsgaard offers an ambitious and influential account of Kant's argument.[8] Since this account is the most forceful one I am familiar with, I consider it in detail.[9] But I do not explore the extent (if any) to which Korsgaard's account departs from the letter of Kant's text in the *Groundwork*. My view, which I do not here try to defend, is that, though the argument Korsgaard presents is available to Kant, it is not in all of its details one that Kant actually gives.[10] It is, I think, a reconstruction of *Groundwork* 428–429.

The key concept in Korsgaard's reconstruction is that of a good end.[11] The reconstruction can be divided into two claims, both of which invoke this concept. First, if we (rational agents) take there to be a categorical imperative and thus something unconditionally good, then we must hold ourselves to have good ends. Second, if we take ourselves to have good ends, then we must hold humanity to be unconditionally good. If these two claims are successful, then, in effect, a big part of Kant's argument goes through. He manages to establish that if we take there to be a categorical imperative, then we must hold that humanity is unconditionally good.

To evaluate the first claim, we need to have in view what Korsgaard means by a good end. She attributes to Kant a robust conception of one. To count as good, an end must meet the following criteria. First, it must be the object of rational choice. By this, Korsgaard apparently means that reason must play a role in the process through which the agent comes to have the end.[12] A good end is not one simply laid down by instinct. It is, rather, one an agent sets himself after some reflection on whether it is worthy of pursuit.[13] Second, a good end must provide "reasons for action that apply to every rational being."[14] This is Korsgaard's gloss on Kant's claim that the good "must be an object of the capacity of desire in the judgment of every rational human being." According to Korsgaard, the claim entails that, for Kant, to be good an end must be one that we can share.[15] Third, a good end must be fully justified.[16] It appears that, for Korsgaard, a fully justified end would be one that was either itself unconditionally good or that derived its goodness from something unconditionally good.

Since the notion of a fully justified end is crucial to the argument, let us view in detail Korsgaard's explication of it:

> An end provides the justification of the means; the means are good if the end is good. If the end is only conditionally good, it in turn must be justified. Justification, like explanation, seems to give rise to an indefinite regress; for any reason offered, we can always ask why. If complete justification of an end is to be possible, something must bring this regress to a stop; there must be something about which it is impossible or unnecessary to ask why. This will be something unconditionally good. Since what is unconditionally good will serve as the condition of the value of other good things, it will be the source of value.[17]

Justifying a claim that an end is good involves answering the question of why it is good. Suppose someone asserts that his jogging is good. To justify this assertion, the person would need to explain why it is good, perhaps by pointing out that it keeps him in shape. Yet then the question arises: why is it good for him to be in shape? Perhaps he answers that he feels better when he is fit, an assertion that might, in turn, give rise to a question of why his feeling better is good, and so forth. In Korsgaard's view, a full justification of the goodness of an end would bring this sort of regress to a close. It would show us that the goodness of the end depended on the goodness of something regarding which it would be either impossible or unnecessary to pose the question: why is this good? In Korsgaard's view, this special something will be unconditionally good.

In sum, for an agent to count his end as good, the end must be the object of his rational choice; provide reasons for action to every rational agent; and be fully justified.

3.5 The Supreme Principle of Morality and Good Ends

If an agent holds there to be a categorical imperative, must he hold that he has at least one end that meets all three criteria? Korsgaard herself explains

what she (and presumably Kant) mean by a good end, and she suggests that the notion that we have good ends serves as an assumption in Kant's derivation of the Formula of Humanity.[18] But she does not focus on the question of whether taking there to be a categorical imperative *rationally compels* one to believe that he has some good end(s): at least one end that meets all of the criteria.

So let us consider this question. Clearly, an agent must have at least one end that fulfills the first criterion. In Kant's view, being a rational agent involves setting ends for oneself. And any end an agent sets for himself would be an object of his rational choice. For the power of rational choice (humanity) is the power to set ends (MS 392). Any rational agent, let alone any agent who holds there to be a categorical imperative, would have to hold that he has ends that are objects of his rational choice.

Next, does assuming there to be a categorical imperative rationally compel an agent to hold that he has an end that everyone has a reason to promote? If holding there to be such an imperative required each of us to maintain that some particular object was unconditionally good, then the answer would be yes. For in this case, all of us would be rationally compelled to promote (or perhaps preserve) this unconditionally good thing. Suppose this unconditionally good thing were everyone's happiness. I would have reason to further your end of promoting the general welfare, and you would have reason to further my end of promoting the general welfare and so forth. However, I have argued that maintaining there to be a categorical imperative is consistent with denying that there is anything unconditionally good. By virtue of holding PW to be a categorical imperative, Fred does not rationally commit himself to the view that everyone has a reason to promote his (Fred's) having maximum power over rational beings. But Fred is committed to the view that everyone has a reason to promote his own power over such beings. An agent's assuming that there is an unconditionally and universally binding practical principle does not entail that he must affirm that there is something unconditionally good that everyone has reason to promote.

What about the third criterion: must an agent who holds there to be a categorical imperative also hold that he has some fully justified end? A fully justified end would be one that was either itself unconditionally good or that derived its goodness from something unconditionally good. Since an agent who maintains there to be a categorical imperative does not have to profess there to be anything unconditionally good at all, I see no reason why she would need to hold that she had any fully justified end (in Korsgaard's sense).

In short, an agent's assuming that there is a categorical imperative does not require her to agree that she has any ends that meet the second and third criteria Korsgaard sets out. It seems that an agent can at the same time hold there to be a categorical imperative yet deny that she has any good ends.

However, that does not preclude the possibility that there are other grounds for holding that one has good ends. In *The Sources of Normativity*, Korsgaard suggests that unless one holds that one does, one is committed to complete practical skepticism – that is, to the view that one has no reason to do anything at all.[19]

Yet I do not see why one would be committed to this. Suppose that someone takes himself to have good ends, though not in Korsgaard's sense according to which such ends must be "fully justified." He holds that the goodness of his good ends derives from his reflectively, as opposed to impulsively, choosing them as ends. He believes that his good ends are good because they are objects of his reflective choice – that is, his choice to preserve, promote, or realize them. Yet this person is committed to the view that there is nothing unconditionally good from which the goodness of things derives. In particular, he denies that his power of reflective choice is unconditionally good. The agent thinks that the goodness of a thing is conditional on its being an object of his reflective choice. Therefore, according to him, the goodness of his power of reflective choice is itself conditional on his exercise of this power. In his view, his power of reflective choice does not count as good unless he at least makes a reflective choice to preserve it, for example, to keep himself alive. Yet he can easily envisage a context in which he would not choose to preserve his power of reflective choice. For example, he imagines that he will die in a matter of months unless he takes steps to procure an experimental medication. The medication is expensive and would consume resources desperately needed right now to preserve the lives of his loved ones. In this situation, he concludes, he would not choose to preserve his power of reflective choice. Since the agent can conceive of circumstances such as this, he can conceive of contexts in which his power of reflective choice would not be good. Korsgaard apparently thinks that such a person would be condemned to complete normative skepticism. But the question is: why would he be? It seems that he would have reasons to do certain things – for example, to preserve, promote, or realize objects of his reflective, as opposed to his impulsive, choice.

Perhaps Korsgaard would respond to this example by agreeing that in light of the agent's account of the conditions of value, he is not compelled to embrace normative skepticism. Nevertheless, she might claim, the example does not realize its aim: it does not show that one who is committed to denying there to be anything unconditionally good from which the goodness of his good ends derives can avoid normative skepticism. For though the agent might not have reflected deeply enough to realize it, he is, by virtue of his account of the conditions of value, committed to affirming that there is something unconditionally good from which the goodness of his good ends derives. This something is not the power of reflective choice, but the *exercise* of this power: his reflective choice (i.e., choosing) itself. After all, the agent takes his good ends to be good because they are objects of his reflective

choice. And, Korsgaard might conclude, if he holds reflective choice to have this status, he must also hold it to be unconditionally good.

This response seems inadequate. For Korsgaard does not explain what would be irrational in the agent's holding that though his reflective choice of an object is what confers value on it, reflective choice is not itself unconditionally valuable. In general, that one thing confers a property on a second thing does not entail that the first thing possesses the property at all, let alone unconditionally. Some university presidents confer the Ph.D. on graduate students. That does not entail that these presidents themselves possess a Ph.D.[20] Of course, Korsgaard might insist that value is a special property; unlike many other properties, it is such that whatever confers it must possess it. But it is highly questionable whether this is the case. Suppose I hold that what confers badness on something is that it be an object of rational disapproval. I would not thereby have to hold that rational disapproval is bad at all, let alone unconditionally bad.[21] Korsgaard has given us inadequate grounds for thinking that, upon reflection, the agent does take there to be something unconditionally good from which the goodness of good ends derives. Therefore, she fails to rescue her claim that unless we are committed to there being such a thing, we push ourselves into utter normative skepticism.

In short, it is questionable both whether assuming there to be a categorical imperative itself compels one to hold that he has good ends (in the robust sense in question) and whether the only way to deny that one has such ends is to embrace complete normative skepticism.

3.6 From Good Ends to the Unconditional Value of Humanity: The Regressive Argument

Nevertheless, it would obviously be very significant if, on Korsgaard's reconstruction, it turned out that if we take ourselves to have good ends, then we must hold humanity to be unconditionally good. In this section I set out the argument; in the next I criticize it.[22]

Korsgaard characterizes the argument as "regressive," which means that "something is taken as given or actual and the conditions of its possibility are explored."[23] In this case, what a person engaged in the argument takes as given is that she has good ends. The burden of (what I call) the "regressive argument" is to show the following: if an agent takes as given that she has good ends, then she must (is rationally compelled to) hold that humanity is unconditionally valuable. She must hold this because, upon reflection, she will find that these ends ultimately derive their goodness from something unconditionally good, namely from humanity.

Korsgaard's interpretation of "humanity" coheres with that offered at the beginning of this chapter (3.1). Korsgaard embraces the view that the "characteristic feature" of humanity is the capacity to set ends.[24] It is through

practical reason that we set ends. Whenever we act, we do so on some self-given principle of practical reason, that is, some maxim. In giving ourselves maxims of acting, we set ourselves ends; for each maxim contains (a description of) an end (see section 1.2).[25] "Human beings are distinguished from animals," says Korsgaard, "by the fact that practical reason rather than instinct is the determinant of our actions."[26] According to Korsgaard each and every one of an agent's ends is set by reason, though only his morally obligatory ends are set entirely by reason.[27] Korsgaard insists that we should not understand humanity merely as a capacity to set moral ends, but, more generally, as a capacity to set ends for our actions, as opposed to behaving on instinct as do other animals. To value humanity is to value the capacity to set ends, wherever it manifests itself. In the context of the regressive argument, Korsgaard sometimes substitutes for "humanity" the terms "rational nature" or "the power of rational choice." She employs these terms as equivalent.[28]

Although Korsgaard summarizes the regressive argument in various works, she offers her most thorough account of it in "Kant's Formula of Humanity."[29] I believe that the regressive argument unfolds as follows:

i. You take it that some of your ends are good.

Therefore,

ii. You hold there to be a sufficient condition of their goodness: something that is either itself unconditionally good or that derives its goodness from something unconditionally good.
iii. The sufficient condition of the ends' goodness does not lie in the ends themselves.
iv. It does not lie in your having an inclination for them.
v. The sufficient condition of the ends' goodness is not that they contribute to your happiness, or even to everyone's happiness.

On reflection,

vi. You hold that the sufficient condition of the goodness of the ends you take to be good is that they be objects of your rational choice.

So,

vii. You must hold your power of rational choice (humanity) to be unconditionally good.

On reflection,

viii. You must hold that the sufficient condition of the goodness of each agent's good ends is that they be objects of the agent's rational choice.

Therefore,

ix. You must hold everyone's power of rational choice (humanity) to be unconditionally good.

In embracing step i, an agent sets out his assumption that he has good ends. According to the argument, given this assumption, he is compelled to embrace ii, namely the idea that there is a sufficient condition of the ends' goodness – something that is either itself unconditionally good or that derives goodness from something unconditionally good. Steps iii–v are supposed to eliminate various candidates the agent might consider for the sufficient condition of the goodness of the ends he takes to be good. Step vi represents what Korsgaard calls the crucial step of the argument – that is, the notion that, upon reflection, an agent takes the sufficient condition of the goodness of these ends to be their status as objects of his rational choice. To allay possible misunderstanding, let me emphasize from the outset that the notion of sufficiency Korsgaard employs appears to be what we might (rather awkwardly) call "becausal" sufficiency. To affirm that *A* is the "becausally" sufficient condition of *B* is to affirm that if *A*, then *B* because *A*. So it appears that we might paraphrase vi as follows. Suppose you have an end and you take it to be a good one. You hold that if this end is an object of your rational choice (as, according to Kant, all of your ends are), the end is good *because* it is an object of your rational choice. In effect, you hold that what confers value on any end of yours that you take to be good is its being an object of your rational choice.[30] Moving forward in the argument, the combination of vi and ii is supposed to yield vii, namely that an agent must take his power of rational choice (humanity) to be unconditionally good. Moreover, since an agent embraces vi, suggests Korsgaard, he must also accept viii, namely that the sufficient condition of the goodness of each agent's good ends is that they be objects of the agent's rational choice. The move from viii to ix, the conclusion, parallels that from vi to vii. According to Korsgaard (who is, of course, following Kant), if an agent embraces the conclusion of the regressive argument, he must recognize moral obligations to himself and others. It is debatable precisely what these obligations are, but I do not focus on this issue until Chapter 8.

Turning to the details of the regressive argument, we find that ii follows from i. In i, we assume that we have good ends. Good ends are, on the conception we are employing here, fully justified. That they are yields ii. To hold an end to be fully justified is, says Korsgaard, to hold there to be some ("becausally") sufficient condition of its goodness that is itself unconditionally good or which derives its goodness from something unconditionally good. The question is: what is this sufficient condition? Steps iii–vi arise from efforts to answer this question.

The third step of the regressive argument rejects a form of realism regarding the good – the notion that goodness is simply inherent in certain ends themselves. It is easy to sketch an example of the kind of position iii disclaims. An environmentalist who has the end of preserving the maximum number of living species on earth might hold that this end not only meets each of Korsgaard's criteria for goodness but is itself unconditionally good. It is, he thinks, good in every context that a maximum number of (currently existing) species be preserved. In "Kant's Formula of Humanity," Korsgaard briefly underscores a Kantian reply to this kind of position. The environmentalist, goes the reply, is confused about the source of his end's goodness. He believes that he wants to maximize species preservation because such preservation is intrinsically good. Yet upon reflection he would find that any goodness had by species preservation would actually stem from his desiring it. Korsgaard says: "[I]t looks as if the things you want, if they are good at all, are good because you want them – rather than your wanting them because they are good."[31] But, the Kantian reply continues, if the goodness of species preservation derives from the agent's desire for it, then it is not unconditionally good. Korsgaard cites approvingly Kant's claim that: "All objects of the inclinations have only a conditional worth; for if there were not inclinations and the needs based on them, their object would be without worth" (GMS 428). Since, if some rational being did not want maximum species preservation, it would be devoid of worth, it is not unconditionally good. Of course, Korsgaard has at her disposal another means of showing that maximum species preservation fails to be unconditionally good: Kant's famous (and much criticized) *Groundwork* I argument that nothing except a good will can even be conceived as unconditionally good.[32]

Having assumed that we have good ends, we are inquiring into what constitutes the sufficient condition of their goodness. In accepting step iii, we have endorsed the notion that their goodness must derive somehow from the nature or concerns of rational beings. A natural proposal for an agent to make at this point is that his good ends are good simply because they are objects of his desire. In other words, a sufficient condition of his ends' goodness is that he have an inclination for them. In step iv Korsgaard denies that this is the case. Her denial seems very plausible. That an agent has an inclination for an object does not entail that the object is good. Someone might have a craving to smoke cigarettes, but her having it might not, even in her own view, make smoking good. For she might herself acknowledge that though smoking gives her a momentary pleasure, it ultimately fails to promote her happiness and is therefore not good.[33]

Yet what about happiness itself? Could it not be the case that a good end is good because it contributes to happiness? There are two possibilities here, both of which are addressed in step v. According to the first, an agent's end is good by virtue of its contributing to his own happiness. Korsgaard rejects this possibility mainly by appealing to Kant's claim that "we do not believe

that happiness is good in the possession of one who does not have a good will."[34] Recall that in Korsgaard's view a good end must be fully justified. For her this means that it must derive its goodness from something that is unconditionally good. If contributing to an agent's own happiness is to justify the goodness of his end, then, she thinks, the agent's happiness must be unconditionally good. Yet, according to Kant, an agent's happiness is not unconditionally good. There is a context in which his happiness is not good, namely when it is not accompanied by a good will. A rational egoist might object to this contention, arguing in what Korsgaard calls "a remarkable feat of egocentrism" that his own happiness is unconditionally good, but I do not pursue this point here.[35]

According to the second way of trying to use happiness to bring the regress to a close, an end is good by virtue of its contributing to everyone's happiness. We might claim that everyone's happiness – that is, the state of affairs in which every individual is happy – is unconditionally good. A good end is good because it contributes to the realization of this unconditionally good state of affairs.

Against this suggestion, Korsgaard appeals to Kant's notion that the good must provide reasons for action that apply to every rational being. In particular, she emphasizes something that she takes to follow from this requirement, namely that if an end is good, then all rational agents must be able to share it. The end must be a "consistent, harmonious object."[36] What is a "consistent, harmonious object"? This much is clear. In Korsgaard's view, we cannot say that in pursuing *his own* happiness, each agent would be pursuing a consistent, harmonious object. Suppose each agent were pursuing his own happiness. Korsgaard endorses Kant's view that what would result is a harmony like that suggested in the pledge of King Francis I to Emperor Charles V: "What my brother Charles would have [Milan], that I would also have" (KpV 28). The brothers do not really have a consistent object – the one wants to get Milan for himself, which would prevent his brother from getting it, and vice versa. In a similar way (as we have already noted in section 3.2), all agents pursuing their own happiness would not have a consistent object; each agent wants his own happiness to be promoted, which, in Kant's view, would prevent (at least some) other agent from promoting his. For example, part of Pete's happiness would be winning this year's tournament. Yet if he wins, then Boris couldn't be happy, since he was counting on victory as well. It seems that if an object is consistent and harmonious, then one agent's promoting it would not itself preclude any other agent from doing so.

Korsgaard concludes that everyone's happiness "does not form a consistent harmonious object."[37] This conclusion, however, does not follow from the understanding we have thus far attained of what it means to form one. For it is not clear that one agent's securing everyone's happiness (if we assume for a moment that it would be practically possible for one agent to do this) would itself preclude another agent from doing so. One agent (angel 1)

might initially bring it about that everyone was happy, while another agent (angel 2) might thwart a threat to everyone's happiness, for example, a threat from a natural disaster. So far, everyone's happiness does seem to be a consistent, harmonious object. Yet I suspect that there is a further condition on being such an object that we have not yet captured; a consistent, harmonious object must be realizable. And, according to Korsgaard, everyone's happiness is not. For each person to be happy, each person would have to have all of his desires satisfied. But the satisfaction of some agents' desires necessarily precludes the satisfaction of some other agents' desires. Not everyone can be happy. In short, Korsgaard argues that since unconditionally good objects must be harmonious, and everyone's happiness is not harmonious, everyone's happiness is not unconditionally good.

Moving forward in the argument, we have assumed that there are good ends – good in the very robust sense Korsgaard has specified. Given that there are good ends in this sense, we must be able to pinpoint the sufficient condition of their goodness. Yet the question remains: what is it? It will be helpful to cite the passage in which Korsgaard answers this question:

> Now comes the crucial step. Kant's answer, as I understand him, is that what makes the object of your rational choice good is that it *is* the object of a rational choice. That is, since we still *do* make choices and have the attitude that what we choose is good in spite of our incapacity to find the unconditioned condition of the object's goodness in this (empirical) regress upon the conditions, it must be that we are supposing that rational choice itself *makes* its object good. His idea is that rational choice has what I will call a value-conferring status.[38]

Here Korsgaard seems to be saying: a good end derives its value from being the object of an agent's exercise of a certain capacity, namely his power of rational choice. Suppose an agent takes one of his ends to be good. Upon reflection, suggests Korsgaard, he will conclude that it has this status by virtue of his having exercised his power of rational choice with respect to it. That the agent, not driven by impulse but rather guided by reason, chose this end suffices, in the agent's considered view, to make it good. "We act as if our own choice were the sufficient condition of the goodness of its object: this attitude is built into (a subjective principle of) rational action."[39] An agent holds that a sufficient condition of the goodness of his good ends is that they be the object of his rational choice. This is the sixth step of the regress argument.

How does Korsgaard move from step vi to vii? Suppose an agent embraces the idea that a sufficient condition of the goodness of his good ends is that they be objects of his rational choice (vi). According to ii, he is then committed to the view that his rational choice is either itself unconditionally good or derives its goodness from something unconditionally good. The regressive argument has the agent affirm the latter. He affirms that his exercising his power of rational choice derives its goodness from this power itself, which is

unconditionally good: "[R]egressing upon the conditions, we find that the unconditioned condition of the goodness of anything is rational nature, or the power of rational choice."[40] It is not rational choosing, but the power of rational choice that the agent holds to be unconditionally good. The argument's seventh step finishes the regress on conditions of the goodness of the agent's ends. It maintains that an agent must view his power of rational choice to be unconditionally good.

Korsgaard's transition from this step to the conclusion that you must hold everyone's power of rational choice to be unconditionally good appears to go as follows. According to step vi, you (an agent who has affirmed that she has good ends) are rationally compelled to view yourself as having "value-conferring status" in virtue of your power of rational choice. But "[i]f you view yourself as having a value-conferring status in virtue of your power of rational choice, you must view anyone who has the power of rational choice as having, in virtue of that power, a value-conferring status."[41] In short, you must embrace viii. Moreover, just as your holding yourself to have value-conferring status requires you to hold your power of rational choice to be unconditionally good (the move from vi to vii), so your holding others to have value-conferring status requires you to hold their power of choice to be unconditionally good. In effect, as step ix states, you must hold everyone's power of rational choice (humanity) to be unconditionally good.

3.7 The Failure of the Regressive Argument

I do not believe that this argument succeeds in showing that, if an agent assumes that he has good ends, then he must hold that humanity is unconditionally valuable. I try to highlight two problems with the argument.

The first difficulty concerns step v, specifically the denial that a sufficient condition of the goodness of your good ends is that they contribute to everyone's happiness. As a basis for this denial, Korsgaard appeals to the notion that such a condition would have to be unconditionally good. However, she argues, if something is unconditionally good, then it is a "consistent, harmonious object," which entails that it is realizable. But everyone's happiness is not realizable, since making some people happy necessarily involves precluding others from being happy. Therefore, everyone's happiness is not unconditionally good.

For the sake of argument, let us grant that if we characterize happiness as the complete satisfaction of all inclinations, as Kant sometimes does, then happiness is not a harmonious object.[42] Given the conflicting set of desires people have (and, let's say, necessarily will have), the happiness of some would always prevent the happiness of others.[43] Yet why should we embrace this desire-satisfaction account of happiness in the first place? Philosophers who argue that everyone's happiness is unconditionally good need not employ such an account. They might, rather, invoke a conception according to

which happiness is a harmonious object. These philosophers might contend, for example, that being happy amounts to having a number of goods – for example, loving relationships, a sense of self-respect, security – such that one person's having them would not preclude anyone else from having them. In order for her rejection of happiness as unconditionally good to be effective, Korsgaard must, it seems, show not only that happiness on a Kantian conception fails to be a harmonious object but also that happiness on other (plausible) conceptions fails as well.

Korsgaard might here appeal to Kant's view that the only thing we can conceive of as good without qualification is a good will. Everyone's happiness, no matter how we define it, would not be good without qualification, Korsgaard might argue. Kant suggests a thought experiment for determining that something fails to be good without qualification. In it, we ask ourselves whether, in some possible context, an "impartial rational spectator" would find that the thing was *not* good (GMS 393). If, in our view, there is such a possible context, then we conclude that the thing is not good without qualification. According to Kant, the notion that only a good will is unconditionally good is to be found in the "moral cognition of common human reason" (GMS 403). Kant defends the notion with an appeal to our everyday moral intuitions. Korsgaard might claim that there is a possible context in which an impartial rational spectator would find that everyone's being happy is not good, namely when some happy individuals did not have a good will.[44]

But what is a good will? Interpreting Kant's notion (or notions) of a good will is a challenging task, and I do not attempt to do so thoroughly here. For our purposes, we can take note of two ways in which Kant seems to employ "good will," as it applies to us, agents who can be tempted by their inclinations to act contrary to the moral law. According to the first way, a good will is a particular sort of *willing* or, what for him amounts to the same thing, of acting (section 1.4). Kant writes of "the unqualified [*uneingeschränkten*] worth of actions" (GMS 411), presumably of actions done from duty, which he has previously stated to have "unconditional and moral worth" (GMS 400). Since, according to Kant, the *only* thing good without qualification (*ohne Einschränkung*) is a good will, it appears that sometimes "good will" refers to a certain kind of action, that is, that done from duty.[45] I call this usage the "particular action" understanding of a good will.

According to a second way in which Kant employs "good will," it refers not to a particular kind of action an agent might perform but rather to a kind of character she might have. An agent has a good will on this usage just in case she is committed to doing what duty requires, not just in this or that particular action, but overall. Presumably if an agent has this commitment, then she will sometimes act from duty. (For example, she will invoke duty as her incentive to do what is morally required in cases where she is tempted by her inclinations to act contrary to what morality demands.) Kant intimates

that having a good will amounts to having a certain kind of character in the first paragraph of *Groundwork* I. Right after suggesting that the only thing good without qualification is a good will, he tells us that certain qualities of temperament (e.g., courage or resolution) "are undoubtedly good and desirable for many purposes, but they can also be extremely evil and harmful if the will which is to make use of these gifts of nature, and whose distinctive constitution is therefore called *character*, is not good" (GMS 393). Later Kant is discussing a man who is by temperament cold and indifferent to others, but who, from duty, acts beneficently. "It is just then," says Kant, "that the worth of character comes out, which is moral and incomparably the highest" (GMS 398–399). These passages suggest that "good will" refers not merely to a particular kind of action, but to a kind of character that can be expressed in action. Sometimes Kant employs what I (following Karl Ameriks) call the "whole character" conception of a good will.[46]

It appears that Kant employs (at least) two *different* notions of a good will. For it seems that one could have a good will on the particular action understanding, yet not have a good will on the whole character conception. After all, why could one not act from duty in a particular case, yet not be committed overall to doing what morality requires? I do not pursue this question here. For our purposes, it suffices to make clear which of these notions of a good will we are employing at a given point, leaving aside the issue of whether, ultimately, they coincide.

Let us now return to the argument we were considering before our brief discussion of a good will. Although I am not entirely sure, I believe that when Korsgaard invokes the notion of a good will, she has in view the whole character conception.[47] An impartial rational spectator, Korsgaard might claim, would not find everyone's happiness to be good if some happy individuals did not have a good will in the sense of an overall commitment to doing what morality requires. On Kant's view, of course, if an agent does not have a good will, then she might not only stray from duty sometimes, but actually make a habit of doing so. But do we really hold that an impartial rational spectator would not approve of everyone's being happy if some did not have a good will? Some of us might imagine such a spectator reacting to this scenario as follows: "The agents who do not have a good will do not morally deserve their happiness. However, this does not mean that everyone's happiness is not good. For in the scenario in question, the actions of those without a good will – their lying, cheating, and so forth – do not prevent others from being happy. Since they do not, the scenario is actually still good. Granted, a scenario in which everyone is happy but some are without a good will is not *as good* as one in which everyone is both happy and has a good will. Yet the former scenario is still good." In Kant's view, of course, this reaction does not conform to ordinary moral reason. But this view seems dubious.

In any case, an appeal to the thought experiment in question would be a dangerous tactic for a defender of the regressive argument to take.

For employing it might show that the power of rational choice is itself not good without qualification. Recall that the power of rational choice is the capacity to set all sorts of ends, including, but not limited to, morally good ones. Consider an agent who has the power of rational choice, but employs it with no concern for whether he is conforming to moral requirements. Whenever this agent is inclined to realize a morally bad end, he does all he can to do so. In this context, which certainly seems to be a possible one, would an impartial rational spectator judge the power of rational choice to be good? That she would is, I think, doubtful.[48] For it is his power of rational choice that enables the agent to choose the morally bad ends. If he did not have this power, yet instead sought merely instinctual gratifications, then he might cause much less harm. It would not be helpful to respond here that what makes the agent's power of rational choice unconditionally good is that by virtue of having it, he has a further capacity, namely that to develop a good will. For we are imagining a case in which the agent never exercises his capacity to develop a good will. And in this case, why should we hold this capacity to be good, rather than, say, indifferent?

This discussion allows us to see in Kant's doctrine an apparent tension that I am unsure how to resolve. Kant holds that the good will alone is good without qualification (GMS 393). He also holds that rational nature is unconditionally good (GMS 428). So unless I am overlooking some subtle distinction between being good without qualification and being unconditionally good, Kant seems to be identifying the good will and rational nature. But on the conceptions of the good will I sketched above – that is, the particular action and whole character conceptions – it seems that a being could possess rational nature and yet not have a good will. As we have just seen, that a being has rational nature does not entail that he ever acts from duty, let alone that he has committed himself to an overall policy of doing what duty requires. As Ameriks notes, some philosophers, perhaps based on such considerations, have attributed to Kant the view that the good will *simply is* rational nature.[49] But if we substitute this understanding into the beginning of *Groundwork* I, we get nonsense. Since Kant's notion of the good will is problematic, so is appealing to this notion in an effort to rescue step v of the regressive argument.

The second difficulty with the regressive argument concerns step iii. In steps i and ii, you have assumed that some of your ends are good and that their goodness must derive from something unconditionally good. Step iii aims to rule out the possibility that your ends themselves count as this unconditionally good thing. Recall our example of the kind of position iii disclaims. An environmentalist who has the end of preserving the maximum number of living species on earth might hold that this end not only meets all of Korsgaard's criteria for goodness but is itself unconditionally good. Korsgaard suggests a Kantian response to this position. The environmentalist is mistaken about the source of his end's goodness, believing that he

wants to maximize species preservation because such preservation is intrinsically good, when, in reality, any goodness had by species preservation would actually stem from his desiring it. But given that the goodness of species preservation derives from the agent's desire for it, it is not unconditionally good, since the agent may well cease to desire it.

This argument does not threaten value realists who hold goodness to be inherent in their ends themselves.[50] For Kant does not here establish that these realists must accede that the goodness of what they take to be unconditionally good derives simply from their desiring it. Why, for example, must the environmentalist agree that the goodness of species preservation depends on his wanting it? Why can he not maintain that species preservation is unconditionally good, and thus good regardless of whether he (or anyone else) desires it?

It is once again open to Kant to appeal to his claim that the only thing good without qualification is a good will. Such an appeal might be more effective here than it was in eliminating the possibility that everyone's happiness was unconditionally good. Would an impartial rational spectator hold that the maximum number of currently existing species being preserved was good in every context?

There are some contexts in which such preservation would have what (we might plausibly think) the spectator would take to be bad effects. For example, in environmentally sensitive areas, preserving species might require closing businesses and thus causing hardship to workers and their families. However, that species being preserved would in some contexts have bad effects does not itself preclude it from being unconditionally good. As many commentators have remarked, a good will can also have what (we might plausibly think) a rational spectator would consider to be bad effects. Someone with a good will might be, as it were, cursed. When she acts from duty, she might not only fail to realize her ends but, by a "special disfavor of fortune," bring about the opposite of her aim. For example, her effort to save a choking victim might actually result in his death. Or her large donation to an emergency relief fund might end up in the hands of terrorists, financing their destruction of innocent civilians. If Kant ruled out something's being good without qualification on the grounds that in some contexts it had bad effects, then he would be compelled to rule out a good will itself.

But Kant suggests another basis for ruling things out: if an object, *disregarding its effects*, is good in all contexts, then it is good without qualification. Qualities such as "moderation in affects and passions," "self-control," and "calm reflection" are, Kant acknowledges, helpful in attaining all sorts of ends. Yet he denies that they are good without qualification, "for, without the basic principles of a good will they can become extremely evil, and the coolness of a scoundrel makes him not only far more dangerous but also immediately more abominable in our eyes than we would have taken him to be without it" (GMS 394). Kant seems to be suggesting here that when

"coolness" belongs to a scoundrel, it undergoes a "value reversal," to borrow a phrase from Berys Gaut.[51] What we (presumably imagining ourselves to be impartial rational spectators) often take to be good (e.g., coolness in an astronaut) becomes bad in some contexts. And it is bad considered independently of its effects. The coolness of a scoundrel is presumably bad even if, with its help, the scoundrel never manages to do anyone any real harm. Kant, of course, claims that a good will is the only thing that never undergoes a value reversal; an impartial rational spectator would hold that in every context it is good.

Would an appeal to this Kantian argument force value realists – those who take their ends themselves to be unconditionally good – to abandon their positions? In returning to our environmentalist, do we find that a maximum number of (currently existing) species being preserved undergoes a value reversal? Considered independently of its effects, is there a context in which an impartial rational spectator would not take this to be good? I suspect that answers to this question will differ. Those who see no inherent value in biodiversity will be drawn to the view that in many contexts an impartial rational spectator would take maximum species preservation to be indifferent rather than good. They might, for example, ask us to imagine the following world. The *human* species is fully flourishing and a maximum number of species have been preserved. Moreover (in the imagined world), if it comes to pass that a maximum number of species is no longer preserved (e.g., if thousands go extinct) humans would fully flourish just the same. Since maximum species preservation is important only insofar as it affects human flourishing, they might conclude, in the imagined world maximum species preservation would have no value to an impartial rational spectator. Others would disagree with this view, however, contending that even in that world the existence of the maximum variety of life would itself be valuable. To deny this would, in effect, be to embrace the appallingly prideful view that human beings are all that really matters, the others might say; and this is surely not a view that an impartial rational spectator would adopt. My aim here is not to settle the issue. It is merely to illustrate that Kant's argument here is controversial at best. Through his appeal to ordinary moral reason, he falls far short of showing that the environmentalist is rationally compelled to give up the notion that species preservation is unconditionally good.

Of course, there are many other candidates for unconditional goodness besides a maximum number of species being preserved. Someone might, for example, defend the view that knowledge, courage, friendship, beauty, and so forth are good in themselves, independently of any agent's desiring them. It is open to Kant to challenge any item on such a list on the grounds that, unlike a good will, it undergoes a value reversal in some context. But I suspect that this tactic would be no more effective with regard to these purportedly unconditionally good things than it was regarding species preservation.[52] I hope that my discussions of species preservation as well as

universal happiness have illustrated the vulnerability of Kant's claim that nothing other than a good will can be considered unconditionally good.

If I am correct that this claim does not really threaten the value realist position, then step iii of the regressive argument is without sufficient support. We are free to hold that some of the objects of our desires are good not because we desire them, but are rather good in themselves. Given that iii lacks sufficient support, the regressive argument does not prove that if we take ourselves to have good ends, we must hold humanity to be unconditionally good.

3.8 Shortcomings in the Derivation of the Formula of Humanity

Let me now crystallize my main findings regarding Kant's derivation of the Formula of Humanity. There are two main reasons why I do not believe that this derivation, even as reconstructed by Korsgaard, is successful.

First, Kant does not prove that if we take there to be a principle that conforms to his basic concept of the supreme principle of morality, then we must hold there to be something unconditionally good. Granted, if there is a supreme principle of morality, then every agent must always have a motive available to him for conforming to it. As our example of PW illustrated, however, this motive need not be the notion that conforming to the principle is itself unconditionally good or enables the agent to secure something unconditionally good. The "ground" of a categorical imperative might be each agent's being rationally compelled to view his conforming to this principle as something good for him (though not necessarily good from an impartial perspective).

The example I have offered of a principle "grounded" in this way – "Maximize your power over rational beings" – is, in my view, a repellent candidate for the supreme principle of morality. I venture that most readers would agree. Nevertheless, it is illegitimate to infer that if a principle conforms to Kant's basic concept of the supreme principle of morality, then we must take there to be something that all agents are rationally compelled to hold to be unconditionally good. Kant does not establish that unless we hold there to be something unconditionally good (in his agent-neutral sense), we cannot hold there to be a universally and unconditionally binding practical principle (a categorical imperative).

The second main difficulty with the derivation of the Formula of Humanity is, I think, more important than the first. Even if holding there to be a categorical imperative requires holding there to be something unconditionally good, Kant does not establish that this must be humanity. Even in Korsgaard's ingenious reconstruction, we find no good reason to rule out the possibility that the unconditionally good "ground" of a categorical imperative is everyone's happiness. The regressive argument fails to threaten utilitarianism. Moreover, it contains no compelling arguments against various forms of value realism. Kant's response to those who would hold that

what is unconditionally good is the objects of their inclinations – that is, their ends such as maintaining biodiversity or gaining systematic knowledge of the universe – is that these objects are really only conditionally valuable, for if these persons did not value them, then the objects would be devoid of worth. But why must we agree that the value of such objects derives solely from our wanting them? It would hardly seem unreasonable for someone to maintain that Kant has things backward; it is not that environmental preservation is valuable (just) because we desire it, but rather that we desire it because it is valuable. Yet the argument Kant suggests against the notion that some objects of our inclinations are unconditionally good is the controversial and, in my view, ineffective one that a good will alone is unconditionally good. Kant does not show that humanity alone is capable of being the unconditionally good "ground" of the supreme principle of morality.

In sum, as I have argued, the derivation of the Formula of Humanity contains two highly questionable steps. First, Kant does not establish that if there is a supreme principle of morality, then there is something unconditionally good. Second, even if we assume that his first step succeeds, he does not show that this unconditionally good something must be humanity.

It is worth pointing out that even if these two steps succeed, the derivation might falter in its third step. If humanity is unconditionally good, then must we always treat it not merely as a means but also as an end? In Chapter 8, we explore what it means to treat humanity as an end or, equivalently, as an end in itself. It seems that in Kant's view treating humanity as an end in itself involves treating it not only as something of unconditional worth, but also as something of *incomparable* worth. Something has incomparable worth if it cannot be legitimately sacrificed for or replaced by anything else. Now let us assume that we hold humanity to be unconditionally valuable and that, since we do, we are rationally compelled to treat it as such. Are we also rationally compelled to treat humanity as incomparably valuable? That is not at all clear. Take a case of an individual who, to preserve the humanity in twenty innocent hostages, sacrifices the humanity in one person, a terrorist, by killing him. It seems that the individual *might* reasonably contend that he treated humanity as unconditionally valuable, though he did not treat it as incomparably valuable. He treated humanity as unconditionally valuable in that he attempted to preserve as much of it as possible, the individual might maintain. But he did not treat it as incomparably valuable, since in his own view he sacrificed the humanity in one person to preserve the greater value inherent in the humanity of twenty people. In short, even if Kant's derivation showed that we are rationally compelled to treat humanity as unconditionally valuable, he would need a further argument to show in addition that we must treat it as incomparably valuable. So, in effect, Kant would need an additional argument to show that we must treat humanity as an end in itself.[53]

4

The Derivation of the Formula of Universal Law: A Criterial Reading

4.1 Main Steps of the Derivation on the Criterial Reading

According to the traditional reading, Kant's *Groundwork* derivation of the Formula of Universal Law has an obvious flaw. It thus makes sense to look elsewhere for more promising derivations of a Kantian principle. Allison reconstructs Kant's second *Critique* derivation of the Formula of Universal Law, Korsgaard his *Groundwork* derivation of the Formula of Humanity. Yet we have found that neither of these reconstructed derivations succeeds. The prospects for a derivation of a Kantian principle seem very dim. The rest of this book aims to show that they are brighter than these results suggest.

I challenge the traditional reading of Kant's *Groundwork* derivation of the Formula of Universal Law. According to the "criterial reading" I defend, Kant's *Groundwork* I derivation of this formula can be broken down into three main steps. First, Kant tries to pinpoint criteria that we, on reflection, believe that the supreme principle of morality must fulfill. Second, Kant attempts to establish that no possible rival to the Formula of Universal Law fulfills all of these criteria. Third, at least implicitly Kant argues that the Formula of Universal Law remains as a viable candidate for a principle that fulfills all of them. With these three steps, Kant strives to prove that if there is a supreme principle of morality, then it is this formula. In short, Kant argues by elimination. When we have before us a clear notion of the characteristics the supreme principle of morality must possess, Kant suggests, we are able to eliminate every candidate for this principle except the Formula of Universal Law (or equivalent principles).

This chapter aims to make room for the criterial reading of Kant's derivation.[1] It starts by examining a reading of the *Groundwork* I derivation that has been offered by Christine Korsgaard. Korsgaard does not explicitly confront the traditional interpretation of this derivation. Nevertheless, if her reading were successful, then it would constitute an alternative to the traditional interpretation that might render the criterial reading unnecessary.

However, I argue against Korsgaard's reading on textual as well as philosophical grounds (section 4.2). I then turn to the traditional interpretation itself. A brief examination of the structure of *Groundwork* I (4.3) helps us to see the serious flaws in one version of the traditional interpretation, that is, a version discussed by Allison (4.4). But the bulk of the chapter is devoted to developing the criterial reading as an alternative to the other version of the traditional interpretation, the version offered by Aune (4.5–11). We must acknowledge that Aune's version has considerable force. However, I try to show that despite initial appearances to the contrary, the criterial reading is compatible with Kant's *Groundwork* I (and even his *Groundwork* II) derivations of the Formula of Universal Law. Later, by the end of Chapter 7, I hope it will be clear that the criterial reading renders Kant's argument far more philosophically powerful and interesting than it is under the guise of Aune's interpretation.

4.2 Korsgaard's Reading of the Derivation

It seems that according to Korsgaard Kant's *Groundwork* derivation not only suffers from no obvious gaps, but actually succeeds.[2] If her interpretation yielded a compelling, textually grounded argument, then there would be little reason to develop the criterial reading. In my view, however, it does not.

Korsgaard's interpretation seems to go as follows:[3]

i. Kant is engaged in "motivational analysis of the notion of duty or rightness. Kant is analyzing the good will, characterized as one that does what is right because it is right, in order to discover the principle of unconditionally good action," and he assumes that "the reason why a good-willed person does an action, and the reason why the action is right, are the same."[4]
ii. The reason in both cases is constituted by what Korsgaard calls the "legal character" of the good-willed person's maxim – that is, the maxim's capacity to express a demand on us, its normative force.[5]
iii. Kant holds that the legal character (normative force) of the agent's maxim must not derive from any external source, such as God's commands. The reason is that "if there were an outside source of legal character, then that source, rather than legal character itself, would be what makes the action right."[6]
iv. And if that were so, by the equivalence mentioned in step i, the agent would not be acting on the maxim because of the maxim's normative force, but because of the normative force of the outside source. For example, the agent would not be acting on the maxim because it was right to do so, but because God commanded that she act on it. So, given step i, the normative force of the maxim of the action cannot derive from any external source.

v. What then constitutes the maxim's normative force? The only alternative to dependence on an external source is that the maxim's normative force is constituted by the fact that the maxim has "intrinsic lawlike form."[7] (Apparently, a maxim has an intrinsic lawlike form when acting on it is required by some law that does not owe its validity to anything external to the will [e.g., to God].)
vi. This lawlike form must be specified by the universalizability test, that is, by the Formula of Universal Law. If acting on a particular maxim is required by the universalizability test, the maxim is one of duty; it is "one that you must will as universal law. And this means that the maxim is a law to which your own will commits you. But a maxim to which your own will commits you is normative for you."[8]
vii. Hence only the Formula of Universal Law (and, presumably, equivalent principles) can confer lawlike form on the maxims of duty, and hence only it can be the supreme principle of morality. And that is the conclusion of the derivation.

Korsgaard's interpretation is correct in laying stress on the importance of motivational analysis and the good will in the derivation. But her interpretation has little textual support where it is most innovative, and it also yields an extremely problematic argument.

First, crucial to the interpretation is that Kant sets up a sharp dichotomy between intrinsic lawlike form and an external source of normative force (a dichotomy deployed in steps iii–vi). But there is no firm textual evidence that he exploits this dichotomy in the *Groundwork* I derivation. The only textual evidence for deployment of the dichotomy that Korsgaard cites is GMS 402:

> Since I have deprived the will of every impulse that could arise for it from obeying some law, nothing is left but the conformity of actions to universal law as such, which alone is to serve the will as its principle, that is, *I ought never to act except in such a way that I could also will that my maxim should become a universal law.* Here mere conformity to law as such, without having as its basis some law determined for certain actions, is what serves the will as its principle, and must so serve it, if duty is not to be everywhere an empty delusion and a chimerical concept.

She asserts that with the words "without having as its basis some law determined for certain actions" (or "without assuming any particular law applicable to certain actions" in the translation she employs), Kant means to block the claim that the law could be an independent one – that is, could derive its validity from any external source.[9] But this is not what Kant says; there is no mention of independence or externality here. Why should we take "some law determined for certain actions" to refer to an external law? Surely some explanation is needed, especially since Korsgaard does not provide any other textual evidence that Kant is in this passage concerned

with blocking the possibility that the principle of a good will could have an outside source.[10]

Second, moving from a concern regarding the textual basis of the argument to one regarding its substance, steps iii and iv do not succeed in ruling out an external source of normativity. Take Korsgaard's example of an external source: acting on some maxim because God has commanded it. The dutiful agent who believes in the divine command view of morality holds, just as the Kantian agent does, that he is acting on a maxim of duty because it is right to do so; but what its rightness consists in, according to the divine command moralist, is its being commanded by God. That addition does not affect his motivation to do what is right; it merely tells him what the property of rightness is. Korsgaard replies to this kind of objection that the maxim's "conformity to divine law can only make a maxim extrinsically, not intrinsically, legal."[11] But given that an intrinsic property is one that is necessarily possessed, the divine command moralist can simply deny the quoted claim. It is not a contingent fact according to him that God wills what is right; on the contrary, it is precisely because God wills something that it is right. One may of course dispute this view, but then the objection is to the substance of the divine command moralist's analysis, not to its making normative character extrinsic.

Finally and most importantly, Korsgaard's interpretation fails to give Kant a reasonable justification for the introduction of the Formula of Universal Law. For the argument in step vi would at best show that this formula is one principle that could test for the intrinsic lawlike form of a maxim. It does not show that it is the only principle that could do this. (The difficulty with establishing the uniqueness of the Formula of Universal Law as a viable candidate for the supreme principle of morality is familiar to us from our discussion of Allison in section 2.5.) In fact, there are many principles that would not derive their normative force from an external source, yet are not equivalent with the Formula of Universal Law. For example, consider two principles we have already recognized not to be equivalent to this formula (2.5): the weaker principle WU, "Act only on that maxim which, when generalized, could be a universal law," and the bizarre principle BP, "Act only on that maxim that you *cannot*, at the same time, will that it become a universal law." Why could not a maxim's being required by one of these principles signal that the maxim has intrinsic lawlike form? We have been given no explanation for how Kant can rule out these other principles. In short, there is a gap in Korsgaard's argument between the notion that a maxim of duty must have an intrinsic lawlike form and the notion that it has this form only if it is required by the Formula of Universal Law. And to me this gap seems almost as large as the one the traditional interpretation finds in the derivation.[12] Korsgaard's reading of the *Groundwork* I derivation does not constitute a viable alternative to the traditional interpretation.

4.3 The Structure of *Groundwork* I

In his preface to the *Groundwork*, Kant sets out his goals: to locate and to establish the supreme principle of morality (GMS 392). In Section I Kant attempts to locate the supreme principle of morality in the sense of specifying what it is, *if* there is one.[13] Appealing to (what he takes to be) ordinary moral views, Kant tries to find the principle that, on reflection, we hold to be at work in our moral practice. It is easy to overlook that this is what Kant is attempting to do. Philosophers have focused so much on Kant's discussion of the value of acting from duty – as opposed to acting from sympathy, for example – that one gets lulled into assuming that Kant's foremost interest is in specifying necessary and sufficient conditions for an action's having moral worth.[14] Near the end of *Groundwork* I, Kant proclaims success at isolating the principle we hold to be at work in our moral practice: "Thus, then, we have arrived, within the moral cognition of common human reason, at its principle, which it admittedly does not think so abstractly in a universal form, but which it actually has always before its eyes and uses as the norm for its appraisals" (GMS 403–404).[15] This principle is the Formula of Universal Law.[16] In *Groundwork* I, Kant's main concern is to show that if there is a supreme principle of morality, then it is this formula.

What, in broad outline, is Kant's route to the Formula of Universal Law? Before arriving at it, Kant discusses at length the good will, duty, and moral worth. Since his primary aim in *Groundwork* I is to construct an effective derivation of this formula, it is reasonable to suppose that he thinks this discussion to be necessary if he is to do so. The discussion includes the claims (roughly) that only a good will is good without qualification (GMS 393); that all and only actions from duty have moral worth (GMS 397–399); that the moral worth of actions from duty stems not from their effects, but from their maxim (GMS 399–400); and that duty is the necessity of an action done from respect for the law (GMS 400). A plausible interpretation of *Groundwork* I must explain why in Kant's view at least some such claims must turn out to be true if he is to succeed in his derivation of the Formula of Universal Law.

4.4 The Failure of One Version of the Traditional Reading of the Derivation

This brief reflection alone leads us to a ground for rejecting one version of the traditional interpretation of *Groundwork* I, namely the one according to which Kant invokes the "principle of rightness universalism." According to this version (section 1.4), Kant presents the Formula of Universal Law in a parenthetical clause aimed at elucidating the prescription that the will conform its actions to universal law as such. This prescription is interpreted to be the principle of rightness universalism RU, namely: "If a maxim or

action is judged permissible for a rational agent in given circumstances, it must also be judged permissible for any other rational agent in relevantly similar circumstances." Kant tries to reach the Formula of Universal Law by embracing RU, then claiming (without argument) that RU entails the Formula of Universal Law. Does this interpretation of Kant's argument explain how his discussion of the good will, duty, and moral worth are necessary to his locating the Formula of Universal Law? Since, according to the interpretation, Kant moves directly from RU to the Formula of Universal Law, an advocate of the interpretation might hope to find these discussions necessary for a Kantian defense of RU. However, the discussions seem totally irrelevant to the issue of whether we should embrace RU. For example, two of the central claims Kant makes in them are that all actions from duty have moral worth and that the moral worth of actions done from duty does not at all depend on the actions' effects. Without in the least threatening RU's legitimacy, we can deny these claims. We can maintain instead that only some actions done from duty have moral worth and that these actions have such worth because they bring about good effects – for example, an increase in the general welfare. In short, this version of the traditional interpretation is to be rejected on the ground that it does not account for the role Kant's complex discussion of ordinary moral views plays in his route to the Formula of Universal Law.

Two additional reasons support rejection of this version. First, according to it, Kant suggests that the supreme principle of morality (whatever it is) must require "the conformity of actions to universal law as such" (GMS 402). In this suggestion we are supposed to find an endorsement of RU. But is Kant really endorsing it there? It is far from obvious that for Kant the requirement to conform one's actions to universal law as such amounts to RU. After all, Kant makes no mention here of the concept of relevantly similar circumstances. The issue is not whether Kant would accept RU. There is no reason to doubt he would. But there is a gap between what Kant actually says and the interpretation of it as an endorsement of RU. Second, it is *clearly* fallacious to identify RU with the Formula of Universal Law, or to hold that the latter is entailed by the former. We have no difficulty at all in demonstrating that this is a fallacy (section 1.4). Given Kant's status as a philosopher, to accuse him of what is a simpleminded error defies credibility here, especially when there is no compelling textual reason to attribute the error to him. One version of the traditional interpretation is relatively easy to dismiss.

4.5 The Challenge Posed by Aune's Version of the Traditional Reading

The other version we sketched (section 1.4), however, poses a greater challenge. According to this version, which has been developed by Aune (and recently reaffirmed in its essentials by Allen Wood, among others), Kant

argues for the principle L: "Conform your actions to universal law." Kant then jumps without argument to the Formula of Universal Law. He simply assumes that an agent abides by L just when he abides by the Formula of Universal Law – that is, he conforms to universal law just when he acts on a maxim that he can at the same time will to be a universal law. But this assumption is highly questionable. In this context, a universal law is a practical principle that is binding on all of us. Why would this universal law have to be the Formula of Universal Law, instead of, say, a principle prescribing us to maximally promote the general welfare or the perfection of rational beings? According to Aune, Kant's argument contains a crucial gap.

Earlier I suggested that a plausible interpretation of *Groundwork* I must show why in Kant's view at least some main points in his discussion of the good will, duty, and moral worth are necessary if he is to succeed in his derivation of the Formula of Universal Law. Aune seems to be cognizant of this requirement. For he attempts to show how L emerges from a central line of argument in *Groundwork* I. Here is a slightly simplified sketch of Aune's account. *Groundwork* I contains an argument that L is the principle of a good will – the one that motivates morally valuable actions. In his first and second "propositions," Kant contends that all and only actions from duty have moral worth and that their worth does not stem from their effects but rather from their motive. Then, in his discussion of his third proposition, Kant suggests that all actions from duty are done from (the motive of) respect for law. In effect, he suggests that to act from duty is to be motivated to act by the notion that one's action conforms to universal law. So for Kant all morally worthy actions are motivated by the principle L, "Conform your actions to universal law." Kant is embracing L when, right before stating the Formula of Universal Law, he says, "nothing is left but the conformity of actions to universal law as such, which alone is to serve the will as its principle" (GMS 402). Since L is what motivates all morally worthy actions, L is the basic moral requirement.[17] In this defense of L, Kant's propositions seem to be necessary. If, for example, contrary to the second proposition, the moral worth of actions was merely a function of their effects, then Kant would not be able to claim with any credibility that L is what motivates all actions having moral worth. Actions done from other motives (e.g., sympathy) could presumably bring about good effects, and thus have moral worth. In sum, it would be unfair to dismiss Aune's version of the traditional interpretation on the grounds that it fails to show the relevance of key claims in *Groundwork* I to the derivation of the Formula of Universal Law.

But there are good reasons for rejecting Aune's reading of *Groundwork* I. As a first step to showing this, let me contrast the basic structure of Aune's reading with that of the one I propose, the criterial reading. According to Aune, Kant employs his discussion of the good will, duty, and moral worth to establish that, upon reflection, we recognize L as the basic moral requirement. Once L has been located, Kant makes no further appeal to

this discussion. Kant simply assumes that the only way we can conform to universal law is to conform to the Formula of Universal Law. On the criterial reading, Kant's argument unfolds differently. Kant employs his discussion of the good will, duty, and moral worth to develop criteria that, upon reflection, we see must be fulfilled by any viable candidate for the supreme principle of morality. These criteria supplement those with which Kant begins – the ones that are contained in his basic concept of the supreme principle of morality (see section 1.2). At least implicitly, Kant relies on the full set of these criteria to eliminate rivals to the Formula of Universal Law. Only this formula (and its equivalents), claims Kant, remain as viable candidates for meeting the full set. Whether or not Kant adequately defends this claim, the *Groundwork* I derivation contains no obvious gap between a practically uninformative principle and the Formula of Universal Law.

Of course, that the criterial reading has a different structure than Aune's reading does not entail that the former is superior to the latter. It is fair to maintain that we would show the criterial reading to be superior if we accomplished three tasks. First, we need to meet the requirement introduced in section 4.3 by explaining why Kant might view his main discussions in *Groundwork* I to be necessary for his derivation of the Formula of Universal Law. Second, we need to offer a plausible alternative interpretation of Kant's murky suggestion that "nothing is left but the conformity of actions to universal law as such, which alone is to serve the will as its principle" – that is, an alternative to Aune's reading of it as an endorsement of L. Finally, we must show that on the criterial reading, Kant's argument is philosophically more powerful and interesting than on Aune's construal. I hope to attain each of these aims in the course of this chapter and those which follow.

4.6 From Duty and Moral Worth to Two Criteria for the Supreme Principle of Morality

According to the criterial reading, it is through his discussion of the good will, duty, and moral worth that Kant pinpoints criteria that the supreme principle of morality must fulfill. He then relies on these criteria to eliminate all candidates for the supreme principle of morality except the Formula of Universal Law (and its equivalents). Kant's argument by elimination is the focus of Chapter 7. Here I would like to defend the view that through his exploration of ordinary moral views Kant is indeed developing criteria for the supreme principle of morality.

To begin, let us look back in the text from the point at which Kant initially formulates the Formula of Universal Law. In the preceding sentence, Kant asks: "But what kind of law can that be, the representation of which must determine the will . . . so that the will can be called good absolutely and without qualification?" (GMS 402). Kant is here supposing that the supreme

principle of morality (the law) must meet a certain condition, and then asking which principle could meet it. Here is the condition. The supreme principle of morality (the law) must be such that the will is good without qualification if and only if it is determined by this principle. So for Kant we cannot hold a principle to be the supreme principle of morality (the law) unless we can hold that a will is good without qualification just in case it is determined by the principle.

This condition as stated at GMS 402 actually crystallizes criteria for the supreme principle of morality that Kant implicitly embraces in section I. Meeting the condition involves meeting *at least* two criteria, both of which concern duty and moral worth. To show this, I will make a very brief pass through Kant's difficult and controversial account of duty and moral worth. My aim is not to evaluate the account, or even to clarify its details (these tasks are left to Chapters 6 and 5 respectively), but merely to show that in it Kant suggests two criteria that any viable candidate for the supreme principle of morality must meet.

According to Kant's condition, the supreme principle of morality (the law) must be such that the will is good without qualification if and only if it is determined by this principle. By "the will" here I take Kant to be referring to an instance of willing. But when is willing unconditionally good? Kant answers this question as it applies to the willing of rational agents like humans who, unlike other (possibly extant) agents such as God and angels, can be tempted by their inclinations to act immorally.[18] Kant suggests that willing is unconditionally good if and only if it is done from duty. All and only actions from duty have moral worth, which is unconditional worth.[19] This is widely taken to be Kant's "first proposition," which he implies, but does not state, in *Groundwork* I (GMS 397–399).[20] Kant takes this point to yield a further, related, one (GMS 399). Given that acting from duty is unconditionally valuable, its value cannot stem from its producing certain effects. For if it stemmed from this source, then there would be contexts in which it was not valuable, namely those in which the action did not produce these effects. Thus Kant intimates in his "second proposition" that the moral worth of actions from duty stems not from their effects but rather from the principle of volition on which they have been done (GMS 399–400).

But what kind of principle is such that acting on it has moral worth? Kant's "third proposition" is that: "*Duty is the necessity of an action from respect for law*" (GMS 400). In his discussion of this proposition, Kant suggests that an action has moral worth if and only if it is determined by the law (the supreme principle of morality).[21] Acting from duty involves conforming to the supreme principle of morality because this principle requires that one conform to it. In sum, Kant suggests that the supreme principle of morality must be such that willing is good without qualification if and only if it is determined by this principle. For us, willing (or, equivalently, acting) is unconditionally valuable just in case it is done from duty. (Here Kant is

employing what I earlier called his "particular action" account of the good will (section 3.7).) The value of action from duty does not lie in its effects but rather in its grounds. When we act from duty, we act because the supreme principle of morality (the law) requires it. In this sense, the supreme principle of morality determines our action.

I claimed that in Kant's view to meet the condition implicit at GMS 402, a principle would have to meet at least two criteria. We can now see what those criteria are. First, the supreme principle of morality must be such that all and only actions conforming to it because the principle requires it – that is, all and only actions done "from duty" – have moral worth. We cannot (rationally speaking) hold a principle to be the supreme principle of morality unless we can maintain that it fulfills this criterion. Second, the supreme principle of morality must be such that the moral worth of actions conforming to it "from duty" stems from the actions' motive – that is, the principle on which they are performed – rather than from their effects. Kant suggests that the first criterion entails the second. If the moral worth of actions done from duty stemmed from their effects, then, contrary to the first criterion, some actions done from duty would not have moral worth, namely actions that failed to have a certain effect.

As it stands, these criteria are quite abstract. In Chapter 5, we probe what they mean and how Kant defends them. It should now be apparent, however, that we may plausibly interpret Kant to be developing criteria for the supreme principle of morality in *Groundwork* I. The criterial interpretation will meet the first requirement for success sketched earlier if we can, in addition, show that Kant needs to appeal to these criteria to eliminate rivals to the Formula of Universal Law. Chapter 7 focuses on this task.

4.7 Law as Motive: A Third Criterion for the Supreme Principle of Morality

For now, let us turn to the second requirement for success. What might Kant mean when he says that "nothing is left but the conformity of actions to universal law as such, which alone is to serve the will as its principle"? Is there a plausible alternative to the notion that he is embracing L, the imperative "Conform your actions to universal law"? I believe that there is. However, in fairness to Aune, we should note that the obscurity of Kant's remarks here renders it very difficult to arrive at a definitive interpretation. Kant writes:

> But what kind of law can that be, the representation of which must determine the will, even without regard for the effect expected from it, in order for the will to be called good absolutely and without limitation? Since I have deprived the will of every impulse that could arise for it from obeying some law, nothing is left but the

conformity of actions to universal law as such, which alone is to serve the will as its principle. (GMS 402)

In the first sentence, Kant implicitly invokes criteria, which he developed earlier in *Groundwork* I, for our accepting a principle as the supreme principle of morality. We stated two of these criteria in the preceding section. I now suggest that in the second sentence, before introducing the Formula of Universal Law, Kant implicitly invokes another criterion. This criterion for the supreme principle of morality has to do with a motive that must be available to us for conforming to it.

In the sentence immediately preceding the cited passage, Kant distinguishes between two basic ways we can be motivated to conform to a principle. Kant contrasts cases in which the representation of a principle in itself constitutes the determining ground of the will from ones in which some expected effect constitutes this ground. Moreover, he suggests that only cases of the former sort are cases of good willing. Thus he says "nothing other than the *representation of the law* in itself . . . insofar as it and not the hoped-for effect is the determining ground of the will, can constitute the preeminent good we call moral, which is already present in the person himself who acts on this representation" (GMS 401).

Returning to GMS 402, Kant is concerned with the kind of will that is absolutely good. He specifies here that what determines absolutely good willing is not the effects one expects to result from the willing. Kant has "deprived the will of every impulse that could arise for it from obeying some law" in the sense that, in his view, he has shown that absolutely good willing is not at all motivated by some "impulse" (e.g., the sensation of pleasure) that one believes will result from obeying some principle.[22] It is rather motivated by the representation of the law. What determines absolutely good willing is the "conformity of actions to universal law as such." In other words, the motive for absolutely good willing is the notion that it conforms to a universally and unconditionally binding practical principle. This principle is, of course, the supreme principle of morality. We can now see the criterion that Kant is implicitly invoking in the second sentence. The supreme principle of morality must be such that our representing it as a law, that is, a universally and unconditionally binding principle, gives us a sufficient motive to conform to it. If this reading is on target, then Kant is not suggesting here that we must embrace the imperative "Conform your actions to universal law," but rather invoking yet another criterion that he takes himself to have established earlier in his discussion. Whatever the supreme principle of morality is, implies this criterion, we must (rationally speaking) be able to hold that our having sufficient motive to adhere to it does not depend on any effect we expect from doing so.

The notion that Kant indeed takes this as a criterion gains support from a distinction he makes between "material" and "formal" principles. Kant

introduces this distinction in his discussion of the "second proposition" in *Groundwork* I:

> For, the will stands between its a priori principle, which is formal, and its a posteriori incentive, which is material, as at a crossroads; and since it must still be determined by something, it must be determined by the formal principle of volition as such when an action is done from duty, where every material principle has been withdrawn from it.[23] (GMS 400)

Kant holds that when an agent acts from duty, her will is determined by the supreme principle of morality. Kant here says, moreover, that when an agent acts from duty, her will is determined by a formal, rather than a material, principle. So Kant is implying that the supreme principle of morality must be a formal, rather than a material, principle. Kant does not discuss in this passage precisely what a formal principle is. Yet he does suggest that, in one sense, a formal principle is one that determines the will even though "every material principle has been withdrawn." What does this mean? In light of Kant's most thorough discussion of material practical principles, namely the one he conducts in the second *Critique*, we will be able to see that it means the following: a formal principle is a rule such that our representing it as a law governing our actions gives us sufficient motive to conform to it.

Let us turn to the Analytic of the *Critique of Practical Reason*. In Theorem I, Kant claims: "All practical principles that presuppose an *object* (matter) of the capacity of desire as the determining ground of the will are, without exception, empirical and can furnish no practical laws" (KpV 21). No material practical principle, says Theorem I, can be a practical law. Since the supreme principle of morality would have to be a practical law, this theorem (if true) entails that no material practical principle can be the supreme principle of morality. According to Kant, a material practical principle presupposes an object of the capacity of desire as the determining ground of the will. In other words, a material practical principle is a rule that an agent has sufficient motive to adhere to only on condition that, in his view, doing so will enable him to realize some object he desires (section 1.8). Take the rule: "In order to visit Grant's tomb, you ought to travel to New York." To say that it is a material practical principle is to say (in part) that an agent's having sufficient motive to act on it (i.e., to travel to New York) is contingent on his belief that doing so will enable him to realize some object he desires (i.e., his visiting Grant's tomb).[24]

Now, obviously, for Kant formal principles are not material. Whatever else it might be, a formal principle is a rule such that an agent's having sufficient motive to adhere to it does not depend on his expecting that doing so will enable him to realize some object he desires. Nevertheless, according to Kant if a principle is a formal one, an agent does have sufficient motive to adhere to it. What is this motive? Later, under Theorem III, Kant states that "all that remains of a law if one separates from it everything material, that is,

every object of the will (as its determining ground), is the mere *form* of giving universal law" (KpV 27; see also KpV 24). Since a formal principle counts as one from which "every object of the will (as its determining ground)" has been separated, this statement suggests that what serves as a determining ground for abiding by such a principle is "the mere *form* of giving universal law." In other words, what serves as a motive for abiding by a formal principle is the representation of it as a law. So, in one sense, a formal principle is a practical rule such that our representing it as a law gives us sufficient motive to conform to it. (I have not tried to give an exhaustive account of Kant's distinction between formal and material principles, but to pinpoint one way in which Kant draws such a distinction.)[25]

In sum, no material principle, Kant claims, can be the supreme principle of morality. The supreme principle would have to fulfill a criterion of formality; it would have to be such that an agent's representing it as a law – that is, a universally and unconditionally binding principle – gives him sufficient motive to conform to it. Kant offers this criterion in the *Groundwork* and develops it at greater length in the second *Critique*.

We could not show the criterial reading of *Groundwork* I to be superior to Aune's reading unless we could plausibly interpret Kant's assertion that "nothing is left but the conformity of actions to universal law as such, which alone is to serve the will as its principle" in some way other than as an endorsement of the imperative "Conform your actions to universal law." We have found that there is another plausible interpretation of this assertion, namely that it amounts to a statement of one criterion that in Kant's view the supreme principle of morality must meet: this principle must be such that our representing it to ourselves as a law governing our action gives us a sufficient motive for us to conform to it.

On Aune's reading, Kant sees his discussion of the good will, duty, and moral worth as necessary for him to establish the imperative "Conform your actions to universal law." After establishing this imperative, Kant thinks (wrongly) that he can move immediately to the Formula of Universal Law. There is an obvious gap in Kant's reasoning. On the criterial reading, in contrast, Kant sees his discussion of the good will, duty, and moral worth as necessary for him to establish criteria for the supreme principle of morality. The supreme principle of morality must be such that: (1) all and only actions conforming to this principle because the principle requires it – that is, all and only actions done from duty – have moral worth; (2) the moral worth of conforming to this principle from duty stems from its motive, not from its effects; (3) an agent's representing this principle as a law (i.e., a universally and unconditionally binding principle) gives him sufficient motive to conform to it. Kant crystallizes these criteria at GMS 402, right before he sets out the Formula of Universal Law. In effect, Kant holds that only this formula (and equivalent ones) remain as viable candidates for fulfilling each of the criteria he develops for the supreme principle of morality. (The

criteria Kant develops are, I believe, the three just mentioned, those implicit in his basic concept of the supreme principle of morality, and one other I discuss later, in section 4.9.) If Kant's argument fails, it is not because it contains an obvious gap between a practically uninformative principle and the supreme principle of morality.

We have already gone half of the way toward showing the criterial reading to be superior to Aune's version of the traditional reading. We have seen that there is a plausible alternative to Aune's reading of Kant's claim that "nothing is left but the conformity of actions . . . " We have seen that Kant holds his discussions of the good will, duty, and moral worth to be necessary for the development of criteria for the supreme principle of morality. All we need to do to show that he could reasonably hold these discussions to be necessary for the derivation is to establish how he might use the criteria to eliminate various candidates for this principle. Chapter 7 focuses on eliminating these rivals as well as showing that the criterial interpretation makes Kant's argument far more forceful and philosophically interesting than it is on the traditional interpretation.

4.8 The Criterial Reading and *Groundwork* II

Before developing the criterial reading any further, however, I must attend to a worry that readers might have. This reading has been presented as an alternative to the traditional interpretation of the *Groundwork* derivation of the Formula of Universal Law. The main proponent of the traditional interpretation, Aune, bases his contention that this derivation contains a crucial gap on examination of *Groundwork* I. The criterial reading also focuses on *Groundwork* I. To be successful the reading must be consistent with what Kant actually says in the sentences preceding his first statement of the Formula of Universal Law. Yet when I initially sketched the traditional reading of the *Groundwork* derivation of the Formula of Universal Law, I cited not only Kant's argument in *Groundwork* I (culminating at GMS 402) but also a parallel argument in *Groundwork* II (culminating at GMS 420–421). The worry is that, although the criterial reading might constitute a viable alternative to the traditional one in light of *Groundwork* I alone, when we take into consideration *Groundwork* II as well we find that only the traditional reading is permitted by the text. In *Groundwork* II, Kant says that he is going to "inquire whether the mere concept of a categorical imperative" can provide him with a principle that is alone suited to be a categorical imperative (GMS 420). He then says:

> When I think of a *hypothetical* imperative in general I do not know beforehand what it will contain; I do not know this until I am given the condition. But when I think of a *categorical* imperative, I know at once what it contains. For since the imperative contains, beyond the law, only the necessity that the maxim be in conformity with

this law, while the law contains no condition to which it would be limited, nothing is left with which the maxim of action is to conform but the universality of a law as such; and this conformity alone is what the imperative properly represents as necessary.

There is, therefore, only a single categorical imperative and it is this: *act only on that maxim through which you can at the same time will that it become a universal law.* (GMS 420–421)

It appears that Kant might be making just that unacceptable move that Aune claims he makes in *Groundwork* I. Does Kant not here jump without argument from the notion that the fundamental moral requirement is to conform your actions to universal law to the conclusion that the only way to adhere to this requirement is to conform to the Formula of Universal Law? Does Kant not here take an illicit step from the notion that, by virtue of its very concept, a categorical imperative commands conformity to law to the further notion that it commands that you act only on maxims that you can at the same time will to become universal laws?

This worry is reasonable, and I have no quick and easy response to it. However, several considerations show at least that the worry is not as serious as it might initially seem to be.

4.9 Coherence with Ordinary Moral Reason: A Fourth Criterion

To begin, Kant does not take the derivation of the Formula of Universal Law to end with his setting out of the formula. The text continues: "Now, if all imperatives of duty can be derived from this single imperative as from their principle, then, even though we leave it undecided whether what is called duty is not as such an empty concept, we shall at least be able to show what we think by it and what the concept wants to say" (GMS 420–421). The derivation is not complete unless "all imperatives of duty" can be derived from the imperative Kant proposes as the only viable candidate for the supreme principle of morality. By "all imperatives of duty," Kant apparently means all imperatives that we, reflective rational agents, take to express our moral duties. Kant proceeds, of course, to try to show that four such imperatives (e.g., a requirement not to make false promises for financial gain) follow from the Formula of Universal Law. He then says: "These are a few of the many actual duties, *or at least of what we take to be such*, whose derivation from the one principle cited above is clear" (GMS 423–424, emphasis added). If these duties' derivation from the Formula of Universal Law were not clear – for example, if it simply did not follow from the formula that we had them – then, Kant implies, we could not accept this formula as the only viable candidate for the supreme principle of morality. In the short paragraph (GMS 420–421) following his statement of the Formula of Universal Law, Kant not only emphasizes that he has not (yet) established (i.e., given a deduction for) this formula but also indicates an important criterion for any viable candidate for

the supreme principle of morality. We must be able to see how it follows from this candidate that if it were established, we would indeed have moral duties that we are convinced we do have. So, despite initial appearances, Kant is not guilty of moving immediately, without argument, from the notion that the fundamental moral requirement is to conform your actions to universal law to the conclusion that the only way to adhere to this requirement is to conform to the Formula of Universal Law. His transition is, at the very least, mediated by consideration of whether the Formula of Universal Law could generate duties that conform to what we take our moral duties to be.

This sort of consideration is found not only in Kant's *Groundwork* II derivation of the Formula of Universal Law but elsewhere as well. Consider Kant's derivation of the Formula of Humanity. Before stating the formula, he says that "it must be possible to derive all laws of the will" (GMS 429) from the "supreme practical principle" (GMS 428). After stating the formula, he says: "we shall see whether this can be carried out" (GMS 429), implying that if the Formula of Humanity is to be a viable candidate for the supreme principle of morality, we had better see that we can derive all laws of the will from it. Granted, in *Groundwork* I Kant does not explicitly make it a condition of success of his derivation of the Formula of Universal Law that this principle generate moral prescriptions we take ourselves to be bound by. Immediately after setting out the principle, however, Kant turns to supporting the view that common human reason "agrees completely with this in its practical appraisals" (GMS 402). And it is only after Kant supports this view that he claims that "we have arrived, within the moral cognition of common human reason, at its principle" (GMS 403). In light of the evidence we have seen in the other derivations, it seems reasonable to conclude that here as well he holds that for his argument to be successful, the principle he selects must (if it is valid) generate moral requirements that cohere with those we pretheoretically believe ourselves to have. Therefore, I will take it that for Kant any viable candidate for the supreme principle of morality must be such that, if valid, it would generate a plausible set of duties, where plausibility is to be assessed in relation to ordinary moral thinking. In brief, the supreme principle of morality must be such that a plausible set of duties (relative to ordinary moral consciousness) can be derived from it.

Those friendly to the idea that Kant embraces this criterion might wonder why the criterion does not belong to Kant's basic concept of the supreme principle of morality. The term "basic concept" is mine, not Kant's. In my usage, Kant's basic concept contains only those criteria that Kant employs from the very outset of the *Groundwork*, namely its Preface (section I.2). I do not find evidence in the Preface that Kant holds that any viable candidate for the supreme principle of morality must generate moral prescriptions acceptable to ordinary moral consciousness. The evidence emerges later, in the places just highlighted. A philosopher might, however, insist that the criterion is obviously implicit in the notion of a supreme principle of *morality*.

If a practical principle generated duties that clashed dramatically with what we take our moral duties to be, then there would be no sense in which it could be a principle of morality, the philosopher might insist. In reply, I think Kant himself suggests a sense in which such a principle could conceivably be a principle of morality. It could conceivably be a categorical imperative, the kind of imperative Kant himself associates with morality (see GMS 416). The principle could set out unconditional practical requirements, that is, specify (correctly) what all of us are obligated to do, regardless of whether we have an inclination to do it (or even regardless of whether we *believe* that we have a duty *not* to do it). In the end, what is important is not that we agree on the precise point in his argument that Kant embraces the criterion in question but rather that we realize that Kant does indeed embrace it.

4.10 The Apriority of the Supreme Principle of Morality

Before moving on to other reasons for rejecting the notion that, in light of *Groundwork* II, the traditional interpretation of the derivation is the only plausible one, we need to address an issued raised by the first reason. It might seem puzzling that in Kant's view a criterion any viable candidate for the supreme principle of morality must fulfill is that of being capable of generating duties that cohere with the moral duties we take ourselves to have. Does not whether we conclude that a given principle meets this criterion rest on experience, that is, our particular experience of morality, and does not Kant insist that the supreme principle be an a priori one? Already in the *Groundwork* Preface, Kant says that the ground of an obligation to conform to the supreme principle of morality must be sought "a priori simply in concepts of pure reason" and that any principle that "rests in the least part on empirical grounds, perhaps only in terms of a motive, can indeed be called a practical rule but never a moral law" (GMS 389).

To determine how much, if any, real tension exists in Kant's view, we need to understand two senses in which according to him the supreme principle must be an a priori rather than an empirical principle. It must be a priori in both (what I call) a motivational sense and an epistemological sense.

Beginning with the former, the supreme principle of morality must be such that all rational agents always have available to them a sufficient motive for abiding by it. (Whether they actually act on this motive or some other one, such as an inclination, is another question.) But that means that their having sufficient motive available to them to conform to the principle must not depend on anything empirical – that is, on their particular inclinations or even on their nature, insofar as this nature is not necessarily shared with all rational agents (KrV A 806–807/B 834–835). A principle is a priori in the motivational sense just in case any rational agent's having available to him a sufficient motive for abiding by it is not conditional on anything empirical. A principle would be empirical in case a rational agent's having sufficient

motive to abide by it was, for example, conditional on his expectation that abiding by it would give him pleasure (KpV 9, note). The requirement that the supreme principle of morality be a priori in the motivational sense entails that it cannot be a material principle. It will become apparent why, precisely, Kant thinks the supreme principle of morality must be a priori in the motivational sense in section 5.7 when we examine his arguments for his third criterion for the supreme principle of morality (introduced in section 4.7), namely the criterion according to which this principle must be such that an agent's representing it as a law gives him sufficient incentive to conform to it.

At any rate, Kant's appealing to experience in his derivation of the Formula of Universal Law does not seem incompatible with all rational agents having an empirically unconditioned motive at their disposal for abiding by this formula. That we rely on our moral experience in pinpointing the supreme principle of morality does not, for example, seem to entail that our having at our disposal sufficient motive to comply with it is conditional on our expectation that doing so will get us something we want.

The second sense in which, according to Kant, the supreme principle of morality must be a priori is what I call the epistemological sense. Kant states that a practical law, and thus the supreme principle of morality, must be knowable a priori (see GMS 425–426 and KpV 26). In the *Critique of Pure Reason*, Kant defines a priori knowledge as "knowledge *absolutely* independent of all experience" (KrV B 2–3). If we had a priori knowledge of a practical principle, that is, knowledge that it was valid, this knowledge would have to be "absolutely independent" of all experience in the following sense: it would have to be *grounded* or *legitimated* without appeal to any particular set of experiences.[26]

Why does Kant claim that a practical law must be knowable a priori? According to him, a practical principle could be a practical law only if it were unconditionally and universally valid, thus admitting of no possible exception. But, in Kant's view, if a principle can be justified only by appeal to particular experiences, then it cannot be known that no exception to it is possible.[27] That experience has thus far shown that there is no exception to a principle fails to entail that there will be none. To bring the point to the issue at hand, that experience has thus far shown that a given principle generates all the duties we take ourselves to have does not entail that the principle will always generate all these duties. For it to be known that there can be no exception to a principle, the principle's validity must be grounded a priori.

Does this apriority requirement clash with Kant's view that the derivation of the Formula of Universal Law requires an appeal to experience? In section 1.3 we discussed Kant's distinction between the derivation of the supreme principle of morality and its deduction. A successful derivation would show that if there is a supreme principle of morality, then it is a certain principle. A deduction would establish that this principle is universally and

unconditionally binding. In *Groundwork* III Kant offers a deduction of the Formula of Universal Law. (Strictly speaking, he offers a deduction of a principle that resembles this formula and that he takes to be equivalent to it. But this point is not important to the present discussion.)[28] Obviously, in Kant's view for this deduction to succeed, it cannot be grounded in any appeal to experience. But if the deduction depended on the success of Kant's derivation of the Formula of Universal Law, then it would, at least partly, be grounded in such an appeal. For, as we noted, in Kant's view the derivation itself could succeed only if the principle it yielded cohered with our moral experience. Therefore, perhaps Kant's considered view in the *Groundwork* is the following: the *Groundwork* III deduction establishes the validity of the Formula of Universal Law, and this deduction relies not at all on appeals to particular experiences. That this formula is universally and unconditionally binding can be demonstrated a priori. However, what cannot be demonstrated a priori is that we are to think of this principle *as* the supreme principle of morality. For whether we think of it as such depends on the principle's fulfilling an empirical criterion. The principle must be such that (if valid) it would generate a set of duties that would cohere largely with the set we, upon reflection, take ourselves to have. In short, I am suggesting that on Kant's considered view, the deduction of the Formula of Universal Law does not presuppose the success of its derivation.

Of course, I would need to do much more, including a close reading of the deduction, to defend this suggestion. Since my main concern here is the derivation, I hope I will be permitted to stop at suggesting a way in which to accommodate Kant's appeal to experience in the derivation with his apriority condition for a deduction.

In any case, we need to keep in view that not only in the derivation of the Formula of Universal Law, but in that of the Formula of Humanity as well, Kant suggests that unless a principle generates moral prescriptions that accord with those we take ourselves to be bound by, we cannot accept this principle as the only viable candidate for the supreme principle of morality.

4.11 Rejecting the Traditional Interpretation of the *Groundwork* II Derivation

The traditional interpretation of the *Groundwork* II derivation has Kant move directly from the notion that, if there is a supreme principle of morality, we ought to conform to universal law, to the further notion that, if there is such a principle, it is the Formula of Universal Law. But Kant does not move directly from the former notion to the latter. At the very least, he makes his transition conditional on the Formula of Universal Law's ability to fulfill the criterion we have discussed, namely that of generating a plausible set of duties.

There are other grounds for rejecting the traditional interpretation. Granted, if we focus exclusively on the paragraph that spans from GMS 420

to 421, we, indeed, get the impression that Kant jumps without argument from the concept of a categorical imperative simply as an unconditionally and universally binding requirement (a practical law) to the Formula of Universal Law as the only principle that could realize this concept. And such a jump would indeed be problematic, since the concept of such a requirement could be realized in many principles, not just the Formula of Universal Law. Misleading though Kant's presentation might be, however, we need not interpret him to be operating *here* with such a thin concept of a categorical imperative.

Just a few pages before he makes the argument in question, Kant distinguishes between hypothetical and categorical imperatives. Regarding the latter he writes:

> Finally there is one imperative that, without being based upon and having as its condition any other purpose to be attained by certain conduct, commands this conduct immediately. This imperative is *categorical*. It has to do not with the matter of the action and what is to result from it, but with the form and the principle from which the action itself follows; and the essentially good in the action consists in the disposition, let the result be what it may. This imperative may be called the imperative *of morality*. (GMS 416)

In this passage, Kant is obviously explaining his concept of a categorical imperative (as the imperative of morality). And this concept is thicker than that of an unconditionally and universally binding principle. Kant here suggests that a categorical imperative (in the relevant sense) must fulfill each of the criteria for the supreme principle of morality we discovered in our discussion of *Groundwork* I. First, Kant writes about "the essentially good in the action." In which action? Given his discussion in *Groundwork* I, he must be referring to the essential goodness of action from duty. So, Kant here implies, a categorical imperative (as the imperative of morality) must be such that conforming to it because the imperative requires it has moral value. Second, Kant maintains that a categorical imperative in the relevant sense must be such that when conforming to it has value – that is, when such conformity is from duty – this value stems from the principle on which one acts, "let the result be what it may." Third, Kant here hints at his distinction between material and formal principles. It seems plausible to construe his rather vague statement that a categorical imperative "has to do" not with the matter of an action but with its form to be an expression of his view that a categorical imperative must not have any material condition. It must rather be such that an agent's representing it to himself as a law provides him with sufficient incentive for conforming to it. In short, the passage supports what is actually an unsurprising conclusion: Kant's concept of a categorical imperative (as the imperative of morality) echoes his concept of the supreme principle of morality in *Groundwork* I – not merely his basic concept, but the thicker one he develops in his discussion of the three propositions. Although at

GMS 420–421 Kant does not say so explicitly, it seems reasonable to assume that he has this thicker concept in view.

If this is correct, then despite appearances we need not interpret Kant to jump from the thin concept of a categorical imperative as an unconditionally and universally binding principle (a practical law) to the Formula of Universal Law. We may read Kant's argument at GMS 420–421 to be elliptical. Kant moves from a thick concept of a categorical imperative to the notion that this concept could be actualized only in the Formula of Universal Law (or equivalent principles). In defense of this move, Kant suggests the argument that no other principle could meet each of the criteria he has established for the supreme principle of morality. (Much of the remainder of this book focuses on understanding and evaluating this argument.) Since the *Groundwork* II derivation of the Formula of Universal Law admits of a criterial reading, it does not cast doubt on the criterial reading of the *Groundwork* I derivation of this formula. It is at least worth a try to see if on the criterial reading Kant's derivation is more philosophically powerful and engaging than it is on a reading according to which it contains a devastating gap.

Even for a reader who remains convinced that the traditional interpretation accurately reflects Kant's intentions in the *Groundwork* derivation of the Formula of Universal Law, all is not lost. It is open to such a reader to take the criterial reading of the derivation as a reconstruction of Kant's argument. Whether or not the reader agrees that Kant employs his criteria for the supreme principle of morality in his derivations of it, it is clear that he does indeed embrace and, in some cases, defend the criteria themselves. The criterial reconstruction (if that is how one sees it) of Kant's derivation uses materials that Kant himself provides.

4.12 Summary

We have covered a lot of ground in this chapter, so it might be helpful to pause here to get our bearings. I have introduced a criterial reading of Kant's *Groundwork* I derivation of the Formula of Universal Law. On this reading, the derivation has three main steps. First, Kant sets out criteria that any viable candidate for the supreme principle of morality must fulfill. These criteria include, but are not limited to, those which belong to his basic concept of this principle. Second, Kant tries to show that no (possible) rival to the Formula of Universal Law remains a viable candidate for fulfilling all of the criteria. Finally, Kant attempts to demonstrate that the Formula of Universal Law does remain a viable candidate for fulfilling all of them. Therefore, if there is a supreme principle of morality, then it is this formula.

In this chapter, I hope to have accomplished two main goals. The first was to show that there is room for a new approach to Kant's *Groundwork*

derivation of the Formula of Universal Law. Korsgaard's interpretation has serious textual and philosophical shortcomings (section 4.2), and one version of the traditional reading clearly fails (4.4). The other version of the traditional reading, that defended by Bruce Aune, presents greater difficulties. However, I have, I hope, shown that the text of Kant's derivations in both *Groundwork* I and II permit the criterial reading. Whether this reading ultimately prevails is largely a question of whether it renders Kant's derivation more philosophically powerful and interesting than does Aune's construal. To see whether it does, we need to probe the plausibility both of the criteria Kant offers for the supreme principle of morality, and of his argument that, upon reflection, no principle besides the Formula of Universal Law (or something equivalent) remains as a viable candidate for meeting all of them. The chapters that follow do just that.

The second main goal of this chapter has been to sketch a preliminary account of the derivation's first step. I have located in Kant's text several criteria for the supreme principle of morality in addition to those belonging to his basic concept of it. It will be helpful to have them in view. The supreme principle of morality must be such that: (1) all and only actions conforming to this principle because the principle requires it – that is, all and only actions done from duty – have moral worth; (2) the moral worth of conforming to this principle from duty stems from its motive, not from its effects; (3) an agent's representing this principle as a law, that is, a universally and unconditionally binding principle, gives him sufficient incentive to conform to it; (4) a plausible set of duties (relative to ordinary rational knowledge of morals) can be derived from this principle. At this point, the criteria might seem somewhat vague, unmotivated, and disjointed. Chapter 5 attempts to show in detail what these criteria mean, how Kant defends them, and how they relate to one another. Chapter 6 probes whether (or to what extent) we should accept criterion 1. This criterion is obviously controversial. It is also crucial to Kant's derivation. As I hope becomes apparent in Chapter 7, this criterion (or, more precisely, one component of it) serves as the ultimate basis for a strong Kantian argument against many consequentialist candidates for the supreme principle of morality.

5

Criteria for the Supreme Principle of Morality

5.1 Plan of Discussion: Focus on First Criterion

If the argument of Chapter 4 has been successful, then it is apparent that in the *Groundwork*, from the Preface all the way up to the statement of the Formula of Universal Law in Section II, Kant develops criteria that the supreme principle of morality must fulfill. According to his basic concept, already implicit in the Preface, this principle must be practical, absolutely necessary, binding on all rational agents, and serve as the supreme norm for the moral evaluation of action (section I.2). Later in the *Groundwork* Kant develops four additional criteria (Chapter 4). The supreme principle of morality must be such that: (1) all and only actions conforming to this principle because the principle requires it – that is, all and only actions done from duty – have moral worth; (2) the moral worth of conforming to this principle from duty stems from its motive, not from its effects; (3) an agent's representing this principle as a law, that is, a universally and unconditionally binding principle, gives him sufficient incentive to conform to it; (4) a plausible set of duties (relative to ordinary rational knowledge of morals) can be derived from this principle.

Until Chapter 8, I have little more to say about the fourth criterion. Kant appeals to ordinary rational knowledge of morals in developing criteria 1–3. I suspect that part of the reason he introduces criterion 4 is because he makes this appeal. He recognizes that it would be intolerably odd to base criteria for the supreme principle of morality on ordinary moral consciousness, yet to champion a principle that clashed dramatically with this consciousness as the only viable candidate for the supreme principle of morality. After all, if he were prepared to dismiss completely the judgment of commonsense moral reason regarding which moral duties we have, then what grounds would he have to rely on it in developing criteria for the supreme principle of morality?

This chapter focuses on the first three criteria for the supreme principle of morality that Kant develops in addition to those contained in his basic

concept of this principle. How, precisely, are we to understand these criteria, and what are Kant's arguments for them? The bulk of the chapter (sections 5.2–5) explores what is perhaps Kant's most controversial criterion, namely the first, while sections 5.6–7 investigate the second and third criteria respectively.[1] At several points in the chapter, I comment on the relations that obtain between the criteria.

5.2 Moral Worth and Actions Contrary to Duty

According to what is perhaps Kant's most important and controversial criterion, the supreme principle of morality must be such that all and only actions conforming to it because the principle requires it (i.e., all and only actions done from duty) have moral worth.

The first thing to note about the criterion is that, according to it, no action that *fails* to conform to the supreme principle of morality can have moral worth.[2] That this is indeed Kant's view in the *Groundwork* is not hard to see. Kant, of course, distinguishes between actions that are in accordance with duty [*pflichtmäßig*] and actions that are done from duty [*aus Pflicht*]. To perform an action that is in accordance with duty, that is, a morally permissible action, is to do something that violates no duty. For example, we presumably have a duty to deal honestly in financial transactions. Following Kant's discussion in *Groundwork* I, consider a shopkeeper who refrains from overcharging inexperienced customers. Whether he does so because he fears that his overcharging them might come to light and ruin his business or because it is required by moral principle not to overcharge them, he is acting in accordance with duty (GMS 397). Only in the latter case, however, is he acting from duty. Not all actions that are in accordance with duty (i.e., morally permissible) are from duty. In *Groundwork* II, Kant elaborates on his notion of what it means to act in accordance with duty. There it becomes clear that, in his view, whether an action complies with duty depends on its maxim. An agent's action complies with duty if and only if the maxim on which he does it accords with the Formula of Universal Law. In other words, the agent's action is in accordance with duty if and only if he can act on its maxim and at the same time will that it should become a universal law. A maxim such as "From self-love, I will give correct change to all of my customers in order to promote my business" accords with the Formula of Universal Law. Nevertheless, in acting on it an agent would not be acting from duty. According to Kant, not all actions done on maxims that pass the Formula of Universal Law test are done from duty.

On Kant's account, however, all actions done from duty are also done on maxims that pass this test; all actions done from duty are also in accordance with it. In his famous exploration of cases in *Groundwork* I, Kant is attempting

to elucidate the concept of a good will. With the help of the concept of duty, he is trying to clarify when we, imperfectly rational beings, perform actions that have intrinsic value – that is, actions that express good will. Ultimately, Kant aims to pinpoint the principle of a good will: the supreme principle of morality (section 4.3). At the beginning of his discussion, Kant tells us: "I here pass over all actions that are already recognized as contrary to duty, even though they may be useful for this or that purpose; for in their case the question whether they might have been done *from duty* never arises, since they even conflict with duty" (GMS 397). Why on Kant's view does the question never arise as to whether actions that conflict with duty can be done from duty? Kant offers no explicit explanation of this remark. Perhaps, according to him, it simply belongs to the concept of an action done from duty that it be done in accordance with it. It may be that Kant has chosen from the outset of the *Groundwork* to use the expression "from duty" to refer only to actions that one does because one believes they are right and that according to Kant's standard are indeed right.

There is, however, another interpretation of Kant's claim that the question never arises as to whether actions that conflict with duty can be done from duty. This interpretation seems to me to be more compelling because it reveals how remarks Kant makes elsewhere in the *Groundwork* might explain the claim. Consider Kant's emphasis in this work and elsewhere on how easy it is to determine what our duties are. Kant intimates that "cognizance of what every man is obligated to do" is available to each of us, "even the most ordinary" (GMS 404), and that what the supreme principle of morality commands "is plain of itself to everyone" (KpV 36). Perhaps, then, he reasons thus. The ultimate ground of an action done from duty is the agent's notion that the action is morally required. But it is very simple to figure out whether doing something is morally required. Therefore, if someone does something contrary to duty, he has obviously not been motivated by the notion that doing it was morally required. In short, Kant might hold an agent's duties to be so transparent to her that she just could not both be motivated by the notion that she is required to fulfill them yet violate them.[3] On either the interpretation that has Kant simply define actions from duty as in accordance with it, or the one (which I advocate) that highlights Kant's notion of the great ease with which one can determine her duties, Kant holds that no actions from duty are contrary to it.

One might, however, offer a very different reading of the passage at GMS 397. Kant asserts that the question does not arise at all as to whether actions already recognized as contrary to duty can be done from duty. In agreement with other interpreters, I take "already recognized" to mean already recognized by the reader – that is, by "objective" observers – to be contrary to duty.[4] But one might construe "already recognized" as contrary to duty to mean: believed by the agent himself to be contrary to it. On this construal,

Kant would not be implying that no action that conflicts with duty can be from duty. Rather, he would be intimating that if an agent believes an action to be contrary to duty, she cannot do it from duty.[5] Kant would be leaving open the possibility that an agent could, from duty, do something she took to agree with duty, but which actually conflicted with it.

The text fails to support this construal. Perhaps Kant does hold that if an agent believes an action to be contrary to duty, she cannot do it from duty.[6] Nevertheless, the evidence in the *Groundwork* indicates that Kant also embraced the notion that all actions done from duty are actually in accordance with it. As we noted, in examining cases Kant is elucidating the concept of a good will. He is trying to pinpoint when our actions are morally good – that is, when they have intrinsic, moral worth, which is the kind of worth characteristic of a good will. He makes the well-known suggestion that they have moral worth if and only if we do them from duty. Moreover, in the Preface to the *Groundwork*, Kant remarks: "[I]n the case of what is to be morally good, it is not enough that it *conform* with the moral law but it must also be done *for the sake of the law*" (GMS 390). In other words, for an action to have moral worth (be morally good), it must both be done from duty (for the sake of the law) and be in accordance with duty (conform with the moral law). Here Kant implies that if an action has moral worth, it is in accordance with duty. Since for Kant all actions done from duty have moral worth, it follows that all actions from duty are in accordance with it.[7]

As I have suggested, I suspect that in the *Groundwork* Kant has a simple reason for holding actions that are from duty (and thus have moral worth) to include only those that are in accordance with duty. On the view he there maintains, what duty requires is so transparent that any agent who genuinely acts from the notion that doing something is morally required will succeed in abiding by his duty. The Kant of the *Groundwork* did not, I venture, overlook the possibility of acting from duty yet contrary to it; rather, based on his conviction that it is very simple to determine what one's duty is, he rejected this possibility as practically irrelevant.

We have gone some way toward understanding Kant's criterion for the supreme principle of morality. We have seen why for Kant only actions that conform to duty can be done from duty, and we can thus comprehend why, in Kant's view, we cannot hold a principle to be the supreme principle of morality unless we can maintain that no actions that fail to conform to it can have moral worth.

5.3 Two Conditions on Acting from Duty

But we need to inquire further. The supreme principle of morality, says the criterion, must be such that all and only actions conforming to it because the principle requires it (i.e., all and only actions from duty) have moral worth. To clarify the criterion, we need to understand when an agent conforms

to a principle from duty. An agent does so, I suggest, only if each of two conditions is met.

According to the first condition, the agent's incentive for acting must stem from the notion that the principle is universally and unconditionally binding and that it requires the action. In actions from duty, asserts Kant, an agent's will is determined by the "*representation of the law* in itself," not by any of the action's "hoped-for effects" (GMS 401); and, he says, "an action from duty is to put aside entirely the influence of inclination" (GMS 400).[8] Brief discussion of the latter statement will help shed light on the former, enabling us to see that Kant embraces this first condition.

That for Kant inclination has no influence in an action done from duty strongly suggests that in his view only morally required actions can be done from duty.[9] For how in performing a morally permissible but not required action could one put aside entirely the influence of inclination?[10] As an example of such an action, imagine a typical case of someone's cutting his hair. No matter how morally reflective or concerned the person may be, in cutting his hair he would not be putting aside entirely the influence of inclination. There would be other morally permissible yet not required things he could do, for example, watch television or wash dishes. Moral grounds alone would not determine that he cut his hair rather than do one of these other things. As a basis for this choice, he must appeal to his inclinations – for example, his desire to be comfortable in this hot, humid weather. If the person is to act at all, he must have some incentive on the basis of which he chooses between the actions available to him. But for Kant, in morally permissible yet not required actions, this incentive could only be some inclination. Therefore, in Kant's view only morally required actions can be done from duty. Since Kant holds that an action has moral worth if and only if it is done from duty, he holds in effect that only morally required actions can have moral worth.

When Kant says at GMS 401 that in actions from duty an agent's will is determined by the representation of the law in itself, one might be tempted to take him to mean simply that in such actions, the agent's will is determined by the Formula of Universal Law. After all, Kant does often refer to this formula as "the law." But it is important to resist this temptation. At this point in the text, "the law" does not designate the Formula of Universal Law, for Kant has not yet derived this formula. That is what he is in the very midst of doing. What Kant means in this passage is that in actions from duty, an agent's will is determined by her representing a principle to herself *as* a law – that is, as unconditionally and universally binding. Since only actions an agent takes to be morally required can be done from duty, Kant is suggesting that, in actions from duty, an agent's will is determined by her notion that an unconditionally and universally binding principle requires these actions. For an agent to represent a principle as a law, she must be (or at some point have been) conscious of this principle, perhaps in a

rough-and-ready form.[11] For Kant, when we do something from duty, we derive our incentive to do it from the notion that doing it is required by a principle we represent as a law.

There is a further feature of Kant's view that I merely note in passing. Kant holds that in us, human beings, the representation of a principle as a law determines the will through the feeling of respect. Roughly, our representing a principle to ourselves as a law and being aware of what it requires produces in us a feeling of respect for it, a feeling that constitutes an incentive for conforming to the principle. Kant's statement that "an action from duty is to put aside entirely the influence of inclination" continues thus: "hence there is left for the will nothing that could determine it except objectively the *law* and subjectively *pure respect* for this practical law" (GMS 400). Kant's discussion of respect is very complex – he develops the concept in detail in the second *Critique* (see KpV 71–89) – and I do not pursue it here.[12] For our purposes, the important point is that in Kant's view an agent acts from duty only if her incentive for acting stems ultimately from her notion that her action is required by a practical law.

It is helpful to relate the understanding of acting from duty suggested by this first condition to contemporary discussion of related issues. Led by Barbara Herman, some philosophers attribute to Kant a distinction between acting from duty as a "primary motive" and acting from duty as a "limiting condition" (or, equivalently, "secondary motive").[13] Acting from duty as a primary motive involves meeting the first condition we mentioned for acting from duty. It occurs only when an agent's incentive for acting is the notion that the action is required by moral principle. In acting from duty as a limiting condition, however, an agent's will need not be determined by the notion that the action is morally required. An agent acts from duty as a limiting condition when his conduct is governed by a commitment to doing what duty requires. A person who cut his hair acted from duty as a limiting condition if the following was the case. Had cutting his hair been contrary to duty, then, *since it was contrary to duty*, the person would have refrained from doing it. To act from duty as a limiting condition, the person obviously need not have as an incentive the notion that his cutting his hair is morally required. When an agent acts from duty as a limiting condition, he need not have as his incentive the notion that he is morally obligated to act as he does. He is often not morally required to act as he does, for example, in a typical case of cutting one's hair. To use Herman's vocabulary, only actions done from duty as a primary motive have moral worth, in Kant's view.[14]

However, I do not employ this vocabulary myself. Kant would, I believe, recognize a distinction between acting from duty (as a "primary motive") and governing one's conduct by a commitment to do what moral principle requires. According to one of Kant's conceptions (what I call the whole character view in section 3.7), a human being has a good will by virtue of governing his conduct by a commitment to do what duty requires. And

given Kant's view of human nature, especially his conviction that human inclinations are often incentives for immoral action, a human being's good will will sometimes manifest itself in actions from duty. For in cases in which an agent is inclined to do wrong, the only way he can do right, and thus manifest his commitment to morality, is to rely on the incentive provided to him through his representation of moral principle. Of course, a person with a good will does not constantly perform morally required actions; some of what he does is morally permissible but not required, and thus cannot be done from duty (as a "primary motive"). Although Kant would acknowledge a distinction between an agent's acting from duty (as a "primary motive") and her governing her conduct by a commitment to conforming to moral principle, he does not refer to individual cases of her doing the latter but not the former as ones of acting *from duty*.[15] However entrenched the vocabulary of acting from duty as a "limiting condition" or "secondary motive" has become in contemporary discussions, it is not Kant's. For simplicity's sake, I do not adopt it. In my terminology, an agent acts "from duty" only in cases in which his incentive for acting stems from the notion that doing so is morally required. Only in such cases, suggests Kant, does his action have moral worth.

Although Kant holds that actions from duty exclude the influence of inclination, he does not maintain that *having* an inclination to do something is incompatible with doing it from duty. This point has recently been made by several philosophers, and I do not belabor it here.[16] According to Kant, an agent's motive for doing something can be that the supreme principle of morality requires it, even if the agent wants to do it. I might have an inclination to keep my promise to a friendly acquaintance. But that does not entail that my motive for keeping it could not be that the supreme principle of morality requires it.[17] Kant implicitly distinguishes between an action's being accompanied by an inclination and its being motivated by one – that is, its being done *from* inclination.[18] The former is compatible with the action's having moral worth.

But what about actions done both from duty and from inclination? Could there be such actions on Kant's scheme, and would any of them have moral worth? These questions have recently been at the center of intricate and extensive debate.[19] Exploring them in detail would take us far from our central concern: Kant's derivation of the supreme principle of morality. Brief consideration of these questions, however, can lead us to a second condition that must be met if an agent is to act from duty.

As we have noted, Kant says that "an action from duty is to put aside entirely the influence of inclination" (GMS 400; see also KpV 72 and 81). Reflecting on Kant's theory of agency and his account of nonmoral action leads us to a better understanding of this dictum.[20] Kant embraces the Incorporation Thesis, according to which no incentive can determine an agent's will unless she has incorporated it into a maxim (see sections 1.2 and 2.2).

This thesis applies not only to inclinations as incentives but also to moral principle as incentive. The thought that an action is required by moral principle can serve as an agent's motive for acting only if she has taken account of this thought in some self-given rule. Let us now consider an alleged case of acting both from duty and from inclination. Suppose that from both an agent keeps a promise. She has incorporated into her maxim both an inclination, one to preserve professional ties with a business associate, and the thought that keeping the promise is required by a moral principle. Her maxim would be something like this: "Because I want to maintain my business reputation and because keeping promises is morally required, I will do what it takes to keep my promises to my business associates." Contrary to Kant's dictum, in acting on this maxim the influence of inclination is obviously not entirely excluded. For the maxim makes the agent's doing what it takes to keep her promise conditional on her wanting to maintain her business reputation. She will not act on her maxim unless she has a desire to maintain it. Therefore, acting on this maxim would not amount to acting from duty. When Kant suggests that an action from duty excludes entirely the influence of inclination, he is implying that an agent who acts from duty must take the action's being morally required as itself generating enough of an incentive for her to do it. The second condition on conforming to a moral principle from duty – that is, conforming to it because the principle requires it – is that one must take the action's being morally required itself to generate a sufficient incentive for performing it.

But is it really the case that no actions done both from duty and from inclination could meet this second condition? Suppose someone acted on the following, rather awkward, maxim: "Because I want to maintain my good business reputation and because keeping promises is morally required (which is itself sufficient incentive for me to keep them), I will do what it takes to keep my promises to my business associates." It appears that the agent would be acting not only from inclination, but also from duty, thereby fulfilling the second condition. Yet the agent would not actually count as acting from inclination. Kant, I have argued, has a hedonistic view of acting from inclination (sections 1.6–8). According to him, if an agent acts from inclination, his performing the action is conditional on his expectation that realizing its object will give him pleasure. Suppose that from my inclination to preserve my professional ties with an associate, I do what it takes to keep my promise to him. In this case, my taking the necessary steps to keep my promise is conditional on my expectation that preserving these ties will have a hedonic payoff. I do not treat the notion that keeping promises is morally required as a sufficient incentive for doing what it takes to keep my promise. Therefore, I cannot be acting on the maxim in question: one in which an agent does treat the moral necessity of keeping promises as a sufficient incentive for his action. If one genuinely performs an action from inclination in Kant's sense, then she cannot fulfill Kant's second condition on acting from duty – she

cannot, at the same time, hold that the action's being morally required itself engenders a sufficient incentive for performing it. She must acknowledge that a further incentive is needed as well, namely the expectation that doing the action will have some hedonic payoff. Kant *does not recognize the possibility* of an action's being done at the same time from inclination and from duty.

If, contrary to this conclusion, Kant held that a particular action could be done at the same time from duty and from inclination, then he would also be committed to the view that a hedonically conditioned action could have moral worth. For Kant, of course, *all* actions done from duty have moral worth. In the second *Critique,* however, Kant says:

> Now, because all determining grounds of the will except the one and only pure practical law of reason (the moral law) are without exception empirical and so, as such, belong to the principle of happiness, they must without exception be separated from the supreme moral principle and never be incorporated with it as a condition, since this would destroy all moral worth just as any empirical admixture to geometrical principles would destroy all mathematical evidence. (KpV 93)

In the Analytic of the second *Critique,* Kant offers a purely hedonistic account of happiness, according to which happiness is "a rational being's consciousness of the agreeableness of life uninterruptedly accompanying his whole existence" (KpV 22).[21] For Kant the term "agreeableness" (*Annehmlichkeit*) designates a kind of sensation (KpV 22). Since to experience this sensation is to experience pleasure (see, e.g., KpV 23), Kant is suggesting at KpV 93 that an action's being conditional on the expectation that it will result in some hedonic benefit for the agent destroys its moral worth. In effect, Kant denies the possibility that a hedonically conditioned action could have moral worth. That he denies this is, of course, consistent with my contention that he does not recognize the possibility of an action's being done at the same time from inclination and from duty (thus accruing moral worth).

To employ contemporary terminology, I am denying that in Kant's view there can be "overdetermined" actions – actions done from both duty and inclination, where either motive by itself would have sufficed.[22] For Kant an agent simply does not count as acting from inclination unless the motive of duty would *not* suffice for the action. All actions from inclination are hedonically conditioned.

In this section we have addressed some complex issues regarding Kant's conception of acting from duty. But our main aim has been to clarify when an agent conforms to a principle because the principle requires it. We have found that for this to occur, two conditions must be met. First, the agent's incentive for acting must stem from the notion that the principle (represented by the agent as a law) requires the action. Second, the agent's notion that the action is morally required must itself provide sufficient incentive for him to perform it. In short, when an agent acts from duty, his notion that his action is morally required provides him with a sufficient incentive

for acting. In effect, our investigation has shown that a principle to which an agent could conform from duty would have to meet one of the criteria we discussed earlier for the supreme principle of morality. According to this other criterion, which is the focus of section 5.7, the supreme principle must be such that an agent's representing this principle as a law – that is, a universally and unconditionally binding principle – gives him sufficient incentive to conform to it. As we have seen, if, from duty, an agent conforms his action to a principle, then his (in itself sufficient) incentive for acting stems from the notion that the action is required by an unconditionally and universally binding principle: a principle the agent has represented as a law.

Although in this section I have set out two necessary conditions for an action's being done from duty in Kant's sense, it has not been my intention to offer (jointly) sufficient conditions for an action's being done from this motive. Before I try to do this (section 6.9), I need to make explicit two further necessary conditions for an action's being done from duty.

At any rate, this section and the preceding one have led us to a better understanding of what Kant means when he suggests in *Groundwork* I that we cannot hold a principle to be the supreme principle of morality unless we can maintain that all and only actions that conform to it because the principle requires it have moral worth. For Kant, maintaining this commits one to the view that all actions with moral worth must actually conform to the supreme principle of morality (section 5.2). It also commits one to the view that in all actions with moral worth the agent's incentive is ultimately that the action is morally required – an incentive that the agent himself takes to be a sufficient basis for his action.

5.4 All Actions from Duty Have Moral Worth

In addition to understanding Kant's first criterion, we need to isolate his grounds for it. We have already considered Kant's grounds for holding that only actions that are in accordance with the supreme principle of morality can have moral worth (section 5.2). We now need to examine why he claims that whatever the supreme principle of morality is, we must be able to hold that all and only actions conforming to it because the principle requires such conformity have moral worth. Put more simply, we need to examine Kant's grounds for claiming that all and only actions from duty have moral worth. I propose to do so in this and the next section. This section is devoted to Kant's claim that if an action is done from duty, then it has moral worth – that is, all actions done from duty have moral worth. The next section (5.5) focuses on Kant's claim that if an action has moral worth, then it is done from duty – that is, no action done from a motive other than duty has moral worth.

A sufficient condition for an action's having moral worth, claims Kant in *Groundwork* I, is that it be done from duty.[23] But Kant does not there

present an argument for this claim. He sets out grounds for rejecting the notion that actions from motives other than duty have moral worth (section 5.5). Yet he apparently finds it unnecessary to argue that all actions done from duty possess such worth. Consider, for example, Kant's discussion of self-preservation. Kant suggests that we have a duty to preserve our lives and that, the vast majority of the time, when we take steps to preserve them, we are acting from an immediate inclination to stay alive. "But on this account," Kant says, "the often anxious care that most people take of [their lives] still has no inner worth and their maxim has no moral content. They look after their lives *in conformity with duty* but not *from duty*" (GMS 397–398). Kant simply assumes here that, if a person preserves his life not from inclination but from duty, "his maxim has moral content," and thus acting on it has moral worth. It appears that while Kant thinks he needs to help us to see that actions done from immediate inclination fail to have moral worth, he supposes we find it obvious from the very outset that actions done from duty possess moral worth. He assumes that this view is obvious to "ordinary moral reason." If we reflect on our moral judgments, we will very quickly find that, in our view, doing what is morally required because it is morally required has moral value.

At bottom Kant seems to take it as given that, according to ordinary moral reason, if an action is done from duty, then it has moral worth. Kant does, however, suggest an account of what is so special about such actions:

> The human being is a being with needs, insofar as he belongs to the sensible world, and to this extent his reason certainly has a commission from the side of his sensibility which it cannot refuse, to attend to its interest and to form practical maxims with a view to happiness. . . . But he is nevertheless not so completely an animal as to be indifferent to all that reason says on its own and to use reason merely as a tool for the satisfaction of his needs as a sensible being. For, that he has reason does not at all raise him in worth above mere animality if reason is to serve him only for the sake of what instinct accomplishes for animals; reason would in that case be only a particular mode nature had used to equip the human being for the same end to which it has destined animals, without destining him to a higher end. (KpV 61–62)

For Kant, in all acting an agent employs practical reason. (Without the faculty of reason, he could not give himself maxims, that is, the rules on which, in Kant's view, an agent acts.) When an agent acts from inclination, she always to some extent employs her reason as a tool for the satisfaction of her needs as a sensible being. In acting from many inclinations – for example, those for food, shelter, or sex – an agent is in a straightforward way aiming to satisfy such needs. But what about acting from an inclination to write a good novel or to solve a mathematical puzzle? Even in acting from these inclinations, which might not seem to have much to do with her needs as a sensible being, an agent would to some extent be using her reason as a tool to satisfy such needs. For Kant, one of the needs we have

as sensual beings is to experience pleasure and avoid pain. In all actions from inclination, an agent's acting is conditional on his expectation that doing so will have some hedonic benefit. This is true with regard to actions done from the inclination to solve a mathematical puzzle just as it is with regard to actions done from the inclination for sex. Nature has foisted on all animals, including human beings, a need for pleasure. In all of our acting from inclination, we necessarily take account of this need to some extent. When an agent acts from duty, however, he is not necessarily using his reason as a tool for the satisfaction of his needs as a sensible being. His having sufficient incentive to act does not depend on his expectation that acting will enable him to gain some hedonic benefit. That actions from duty are not tied to sensible needs, which we share with other animals, gives them a special value. In acting from duty we fully elevate ourselves above the beasts.[24]

5.5 Only Actions from Duty Have Moral Worth

To defend the view that *only* actions from duty have moral worth, Kant highlights two conditions on actions with such worth, both of which he takes to be accepted by everyday moral consciousness. He then intimates that no action from inclination could meet these conditions.

Kant introduces the first condition in the *Groundwork* Preface:

> [I]n the case of what is to be morally good, it is not enough that it *conform* with the moral law; but it must also be done *for the sake of the law*; without this, that conformity is only very contingent and precarious, since a ground that is not moral will indeed now and then produce actions in conformity with the law, but it will also often produce actions contrary to the law. (GMS 390)

As we discussed earlier (section 5.2), Kant holds that only actions that conform with duty can be morally good, that is, have moral worth. Kant here points to a condition on a morally valuable action: it must be done from a motive that will not produce actions contrary to duty. In the *Groundwork*, Kant maintains that acting "for the sake of the law" – that is, doing something because you take it to be required by moral principle – meets this condition, whereas acting from inclination does not.

Kant invokes this condition in his famous discussion of the "philanthropist" (or "friend of humanity") (GMS 398). Before undertaking this discussion, Kant suggests a distinction between acting from a mediate inclination (self-interest) and acting from an immediate inclination (GMS 397). A mediate inclination to do something is an inclination to do it for the sake of fulfilling some further inclination. The shopkeeper in Kant's example presumably has a mediate inclination to charge his customers fairly. He wants to do it but merely as a means to satisfying another end, for example, that of having a thriving business. An immediate inclination to do

something is an inclination to do the thing itself. Since he is "sympathetically attuned," the philanthropist presumably has an immediate inclination to promote the well-being of others. His inclination to help them is not one that he strives to satisfy merely to fulfill some further desire. Kant, of course, denies that acting from this inclination has moral worth. Doing so, he says, is like acting from other inclinations, for example, the inclination to honor, "which, if it fortunately lights upon what is in fact in the common interest and in conformity with duty and hence honorable, deserves praise and encouragement but not esteem" (GMS 398).[25] Here Kant underscores the possibility that, in acting from an immediate inclination to help others, that is, from sympathy, an agent might do something that conflicts with duty. (To echo a well-known example, someone might, because of his sympathetic temper, have an immediate inclination to help someone he sees late one night quietly struggling to move a sculpture out the back door of an art museum and into his waiting car.[26] Acting from this inclination might presumably be contrary to duty.) Since the philanthropist is acting from an immediate inclination, and thereby doing something that might fail to accord with duty, his action, Kant suggests, does not have moral worth.

Yet, as Herman emphasizes, in his discussion of the philanthropist Kant points to a further condition he places on an action's having moral worth.[27] Kant says that the maxim on which the philanthropist acts "lacks moral content, namely that of doing such actions not from inclination but *from* duty" (GMS 398). Kant does not tell us explicitly what the philanthropist's maxim is. From the description Kant provides, however, we can assume that it is something like the following: "Because I want to help others, I will promote their happiness." This maxim, says Kant, lacks moral content, and it is not hard to pinpoint a reason why. The maxim reflects no commitment to the action's being morally permissible, that is, in accordance with what moral principle requires. In other words, the maxim expresses no interest in the rightness of the kind of action it specifies, namely promoting others' happiness. If we reflect on our ordinary moral understanding, suggests Kant, we find that we are willing to attribute moral worth only to actions done on maxims that (if fully specified) reflect a commitment to doing only what is morally permissible. The grounds of a morally valuable action – its motive – must express an interest in the action's moral rightness. This is Kant's second condition for an action's having moral worth. It is a necessary condition, not a sufficient one. That an agent does something against the background of a commitment to doing what is morally permissible does not entail that his action has moral worth. What the agent does might be morally permissible but not morally required. And for Kant only morally required actions can have moral worth. According to Kant, of course, actions from duty fulfill this second condition. In them, an agent's basis for acting – his maxim – obviously expresses concern for his action's

moral rightness, for it invokes the notion that actions of its kind are morally required.

As Herman has pointed out, Kant would insist that an action might fulfill the first condition for moral value without fulfilling the second.[28] Suppose, for example, that the philanthropist's immediate inclination to help others were such that it served as the basis only for morally permissible actions. In that case, the philanthropist's beneficent actions would fulfill Kant's first condition; they would be done on a motive that always produced actions conforming to duty. Nevertheless, the philanthropist's actions would still not have moral worth; for the grounds of his actions would fail to express concern for their moral rightness, thereby running afoul of the second condition.

In the *Groundwork*, Kant maintains that only actions from duty can have moral worth, since only these actions meet each of the two conditions we have discussed. However, there might be a third condition Kant places on morally valuable actions. This condition is implicit in our discussion of a Kantian ground for assigning special worth to acting from duty (section 5.4). Unlike in acting from inclination, Kant suggests, in acting from duty, we are not using our reason as a tool for the satisfaction of needs foisted upon us by nature. Kant intimates that the special worth of acting from duty derives from such independence from natural desire. Perhaps Kant holds that actions with moral worth must reflect this independence, that is, must elevate the agent above striving to satisfy needs we share with other animals. In Kant's view, no action from inclination could meet this condition; for, as we have seen, all of them are conditional on the agent's expectation that they will result in promoting the satisfaction of a natural need, namely that for pleasure.

The past several sections (5.2–5) have been devoted to the first criterion for the supreme principle of morality that Kant develops in *Groundwork* I: the supreme principle of morality must be such that all and only actions conforming to it because the principle requires it – that is, all and only actions done from duty – have moral worth. We have focused on clarifying this criterion and illuminating Kant's grounds for it. In sum, Kant asserts, to take a principle as the supreme principle of morality, we must be able to hold the following: an agent's action has moral worth when and only when the agent's (correct) notion that an unconditionally and universally binding principle requires the action is his (in itself sufficient) incentive for performing it. In other words, an agent's action has moral worth if and only if it is done from duty. Although Kant seems to take as obvious that all actions done from duty have moral worth, he offers arguments for the view that only such actions have it. Among the arguments are the following. Unless it is from duty, an action is done from a motive that may produce actions contrary to duty, and no action from such a motive has moral worth. Moreover, the maxims of actions not done from duty are devoid of moral

content. Since they do not reflect concern for moral rightness, acting on them cannot have moral worth.

5.6 The Second Criterion and Its Grounds

We need now to turn to two other criteria that, I have argued, Kant advances in *Groundwork* I. Clarifying these criteria and their grounds can be done relatively quickly. This section and the next consider the second and third criteria respectively.

According to criterion 2, the supreme principle of morality must be such that the moral worth of any given instance of conforming to it from duty stems from its motive, not from the effects actually produced by this instance. We cannot affirm a principle to be the supreme principle of morality unless we can hold that the moral worth of actions conforming to it from duty does not stem from the actions' results. That Kant embraces this criterion is clear. In his "second proposition," he says that the moral worth of an action done from duty "does not depend upon the realization of the object of the action but merely upon the principle of volition in accordance with which the action is done" (GMS 399–400, emphasis omitted). Later in *Groundwork* I Kant says that "the moral worth of an action does not lie in the effect expected from it" (GMS 401). The criterion relies on a distinction between an action and its effects. For Kant, to act in the relevant sense is, strictly speaking, to exercise one's will (section 1.4). It is to try, based on some principle (some maxim), to realize a state of affairs (an object or end). This state of affairs (or whatever state of affairs actually results from the action) is an effect of the willing. Acting consists in the willing itself, not in its effects. According to the second criterion, it is not the results of acting from duty – that is, willing to conform to the supreme principle of morality because the principle requires it – that gives it moral value.

Implicit in *Groundwork* I is a straightforward argument for this second criterion. Suppose that, contrary to it, the moral worth of an action from duty *did* stem from its effects. There would, then, be possible circumstances in which an action from duty did not have moral worth, namely ones in which the action failed to produce certain effects. For Kant, however, if an action is done from duty, then it has moral worth, no matter what the circumstances may be. His first criterion incorporates this view. Moral worth is "unconditional," Kant suggests (GMS 400). Therefore, as the second criterion indicates, the moral worth of an action from duty does *not* stem from its effects. For example, suppose that an agent holds the supreme principle of morality to be: "Always do what you believe will please God." Moreover, contrary to the second criterion, the agent maintains that the moral worth of her conforming to this principle because the principle requires it – that is, the moral worth of her acting from duty – stems from its effects. Whether her action has moral worth, she thinks, depends on whether it actually pleases

God. There would then presumably be possible circumstances in which her acting from duty would not actually please God. As a fallible being, she might be mistaken as to what would please God. In these circumstances, the agent would be compelled to maintain, her acting from duty would be devoid of moral worth. But this acknowledgment would contradict Kant's first criterion, one of the constitutive claims of which is that a sufficient condition for an action's having moral worth is that it be done from duty. In short, Kant defends the second criterion by appealing to the first. That the effects of our actions can give them "no unconditional and moral worth," he says, "is clear from what has gone before" (GMS 400). What has gone before, of course, is Kant's discussion of the relations between acting from duty and moral worth: a discussion that lays the basis for his first criterion.

5.7 The Third Criterion and Its Grounds

According to the third criterion, the supreme principle of morality must be such that our representing it as a law provides us with sufficient motive to adhere to it. If this criterion is correct, then we can (rationally speaking) maintain a principle to be the supreme principle of morality only if we can hold that our representing it as a law – that is, a universally and unconditionally binding principle – gives us a sufficient motive to conform to it.

It might seem that this criterion is entailed by criterion 1, according to which the supreme principle of morality must be such that an action has moral worth if and only if it conforms to the principle because this principle requires it. After all, in probing the meaning of criterion 1 (section 5.3), we found that, in Kant's sense, an action conforms to a principle because the principle requires it (the action is done from duty) only if the agent's (in itself sufficient) incentive for acting is the notion that the principle, represented by the agent as a law, requires the action. Strictly speaking, however, we can imagine scenarios in which it would be possible for a principle to fulfill criterion 1 yet fail to fulfill criterion 3. For example, suppose that acting from duty is impossible and that no action can have moral worth. Obviously, on this supposition, no principle could fulfill 3. But all principles would fulfill 1. For criterion 1 just says that a viable candidate for the supreme principle of morality must have the following characteristics. It must be such that *if* there are actions that have moral worth, then they are done because the principle requires them, and *if* there are actions done because the principle requires them, then these actions have moral worth. Against the background of our supposition, the antecedent of each conditional is necessarily false, rendering each conditional trivially true. So no practical principle could fulfill 3, but any such principle would, albeit trivially, fulfill 1. Actually, 1 does not entail 3.

As far as I can tell, Kant suggests two arguments for criterion 3. He hints at one in his *Groundwork* discussion of the "second proposition" (GMS 399–400). Kant finds in ordinary moral thinking the view that conforming to the supreme principle of morality can have unconditional worth. Against the backdrop of this view, the argument unfolds as follows. Denying criterion 3 would, Kant seems to assume, amount to holding that the supreme principle of morality must be such that each agent's expectation of the effects of conforming to it necessarily constitutes (at least part of) the agent's incentive for conforming to it. Now suppose an agent denies 3 and takes a particular principle to be a viable candidate for the supreme principle of morality. She would then be committed to the view that the *value* of her conforming to this principle necessarily derives (at least in part) from its effects. After all, if, in her view, conforming to the principle were valuable in itself, then she would not hold that she *necessarily* needs to look to its effects to find a sufficient incentive to do so. But if the agent inextricably ties the value of her conforming to a principle to its expected effects, then she is rationally compelled to deny that her conforming to it can have unconditional worth. She must hold that its having such worth would always depend on some conditions being met, that is, on whether the expected effects actually occur. But, according to ordinary moral reason, conforming to the supreme principle of morality can have unconditional worth. Therefore, the agent must not deny criterion 3, but instead agree that we can hold a principle to be the supreme principle of morality only if we can maintain that our representing it as a law governing our actions gives us a sufficient incentive to conform to it.[29]

Kant suggests another argument for criterion 3 in the second *Critique* (KpV 21–22). According to him, to reject the criterion is to hold that the supreme principle of morality could be a material practical principle. But, Kant argues, no material practical principle could be a practical law (the supreme principle of morality). The supreme principle of morality must, he maintains, be absolutely necessary (section 1.2). A human agent would always be obligated to conform to the supreme principle, no matter what he desired or took pleasure in.[30]Moreover, Kant maintains that an agent's having an obligation to do something entails that he is able to do it: ought implies can (e.g., KpV 159). Kant thus holds that the supreme principle of morality must be such that each of us is necessarily able to conform to it. But each of us is necessarily able to conform to a principle only if each one necessarily has sufficient motive to conform to it. And, Kant asserts, no material practical principle is such that each of us necessarily has sufficient motive to conform to it. According to Kant's account of such principles, an agent will have sufficient motive to conform to a given one only if she expects that doing so will enable her to realize some object she desires and that realizing this object will give her pleasure (section 1.8). But there is

nothing to guarantee that she will expect these things from conforming to a given material principle. Whether she will is a contingent – Kant might say "empirical" – matter. Suppose, for example, that we (and the agent) interpreted the following as a material practical principle: "In order to perfect yourself, you ought to develop your physical strength and flexibility."[31] The agent's having sufficient motive to develop the capacities in question would, in part, depend on whether she expected doing so to have a hedonic payoff. But instead of expecting this, she might think that she is strong and flexible enough and that more exercise would be a painful waste of time. The agent would, then, have insufficient motive to conform to the principle. She would recognize from her own case that it did not meet the absolute necessity requirement of the supreme principle of morality. In sum, Kant argues that unless a principle meets the third criterion for the supreme principle of morality, it cannot conform to his basic concept of this principle, specifically to the notion that the principle must be absolutely necessary.[32]

5.8 Relations between the Criteria

On my reading, Kant offers a set of criteria for the supreme principle of morality. According to Kant's basic concept, this principle must be practical, absolutely necessary, binding on all rational agents, as well as the supreme norm for the moral evaluation of action. I have argued that, in the course of *Groundwork* I and II, Kant develops four more criteria. The supreme principle of morality must also be such that: (1) all and only actions conforming to this principle because the principle requires it – that is, all and only actions done from duty – have moral worth; (2) the moral worth of (any case of) conforming to this principle from duty stems from its motive, not from its effects; (3) an agent's representing this principle as a law – that is, a universally and unconditionally binding principle – gives him sufficient incentive to conform to it; (4) a plausible set of duties (relative to ordinary rational knowledge of morals) can be derived from this principle. This chapter has focused on understanding what criteria 1–3 mean and how Kant defends them.

At several points, I have discussed relations between Kant's criteria. But it might be helpful for me to summarize them. According to Kant, criterion 2 follows from 1. In brief, if one holds that in all possible circumstances an action done from duty has moral worth, then one is committed to the view that this worth cannot stem from the action's effects. For there are possible circumstances in which the effects do not occur. Strictly speaking, 3 does not follow from 1 or from 2. If either 1 or 2 entailed that some actions actually are done from duty, then it would yield 3. That is because in order for there to be any actions from duty, criterion 3 would have to be fulfilled. However, neither 1 nor 2 entails that there are any such actions. But Kant does suggest two arguments for 3, as we have just seen. One appeals to his

notion that according to ordinary moral consciousness, there are actions that have unconditional value. The other is based on Kant's axiom that ought implies can, as well as a criterion that belongs to his basic concept of the supreme principle of morality, namely that this principle must be absolutely necessary.

Of course, it is one thing to understand how Kant argues for his criteria and quite another to accept them. The next chapter considers several objections to what is perhaps Kant's most controversial criterion, that according to which the supreme principle must be such that all and only actions conforming to this principle because the principle requires it – that is, all and only actions done from duty – have moral worth.

6

Duty and Moral Worth

6.1 Aims of the Discussion

The success of Kant's derivation of the Formula of Universal Law (as well as the Formula of Humanity) depends on his ability to eliminate rival candidates for the supreme principle of morality. To eliminate them Kant appeals to criteria for the supreme principle of morality. He argues that unlike his candidates, the rivals fail to remain as viable candidates for fulfilling the full set of criteria. As Chapter 7 illustrates in detail, the derivation relies on a criterion (or part of one) that has been a main topic for the past two chapters. This principle, the criterion goes, must be such that all and only actions conforming to it because it is morally required – that is, all and only actions done from duty – have moral worth. We now understand what this means and how Kant argues for it. This chapter explores the criterion's plausibility. It addresses objections to the view that an action has moral worth if and only if it is done from duty. The bulk of the chapter focuses on the claim that all actions done from duty have moral worth (sections 6.2–9). The penultimate section (6.10) takes up the claim that only actions from duty have such worth. The chapter focuses more on the former than the latter claim for a couple of reasons. Whereas I want to defend the former claim (albeit understood a bit differently than Kant does), I do not find the latter entirely plausible. Moreover, I think that the former plays a much more central role than does the latter in the elimination of rival candidates for the supreme principle of morality.

Kant claims that if an action is done from duty, then it has moral worth. In the *Groundwork*, however, he does not so much argue for this view as point to it as a fundamental tenet of ordinary rational knowledge of morals, a starting point not really in need of defense. In the second *Critique*, he does suggest a reason why actions from duty have a special value. They are not conditional on our expectation that they will fulfill any sensible needs, and they thus elevate us over other animals, whose behavior is geared toward

fulfilling such needs (see section 5.4). But the force of this suggestion depends on two controversial notions. The first is that acting from inclination is more animal-like than acting from duty. But this is not obvious. While acting from the inclination to slake one's thirst clearly seems more animal-like than acting from duty, acting from the inclination to prove a mathematical theorem does not. Is an action done not from duty but from a desire, such as that to prove a theorem, really always conditional on the agent's expectation that the action will result in some hedonic benefit for him? Kant asserts this, but he does not establish it. The force of Kant's suggestion also depends on the notion that since an action from duty is less animal-like than one from inclination, the former has a special value that the latter lacks. But some might, in a Nietzschean vein, hold that this notion smacks of an irrational devaluation of our animal nature. I doubt whether Kant's second *Critique* suggestion as to why actions from duty have a special value will (or was intended to) change the views of those who do not think they do.

I believe, however, that Kant is fundamentally correct in holding it to belong to ordinary moral consciousness that all actions done from duty have moral worth. There is a significant adjustment to his view that I propose in section 6.6, one that arises from internal critique. Presently I consider external critique of Kant's view (sections 6.2–3). I discuss two objections to it with the aim not of refuting them definitively, an aim that seems out of place with respect to issues that must ultimately be adjudicated by controversial appeals to intuition, but of blunting the objections' force so that we can see that it is at least reasonable, and perhaps even attractive, to hold the Kantian view. The two objections stem from general criticisms of Kantian morality developed by Bernard Williams and Michael Stocker.[1] These criticisms have been thoroughly addressed by Kantians before, and much of my discussion, which concerns only how they apply to Kant's claim that all actions from duty have moral worth, draws on their work.

Before beginning the business of the chapter, it might be helpful to bring together some general points regarding Kant's notion of an action's moral worth. First, for Kant to act is to exercise one's will (section 1.4). It is to attempt, based on some principle, to realize a state of affairs. This state of affairs (or whichever one really results from the action) is an effect of the willing. Acting consists in the willing itself, not in its effects. So to say that a certain kind of action has moral worth is really just to say that a certain kind of willing has such worth. Second, for Kant moral worth is unconditional worth (section 4.6). If a particular type of action, for example, action done from duty, has moral worth, then every possible token of this type has such worth. Third, according to Kant, moral worth is a "preeminent" good (GMS 401). This suggests that if only one particular type of action has moral worth, then actions of this type have higher value than actions of any other type.

6.2 Moral Worth and Helping a Friend from Duty

The first objection I consider to Kant's view that all actions from duty have moral worth can be derived from a well-known scenario sketched by Stocker:[2]

> [S]uppose you are in a hospital, recovering from a long illness. You are very bored and restless and at loose ends when Smith comes in once again. You are now convinced more than ever that he is a fine fellow and a real friend – taking so much time to cheer you up, traveling all the way across town, and so on. You are so effusive with your praise and thanks that he protests that he always tries to do what he thinks is his duty. . . .You at first think he is engaging in a polite form of self-deprecation, relieving the moral burden. But the more you two speak, the more clear it becomes that he was telling the literal truth: that it is not essentially because of you that he came to see you, not because you are friends, but because he thought it his duty.[3]

Stocker goes on to suggest that Smith's action is "lacking in moral merit or value." Stocker might mean by this that, though Smith's action is good, it could be better, and is in that sense lacking in moral value. In other words, his visit to his friend has moral value, but not the most moral value that such an action might have. But if this were Stocker's contention, then it would not threaten the particular claim that all acting from duty has (some) moral worth. Granted, following Kant we have understood whatever has moral value to be unconditionally and preeminently good – that is, good in all possible situations and always better than anything possessing some other kind of goodness. Yet consistent with this understanding is the view that moral value itself might come in differing degrees. At any rate, this initial reading of Stocker's passage seems to me less natural than another, according to which he is charging that Smith's action is simply devoid of moral value. Since this charge would threaten Kant's claim, I focus on it.

Although Stocker does not conceive of Smith's motive precisely in Kantian terms – the target of his criticism is modern ethical theories as a whole, not primarily Kantianism – let us do so. Smith, let us say, makes his visit from Kantian duty. He takes the notion that helping others is morally required as his incentive (and a sufficient one) for visiting his friend, call her Jones, in the hospital. (He acts on a maxim such as this: "Because it is morally required, I will promote others' well-being.") Why, according to Stocker, does Smith's visiting Jones from duty lack moral value? The main reason seems to be that it is not "essentially" because of Jones, not because she and Smith are friends, that Smith visits her. Concern for his friend does not constitute Smith's basis for visiting Jones, and thus his action lacks moral value.

However, do we really hold that since Smith's basis for visiting Jones is not concern for her that it is devoid of moral value? Imagine that Smith does have concern for Jones. But Smith finds that, in itself, this concern is not strong enough to outweigh his great anxiety at the prospect of visiting a hospital. In Kant's terms, Smith's inclination to avoid the hospital is stronger

than his inclination to cheer up his friend. Nevertheless, from duty, Smith brings himself to go to the hospital and do his best to raise his friend's spirits. I think that in such a case most of us would grant that Smith's action had moral worth. So, in the kind of circumstances Stocker describes, that one does not act from concern for one's friend does not itself appear to preclude one's action from having moral worth.

Yet perhaps other ways of filling in the details of the scenario will reveal that upon reflection we hold that Smith's acting from duty might not have moral worth. Suppose that though Smith has no disinclination for hospitals, he does not want to comfort Jones. Smith is like one of the people Kant describes in the *Groundwork*. He is "by temperament cold and indifferent to the suffering of others," including his friends, "perhaps because he himself is provided with the special gift of patience and endurance toward his own sufferings" (GMS 398).[4] Would Smith's action, done from duty, of trying to cheer up Jones be entirely lacking in moral worth? We would condemn as wrong attempts by Smith to deceive Jones into thinking that he sympathizes with Jones's suffering. But Smith appears to be quite frank with Jones in the scenario as Stocker describes it. We would also question whether, given Smith's lack of sympathy, we would choose to have friends like him. Some of us might even insist that Smith is incapable of being a true friend, since genuine friendship necessarily involves having the very sympathy he lacks.[5] Nevertheless, I do not believe that we would deny all moral value to Smith's action. After all, from duty, he did do his best to improve Jones's condition, and there seems to be something morally admirable in that.

If there is a lingering unwillingness to attribute any moral worth to Smith's action, it is, I suspect, based on the worry that in the scenario just sketched, he sees Jones's misfortune as an opportunity of sorts. What matters to Smith, according to the worry, is not that he can raise Jones's spirits, but rather that her suffering provides him with an occasion to discharge his duty. He is using his visit to Jones as an instrument to increase his own moral merit – behavior that some might find to be lacking entirely in moral value. Yet if Smith is doing this, then he is not really acting from duty. When an agent acts from duty, she takes as a sufficient incentive for acting the notion that her action is morally required. If Smith were really using his visit to Jones as a means to increase his moral point total, then he would not be doing this. For he would be treating the notion that his moral merit would increase as (part of) his incentive for acting. He would presumably not visit Smith if he thought that doing so would have no effect on his moral merit or that moral merit was not additive, and so forth.

In short, it does not appear that Stocker's example shows that acting from duty (as Kant envisages such action) sometimes lacks all moral worth. However, we need to be clear on what this entails. Acknowledging that Smith's action has moral value does not *in itself* commit us to the view that *only* visiting Jones from duty, as opposed to some other motive such

as sympathy, would have moral value. Kant might be correct that all, but incorrect that only, actions from duty have moral worth. We explore this possibility in section 6.10. Nor does acknowledging that Smith's action has moral value *in itself* commit us to holding any position regarding the relative moral worth of his acting from duty versus some other motive. Although, in a given situation, acting from duty is morally valuable, acting from some other motive could be more so.

6.3 One Thought Too Many?

One might base a further objection to Kant's claim that all acting from duty has moral worth on the notion that doing so sometimes involves "one thought too many," to use Bernard Williams's phrase.[6] Suppose that an agent's husband, who has accidently fallen overboard, is drowning and that the agent can rescue him, with little risk to herself. She does not simply take action, motivated by a desire to save her husband, but rather she reflects then and there on her duty. She quickly decides that rescuing her husband is morally required and, from duty, jumps in the water to save him. Does the agent's reflection empty her action of all moral worth? Just as we did in the case of the hospital visit, we need to distinguish between the claim that acting from duty has moral value and the further claims that *only* acting from this motive has moral value or that acting from this motive always has the highest moral value. Here, once again, the first claim alone is at issue. And, once again, I believe that it resists counterexample.

Consider three ways of filling in some of the details of the story.[7] First, imagine that the agent is deeply estranged from her husband, who is an abusive drunk. In this case, I take it to be obvious that the agent's reflection would not rob her action of moral worth. Her husband might wish that an inclination to save him would have played a role in motivating her action, but his wish seems to be irrelevant to the issue of whether her action had moral worth. Second, suppose that the agent loves her husband and, on some level, realizes that saving him will pose little risk to herself. But since she has an irrational fear of swimming in the ocean, she is strongly disinclined to jump in. If, nevertheless, the reflection that it is her duty to save her husband steels her for the plunge and, from duty, she saves him, then it seems unproblematic to say that her action has moral worth. In both of these cases, had the agent not reflected on her duty, she would have had not one thought too many but one thought too few. Yet what about the following, third, specification of the drowning case? The agent who loves her husband dearly has a strong inclination to rescue him and none not to do so. But before she dives in to save him, she reflects that helping others in peril is morally required, and then, from duty, she rescues him. There is something odd about this scenario. What would prompt the agent to reflect in that instant that she has a duty of beneficence? What would bring

the agent in this case to act from duty, rather than from inclination? Yet however unusual the situation might be, the question remains: would her rescue be devoid of moral worth? I still do not see why it would. The agent is not treating her husband's peril as an occasion to make a deposit in her moral bank account – otherwise, she would not be acting from duty. She is doing something that conforms with duty because she thinks she morally ought to do it. And this action does seem to have some moral value.

6.4 The Moral Worth of Actions Contrary to Duty

Kant's claim that all actions from duty have moral worth withstands some well-known objections. At this point, however, someone might wonder why I have not considered a further, seemingly obvious objection. Don't people, in acting from duty, sometimes do horrific things? Think of "ethnic cleansers" who, apparently from duty, round up and kill members of a hated minority. Surely we would resist acknowledging that these actions have moral worth.

This objection would, however, be misplaced as a criticism of Kant's understanding of the notion that all actions from duty have moral worth. For Kant all acting from duty is acting in accordance with duty, as he and, he thinks, we conceive of duty (section 5.2). The actions of the ethnic cleansers are contrary to Kantian duty – surely they are not treating their victims as ends-in-themselves – and they thereby cannot be from duty, holds Kant. So from Kant's own perspective, the view that these actions have no moral worth does not threaten Kant's claim.

However, as I argue in sections 6.5–6, things are not so simple. For Kant should acknowledge that actions contrary to duty can be done from duty and thus can have moral worth. If I am right about this, the objection in question does come into play, a point I address in section 6.9.

6.5 A Disturbing Asymmetry in Kant's View of Moral Worth

According to Kant, all actions from duty are morally permissible (section 5.2). No morally impermissible action, he implies, can be done from duty, and none, therefore, can have moral worth. I argue that Kant should relinquish this position. He should hold instead that some actions contrary to duty can actually be done from duty and thereby have moral worth.[8] Key to my argument is the observation that there is an asymmetry in Kant's account of how two kinds of failure affect the question of an action's moral worth. While failure to judge correctly whether one's action is morally permissible precludes it from having moral worth, failure to attain the end of one's action does not. This asymmetry is disturbing because the very considerations that imply the one kind of failure to be irrelevant to the assessment of moral worth suggest the other kind to be irrelevant as well. Both kinds of failure can be due to circumstances beyond an agent's control and thus,

in the spirit of Kantianism, immaterial to an action's moral worth. Kant, I claim, needs to acknowledge that morally impermissible actions can be done from duty and thus can have moral worth.[9]

Let me now explain in detail the asymmetry I find in Kant's view. In *Groundwork* I, Kant insists that an action can have moral worth even if it does not bring about its intended results.

> [A]n action from duty has its moral worth *not in the purpose* to be attained by it but in the maxim in accordance with which it is decided upon, and therefore does not depend upon the realization of the object of the action but merely upon the *principle of volition* in accordance with which the action is done. (GMS 399–400)

Like each of the fundamental claims in *Groundwork* I, Kant bases this one on ordinary knowledge of morality.[10] Kant appeals to our intuition that an action done from duty has moral worth even if it does not succeed in realizing its end. Why doesn't the failure of an action done from duty to bring about its intended effects disqualify it from having moral worth? The answer seems to be: because such a failure is outside the agent's control. In a well-known passage, Kant says:

> A good will is not good because of what it effects or accomplishes. . . . Even if, by a special disfavor of fortune or by the scanty provision of a stepmotherly nature, this will should wholly lack the capacity to carry out its purpose – if with its greatest efforts it should yet achieve nothing and only the good will were left (not, of course, as a mere wish but as the summoning of all means insofar as they are in our control) – then, like a jewel, it would still shine by itself, as something that has its full worth in itself. (GMS 394)

Here Kant suggests that an agent's action can express good will only if she does everything in her power to realize the action's end. Since in Kant's view the actions we do from duty express good will, Kant implies that to count as performing an action from duty, an agent must do her best to realize the action's end.[11] If an agent fails to muster all her resources in an effort to realize an end, her action is not really from duty and is thus devoid of moral worth. Factors presumably within an agent's control, such as the effort she makes to realize an end, count in determining whether her action has moral worth. But factors outside of her control seem not to count.[12] In this passage, Kant mentions conditions such that when they prevent an agent from realizing her end, they do not preclude her action from having moral worth. These conditions are an "unfortunate fate" or the "scanty provision of stepmotherly nature," both of which are clearly outside the agent's control. It seems to be in the spirit of Kant's remark to hold that an agent's action is not to be disqualified from having moral worth by anything we take to be outside of her control.

Kant embraces the notion that an agent does not determine all of the effects of her willing. The best-laid (and executed) plans sometimes come to

naught. But now the question arises: is an agent's choice of a plan of action itself under his control to such a degree that whenever he adopts a morally impermissible one, he has committed an error for which he is morally accountable? In the *Groundwork*, Kant writes of an agent's inclinations as a force, which, if he permits it, will push him to leave his duty unfulfilled. "The human being feels within himself a powerful counterweight to all the commands of duty," says Kant. This counterweight consists of "his needs and inclinations, the entire satisfaction of which is summed up under the name of happiness" (GMS 405). Kant appears to hold that each one of an agent's failures to act rightly stems from his privileging the satisfaction of some inclination over fulfilling his duty.[13] Instead of a question of succumbing to inclination, however, might not whether one succeeds in adopting a principle of action that is in accordance with Kant's standard of morality be a matter of one's circumstances, upbringing, or cognitive abilities? These questions point us toward the asymmetry I have in mind. It is an asymmetry between the way in which two different kinds of failure relate to an action's moral worth. On the one hand, Kant holds that failure to realize its end does not disqualify an action from having moral worth; on the other hand, he holds that failure to act on a morally permissible principle (and thus in a morally permissible way) does disqualify it. There is nothing blatantly contradictory in this asymmetry. But whether we should accept it depends on the plausibility of Kant's implicit view that our failure to perform a morally permissible action is always a failure of will – that is, a succumbing to inclination – and never an unfortunate event ultimately beyond our control.[14]

6.6 Failure of Will or Unfortunate Event?

I believe that this view is implausible, as I try to show with the help of a couple of examples. First consider the well-educated Colonel Mikavitch. A morally reflective person since she was a child, she has embraced the Formula of Universal Law as the supreme principle of morality; she tries to act only on that maxim by which she can at the same time will that it should become a universal law. Colonel Mikavitch, who has studied the *Groundwork*, believes that even though Kant offers several formulas of the supreme principle of morality, he insists that we do best if we adopt "the strict method" and make the basis of our moral appraisal the Formula of Universal Law (GMS 436–437).[15] Unfortunately and unforeseeably, a foreign power has attacked the colonel's country, bent on exterminating one of its ethnic minorities. With the enemy nearly on her doorstep and no hope of escape, she comes to the painful conviction that if she is captured, she will, under the weight of torture, reveal a secret known only to her: the location of several minority families. After careful consideration of the alternatives, she has decided that the only way to save the families is to kill herself. The colonel finds in herself no inclination to do so and, indeed, believes that suicide would require her

last ounce of courage. Although she thinks she has a moral duty to save the families, she wonders whether it is morally permissible for her to take her own life. She asks herself whether it is permissible to act on the maxim: "If my end is to save others but I find no means available but suicide, I will kill myself." After careful thought she judges that this maxim passes the Categorical Imperative test. She could will that it become a universal law. It is not self-contradictory to imagine a world in which, whenever an agent believed taking his own life to be the only means of securing his end of saving others, he killed himself. Furthermore, the colonel reasons that, as a rational being, she could act on the maxim and, at the same time, will that it become a universal law. Her willing would not sink her into rational self-contradiction; that every agent in circumstances like hers committed suicide would not prevent her from attaining her end of saving others by committing suicide herself. With the regretful thought that she must heed the call to save innocent lives, she takes poison.

On Kant's view, would Colonel Mikavitch's action have moral worth? Kant would be quick to make an epistemological point. Even supposing the colonel's action were in accordance with the Categorical Imperative, we could not conclude with certainty that it had such worth. "[I]t is absolutely impossible by means of experience to make out with complete certainty a single case in which the maxim of an action otherwise in conformity with duty rested simply on moral grounds and on the representation of one's duty" (GMS 407). Some "secret impulse of self-love," such as fear of torture, might actually have prompted the colonel's suicide.

Moreover, Kant would insist that the colonel's suicide could have moral worth only if it were in accordance with duty. Was it actually in accordance with Kantian duty? I am unsure. On the one hand, Kant argues that we have a duty to preserve ourselves in our animal nature (MS 421). On the other hand, in his casuistical discussion of this duty he brings up the case of Frederick the Great, who carried a fast-acting poison with him "presumably so that if he were captured when he led his troops into battle he could not be forced to agree to conditions of ransom harmful to his state" (MS 423). Since it is unclear whether Kant would morally condemn a country-saving suicide by a king, it is uncertain whether he would condemn the analogous family-saving suicide by the colonel.[16] Of course, even if Kant himself would always condemn suicide, it does not follow that the colonel's was contrary to the Categorical Imperative. Perhaps Kant was not the best interpreter of his own principle.

In any event, the point this case is designed to illustrate is simple. According to Kant, if the colonel's action was not morally permissible, it would follow that it could not have had moral worth. Suppose that – after due consultation with the world's greatest Kantian casuists – we discovered that, measured by Kant's principle, the colonel's suicide was morally impermissible. On my view, this discovery would not, in itself, warrant the

conclusion that her action failed to have moral worth. Intuitively speaking, as far as moral worth goes, it just would not matter. The colonel did her best to determine whether her course of action was morally permissible. If she did not succeed, her failure stemmed, it seems, not from lack of sincere effort but rather from the limits of her cognitive capacities. Kantian casuistry is hard. Much as it is often beyond her control whether the world cooperates and she succeeds in her efforts to promote the happiness of others, so is it beyond her control whether she succeeds in discerning whether a given action meets the standard of Kantian moral permissibility.

Kant, of course, might insist that if the colonel's action was contrary to duty, then it was really motivated by some inclination of which she was unaware. But this reply seems weak. Note that Kant's epistemological claim – neither the colonel nor anyone else can know for sure that she has acted from the notion that her action was morally required instead of from inclination – does nothing to bolster the reply. That it is impossible for us to know whether the colonel has acted from duty does not entail that she has actually failed to act from duty. Why, then, should we conclude that if, as it turns out, the colonel's action was contrary to duty, its wrongness was due to her succumbing to some inclination? Such a conclusion seems forced to fit Kant's denial of moral worth to all actions contrary to duty. And, to me, at least, it seems to go against "ordinary knowledge of morality": the very basis of Kant's case in the *Groundwork*.

The suicide example revolves around someone who has embraced the Categorical Imperative as the supreme principle of morality and, perhaps, has made an error in applying it. But suppose a person, call him Stram, has not embraced this principle. After years of long, careful, and strenuous reflection, Stram has concluded that a version of Act Utilitarianism is the valid moral doctrine. According to him, the supreme principle of morality is: always do what you believe will maximize the pleasure and minimize the pain of all sentient beings. Stram does his very best to live by this principle. He has internalized it to such a degree that at times he is surprised to find himself calculating the effects on sentient beings of even seemingly trivial deeds. At one point, he finds himself in a situation where he believes that lying is demanded by the combination of his circumstances and the utilitarian principle. After thinking through the alternatives, he decides that he must, in his position as an accounting consultant, lie to a politician about the county's ability to raise funds for a proposed dam. Only by lying, he judges, can he insure that the dam will not be built, wildlife decimated, and three whole towns destroyed. Because he believes it to be the right thing to do, Stram goes ahead and lies to the politician.

The key question here is not whether Stram's action is by Kant's standard morally permissible; let us just assume that it is not – that he did it on a maxim that fails the Categorical Imperative test. The central issue is, rather, whether the moral impermissibility of Stram's action would, as Kant suggests,

be a sure sign that it was lacking in moral worth. What if, as a sincere Act Utilitarian, Stram has used all of his powers in a struggle to determine what is right? What if, in Kant's terms, he did not merely wish but willed to discover the correct course of action? In this case, it appears that his failure might have stemmed not from his inclinations but rather from factors beyond his control: factors that should not, in the spirit of Kant's own view, matter in a determination of whether his action could have had moral worth.[17]

Notice that the cases of Colonel Mikavitch and Stram illustrate two ways an agent's action might fail to be morally permissible, yet have moral worth. In the first case, an agent has (in Kant's view) embraced the correct moral standard, but has failed to apply it accurately. This case illustrates an error in principle application. In the second case, an agent has (in Kant's view) embraced an incorrect moral standard and applied it correctly. This case illustrates an error in principle choice.[18]

Of course, examples like that of Stram and Colonel Mikavitch have an artificial ring. There may not be many colonels or accounting consultants who have explicitly embraced the Kantian, Act Utilitarian, or, for that matter, any one "supreme principle of morality." Nevertheless, there are, I venture, plenty of people who throughout their lives have done their best to determine what is right but, measured by the standard of the Categorical Imperative, have failed. In the determination of an action's possible moral worth, Kant discounts the effects an agent's action actually has in the world, apparently on the grounds that these effects are beyond her control. Nevertheless, at least in the *Groundwork*, he does not discount mistaken moral judgments, even though, when an agent makes her best effort to get it right, these also seem to be beyond her control. It is this questionable asymmetry that the cases of Stram and the colonel were designed to bring into focus.

This asymmetry in Kant's view stems from his limitation of actions done from duty to ones that, by the standard of the Categorical Imperative, are morally permissible. I suggest that we reject this limitation – that we acknowledge that a morally impermissible action can be done from duty and thus can have moral worth.

6.7 Moral Permissibility and Moral Worth in the *Metaphysics of Morals*

In fairness to Kant, we should note that in the *Metaphysics of Morals* he moves toward, though he does not explicitly embrace, the possibility of morally impermissible actions having moral worth. Evidence that he makes this move emerges in his discussion of conscience.[19] For Kant, conscience is practical reason's capacity to hold a person's duty before him, judge him on whether he has abided by it, and even punish him for his failures to do so. Conscience is an internal court where a person's practical reason, in its capacity as judge,

renders a verdict on her deeds. Practical reason renders this verdict on the basis of "the law," namely the Categorical Imperative (MS 438). According to Kant, everyone has a conscience. When we refer to someone's having none, what we mean (or should mean) is that this person never *heeds* his conscience (MS 400). Kant goes on to say:

> [W]hile I can indeed be mistaken at times in my objective judgment as to whether something is a duty or not, I cannot be mistaken in my subjective judgment as to whether I have submitted it to my practical reason (here in its role as judge) for such a judgment. . . . [I]f someone is aware that he has acted in accordance with his conscience, then as far as guilt or innocence is concerned nothing more can be required of him. It is incumbent on him only to enlighten his *understanding* in the matter of what is or is not duty. (MS 401)

Here Kant makes an acknowledgment that, in my view, would have been welcome in the *Groundwork*: without being led astray by his inclinations, an agent can make an error in determining what his duty is. Kant does not tell us explicitly how such an error occurs. It is evident, nevertheless, that the mistake gets made at the level of *applying* the fundamental standard of moral judgment (i.e., "the law") rather than at the level of determining what this standard might be. For Kant, the law on the basis of which each person's conscience reaches its verdict *just is* the Categorical Imperative.

Not only does Kant here acknowledge that an agent's acting contrary to duty might stem simply from an error in principle application, but he also suggests that when it does, the agent is not morally blamable. When an agent is heeding his conscience, that is, doing what he believes the Categorical Imperative to prescribe, he is not morally at fault for acting contrary to duty. In the *Groundwork*, Kant makes no such statement, nor is it clear that he would be amenable to it. As we have seen, Kant is there at pains to emphasize the ease with which each agent can determine what he (morally) ought to do. This passage from the *Metaphysics of Morals* indicates a change in Kant's tone, and it seems to mark a shift in his doctrine as well.

But how great a shift? That a conscience-abiding agent incurs no moral *guilt* in performing an undutiful action does not in itself entail that this action can have moral *worth*. Kant neither states nor plainly implies that an action done contrary to duty can have moral worth. Nevertheless, at least on a charitable interpretation, he seems to be leaning in this direction. It is hard to see what plausible grounds Kant could offer for granting that a person is not morally blamable for a conscience-abiding, undutiful action, yet denying that such an action *could have* moral worth. It is one thing to hold, as Kant might in the *Groundwork*, that we are always morally accountable for acting in a morally impermissible way and, on this basis, to conclude that morally impermissible actions cannot be from duty and thus cannot have moral worth. It would be quite another to acknowledge that we are

sometimes not at all accountable for acting contrary to duty, yet to cling to a view that, in effect, rules out the possibility of an action contrary to duty having moral worth. Recall that in his discussion of moral worth in the *Groundwork* Kant appeals to our intuition that an agent's action is not to be disqualified from having moral worth by anything we take to be outside of his control. By claiming in the *Metaphysics of Morals* that a person is not always morally blamable for acting contrary to duty, he is, in effect, admitting that whether an agent acts contrary to duty can be determined by factors outside of his control – for example, how adept he is at applying the Categorical Imperative. But now suppose that an individual not only acts in accordance with his conscience but does so for its own sake. He obeys his conscience simply because it is the right thing to do. It seems clear not only that this person might end up acting contrary to duty but that her doing so could stem from conditions over which she has no real power. In this case, the very intuition to which Kant appeals in the *Groundwork* would direct him not to disqualify the agent's action from having moral worth. By acknowledging that when he abides by his conscience, an agent can blamelessly violate his duty, Kant sets himself on a path toward the view that morally impermissible actions can have moral worth.

This discussion does not, however, allow us to conclude that the Kant of the *Metaphysics of Morals* would agree entirely with my description of cases where moral worth is at issue. I suspect that he would concur that even if Colonel Mikavitch acted contrary to duty, her action could have moral worth. According to the example, she had the Categorical Imperative in view, and if she acted contrary to it, it was owing solely to a mistake in applying it. But Stram is a different matter. Recall that Kant conceives of conscience as an internal court where a judge renders a verdict on each person's deeds on the basis of "the law" – that is, the Categorical Imperative. According to my description, however, the judge presiding over Stram's internal court seems to have based his decisions not on the Categorical Imperative but rather on a principle of utility. In the spirit of his *Groundwork* contention that ordinary human reason always has the Categorical Imperative in view, I believe that Kant would reject this description. He would, I think, insist that in cases where an agent purposefully acts in accordance with a rival practical principle, but contrary to Kantian duty, he has failed to heed his conscience. And Kant, of course, would not excuse such a lapse. On his view, conscience simply is the court of the Categorical Imperative.

Kant moves toward recognizing the moral worth of some morally impermissible actions, namely those which stem from errors in applying the Categorical Imperative. Nevertheless, he remains steadfast in his denial of moral worth to actions whose moral impermissibility would seem to stem from errors in choosing a moral standard.[20] I believe this denial to be contrary to ordinary (and better) moral judgment. Actions stemming from moral

standards other than Kant's can have moral worth, and can thus express what, intuitively speaking, we might call a good will.

6.8 The (Alleged) Transparency of Moral Requirements

Let me now turn to two objections to my claim that Kant should acknowledge that some actions contrary to duty can be done from duty, and thus that among actions that have moral worth, we might find some that clash with moral requirements. First, whatever intuitive force the claim has derives largely from the following notion: however hard an agent tries to do what is right, she might actually end up doing something that conflicts with Kantian duty. But as we reminded ourselves, Kant rejects this notion. In the *Groundwork*, he says:

> [W]e have arrived, within the moral cognition of common human reason, at its principle, which it admittedly does not think so abstractly in a universal form but which it actually has always before its eyes and uses as the norm for its appraisals. Here it would be easy to show how common human reason, with this compass in hand, knows very well how to distinguish in every case that comes up what is good and what is evil, what is in conformity with duty or contrary to duty, if, without in the least teaching it anything new, we only, as did Socrates, make it attentive to its own principle; and that there is, accordingly, no need of science and philosophy to know what one has to do in order to be honest and good, and even wise and virtuous. We might even have assumed in advance that cognizance of what it is incumbent upon everyone to do, and so also to know, would be the affair of every human being, even the most common. (GMS 403–404)

Here Kant implies that he would deny the possibility of an agent's trying her best (without succumbing to inclination) to do what is right, yet erring either in choice or application of moral standard. Of course, the principle of "the moral cognition of common human reason" to which Kant refers is the Categorical Imperative. Ordinary reason, he believes, has (a version of) the Categorical Imperative always in view. In light of this belief, it is hard to see how, for Kant, a completely sincere and dedicated inquirer could embrace any moral standard other than this imperative. Kant also here seems to reject the idea that someone who had embraced the Categorical Imperative could, his best efforts notwithstanding, misapply this principle. With the compass of the Categorical Imperative in hand, says Kant, ordinary reason "knows very well how to distinguish in every case that comes up . . . what is in conformity with duty or contrary to duty."

In response, suppose for a moment Kant is right in claiming that ordinary reason uses the Categorical Imperative (perhaps in a folksy form) as the standard of moral judgment. It is, nevertheless, a strain to deny that even with the best of intention and effort, we might fail to apply this standard

correctly. After all, haven't we behind us two hundred years of scholarly disagreement on how to employ the Categorical Imperative test?[21] Second, if ordinary reason always employed the Categorical Imperative as the standard of moral judgment, then Kant might have grounds for insisting that anyone who sincerely tried to determine what was right would not embrace or act on any other principle. But it is easy to be skeptical as to whether ordinary reason does employ exclusively something like the Categorical Imperative as the standard for moral judgment. Granted, in contemplating whether to do something, we sometimes ask ourselves the roughly Kantian question, "What if everyone did that?" But we also sometimes pose the roughly utilitarian one: "If I did this, how would it affect the well-being of those I care about?" The passage in question does little to undermine the possibility that a sincere and strong-willed moral inquirer might, either by misapplying the Categorical Imperative or by correctly applying some other principle, act in a way that conflicts with Kantian duty.

At this point one might object that I have not really focused on the crux of Kant's notion that, unless he is swayed by his inclinations, an agent would not embrace a moral standard other than the Categorical Imperative. Kant holds that this imperative is valid (unconditionally binding on all of us) and that it has its source in reason alone.[22] If he is right, goes the objection, then the Categorical Imperative would obviously present itself as the standard of moral judgment to every being who possesses reason. Every such being would legislate this imperative to herself by virtue of the very cognitive equipment she possessed. She couldn't help but embrace it as the standard of moral judgment.

In this objection we find the beginnings of an explanation of Kant's view that no agent could in a sincere quest to discover his duty embrace a moral standard other than the Categorical Imperative. However, we find no genuine justification of the view. Granted, if the Categorical Imperative were valid and had its source in reason, we might have license to conclude that it would be recognized by all of us as the standard of moral judgment.[23] (I say "might" because there seems to be no guarantee that reason is transparent in the requisite sense. That an agent is obligated by her own reason to obey the Categorical Imperative would not in itself entail that she would realize that she is. An agent could conceivably fail to discern what her own reason demands. Why should the transparency of practical reason be taken for granted?) At any rate, the truth of the claim that if the Categorical Imperative were valid and had its source in reason, it would be recognized by all sincere inquirers as the standard of moral judgment fails to justify the view that, actually, all sincere inquirers do recognize it as this standard. In the second *Critique*, Kant himself seems to acknowledge that he does not prove the Categorical Imperative to be valid and to have its source in reason (see section I.3). Moreover, consider a parallel. Suppose that the following claim is true. If an Act Utilitarian principle were valid and had its source in human

reason and intuition, it would be recognized by all sincere inquirers as the standard of moral judgment. The truth of this claim would fail to justify the view that, actually, all sincere inquirers do recognize an Act Utilitarian principle as this standard. What we might be able to conclude if we had a proof of the validity and origins of a moral principle has little relevance to what we have grounds to believe now, in the absence of such a proof.

6.9 Odious Actions and Moral Worth

Let me now turn to a second objection to my suggestion that, even if the Categorical Imperative is indeed the supreme principle of morality, the moral worth of an action should not turn on its Kantian moral permissibility. If we conclude that Colonel Mikavitch's and Stram's actions have moral worth, goes the objection, then we are rationally compelled to admit that any action that could be done from duty can have such worth. But there are actions done from duty that are so odious that we are unwilling to grant them any moral value.[24] In *Eichmann in Jerusalem*, Hannah Arendt writes of the Categorical Imperative in the Third Reich, a principle that was apparently known to some Nazis: "Act in such a way that the Führer, if he knew your action, would approve it."[25] If we remove moral permissibility as a condition for moral worth, then we are forced to conclude that an agent's acting from duty on an imperative such as this would have moral worth. Kant's *Groundwork* account enables us to avoid this unwelcome and disturbing conclusion.

In response, if I am correct in finding an untenable asymmetry in Kant's view of what affects an action's moral worth, then Kant's account does not really give us any philosophically plausible means for avoiding this conclusion at all. However, the question is whether Kant's account, revised in the way I have suggested, can meet the objection. To a large extent, I think it can.

The key to seeing this is to understand what in Kant's view it means to act from duty. In the preceding chapter (5.3) we found that an action is done from duty only if two conditions are met. First, the agent's incentive for acting must stem from the notion that the principle (represented by the agent as a law) requires the action. Second, the agent's notion that the action is morally required must itself provide *sufficient* incentive for him to perform it. In sum, the agent's (in itself sufficient) incentive for acting must stem from the notion that a principle (represented by the agent as a law) requires the action.

The discussion in this chapter enables us, I believe, to make explicit two further conditions on acting from duty. First, as we found in section 6.6, an agent must do his best to realize the end of his action. Expressing a good will through acting from duty involves "the summoning of all the means in our power" to realize our aim. If an agent holds breaking promises to be forbidden by a principle she represents as a law, she would not count as

willing, from duty, to keep a promise unless she made every (in her view morally permissible) effort to do so. Kant finds this condition in ordinary moral reason.

I think we also find in ordinary moral reason an additional condition on acting from duty, namely that an agent must make a genuine effort to determine what her duty is. At least part of why we think an agent's making a halfhearted attempt to attain her end to be inconsistent with its being from duty (and thus having moral worth) is that her action betrays a lack of commitment to doing what is morally required in the case at hand. An agent's failure to make a genuine effort to determine just what her moral duty is betrays a similar lack of commitment. If someone is really interested in *doing* what is morally required, then she must take an active interest in finding out *just what is* morally required. She need not delve into casuistry before every action, but she needs to act against the background of reflection on the moral status of her action, that is, act against the background of what I call conscientious reflection. Since, I venture, it belongs to our everyday concept of an action done from duty that it express a commitment to doing what is morally required (at least in the case at hand), we hold that no action that fails to express such a commitment can be done from duty. We do not allow factors that, intuitively speaking, we hold to be beyond an agent's control to preclude her action from having moral worth. Factors within her control, however, are a different matter. And among these factors we find not only the agent's effort to realize the ends of morally required actions but also her effort to determine just which actions these are. In the spirit of Kant's view, if an action fulfills each of the four conditions we have just sketched, then it has been done from duty and thus has moral worth.

Returning to the objection, I doubt very much whether someone acting in accordance with the Nazi perversion of the Categorical Imperative would fulfill all four conditions we have isolated for acting from duty. In particular, I find it far more likely that slovenliness, rather than sincere effort at reflection, would result in a person's embracing this principle. Moreover, I doubt very much that the agent's (in itself sufficient) incentive for acting would really lie in the notion that this principle, represented by him as a law, required the action. It is, I think, much more likely that greed or ambition would constitute the grounds of his action. In the case of Eichmann, these doubts seem to be confirmed. However, I cannot prove it to be impossible that in performing an odious action, someone might fulfill each of the conditions in question, thereby giving his action moral worth. Acknowledging the possibility of odious actions having moral worth is painful. Yet I see no way of avoiding it while, at the same time, defending a coherent reconstruction of Kant's views.

At this point, someone might object that I have overlooked a very Kantian way of avoiding this disturbing conclusion. In the *Religion within the Limits of Reason Alone*, Kant considers the case of a religious inquisitor "who clings

fast to the uniqueness of his statutory faith even to the point of [imposing] martyrdom, and who has to pass judgment upon a so-called heretic (otherwise a good citizen) charged with unbelief" (Rel 186, English ed. 174). Suppose that the inquisitor condemned the heretic to death and that this action was, by Kant's standard, morally impermissible. Here, it might seem, we have a case in which a morally impermissible action could have moral worth. For it appears that the inquisitor's action might meet each of the Kantian conditions we have discussed: he might have arrived through sincere moral reflection at his action of condemning the heretic, he might have had as his incentive for condemning him the notion that doing so was required by a universally binding principle, and so forth. In short we seem to have just the sort of case the objection worries about – an odious action that (if my view is correct) we must acknowledge to have moral worth.

One might think that Kant's own reaction in the *Religion* to a case such as this provides us with a way of avoiding this acknowledgment. Kant suggests that the inquisitor's action could not meet these conditions. In Kant's view, sincere moral reflection leads an agent to the view that he must be *sure* (*gewiß*) that an action he proposes to perform is right before he performs it. Kant calls this view a "postulate of conscience" (*Postulat des Gewissens*; Rel 186, English ed. 174). However, the inquisitor cannot, Kant argues, sincerely reach the conclusion that he is sure of the rightness of his condemnation. Apparently, Kant holds that earnest reflection would lead the inquisitor to the conclusion that it is wrong, based on a man's religious faith, to deprive him of his life – unless the divine will has ordered it (Rel 186–187, English ed. 175). If the inquisitor believed that the divine will had indeed made such an order, his belief would be based either on what he took to be his personal communication with God or on divine doctrine revealed to someone else. In either case, Kant argues, the inquisitor could not sincerely come to the conclusion that he was *sure* that the condemnation was ordered by God and thus right.[26] Therefore, the inquisitor's condemnation of the heretic to death could not, in Kant's view, be the result of sincere moral reflection, nor, it seems, could the inquisitor be motivated by the notion that his action was morally required. If Kant's views regarding this "postulate of conscience" are correct, then we can say that, appearances to the contrary, the inquisitor is really not acting from duty.

Although I believe that actions like the inquisitor's would almost always fail to meet the Kantian conditions for moral worth, I am not convinced that Kant's considerations here *prove* that they *could never* meet them. First, although I think it highly unlikely that the inquisitor would sincerely conclude that he was sure of the condemnation's rightness, I do not think it to be impossible that he would. We cannot totally discount the possibility that, even after earnest reflection, he takes it to be certain that the condemnation was commanded by God. Second, I do not believe it to be obvious that all sincere, morally reflective agents would embrace Kant's postulate

of conscience. Serious moral reflection might lead an agent to the view that, ideally, one would be sure that an action is right before he did it, but that, sometimes, given the complexity of the moral landscape, one is *forced* to choose between actions none of which one holds with certainty to be right.[27] Appeal to Kant's "postulate of conscience" would not enable us to avoid the painful conclusion that being faithful to the spirit of Kantianism – and, I believe, of ordinary views regarding moral worth – requires us to admit the possibility (though by no means the likelihood) that some terribly wrong actions have moral worth.

6.10 Sympathy and Moral Worth

Against Kant's official view (at the very least his view in the *Groundwork*), I have argued that some actions not in accordance with duty can be done from duty. The logic of Kant's own position, I have contended, requires him to acknowledge this. Kant's claim that all actions from duty have moral worth must in the end be understood to allow that some morally impermissible actions can have such worth. I have tried to defend Kant's claim understood in this way.

But there is a far more familiar criticism of Kant's views regarding moral worth that warrants attention. Kant claims not only that all actions from duty have such worth but that only such actions have it. The criticism is that actions from other motives, typically from sympathy, compassion, and the like, have moral worth. A full treatment of the moral worth (or lack thereof) of acting from such motives is beyond the scope of my project. However, as I try to explain regarding the motive of sympathy, some specifications of this objection seem to have force, whereas others do not.

To begin, we need a rough idea of what critics of Kant mean by acting from sympathy. On one critic's account, namely that of Lawrence Blum, acting from sympathy amounts to acting from an emotion that has three elements.[28] First, it has a cognitive element. Sympathy is intentional; it is directed at another's weal or woe. If an agent has sympathy for another, then he believes that she is in a certain state (e.g., one of suffering). Second (and obviously), sympathy involves feeling. To have sympathy for a person, an agent must at least sometimes be in a certain affective state regarding her (e.g., pained at her suffering). Third, sympathy has a conative element. If an agent has sympathy for another, then he wants to help the person for her own sake. To act from sympathy is to act from this emotion. It typically involves thinking that another is suffering, feeling distress at this suffering, and trying to help the other for her own sake. Acting from sympathy alone does not involve any reflection on the moral status – for example, the moral permissibility or even virtuousness – of one's action.

In his well-known *Groundwork* I discussion of the sympathetically attuned person, Kant does not offer a precise account of what is involved in acting

from sympathy. It might seem clear that what he does say is incompatible with that suggested by his critics. Kant holds acting from sympathy to be a kind of acting from inclination (GMS 398). If my reading of acting from inclination is correct (see sections 1.6–8), his holding this entails that, in his view, all of an agent's acting from sympathy is conditional on his belief that it will have some hedonic payoff for the agent, for example, in relieving his pain at seeing another suffer. Some might think that no action thus conditioned is really done for the sake of another person. If you genuinely act for another's sake, the idea goes, then your expectation of a hedonic payoff for yourself does not enter into your motivation. But I am not convinced. Sometimes, at least, a particular action can be conditional yet done for the sake of another. Suppose for example that given my current financial goals, I decide that my birthday gift to a friend must meet a certain condition: it must cost less than fifty dollars. It seems that, nevertheless, I might buy the present for the friend's sake. Analogously, my helping a stranger might be conditional on my expectation of hedonic benefit for myself – for example, the disappearance of my pain of seeing him suffer – yet its being so seems compatible with my helping the stranger for his own sake. For it does not seem to prevent me from having as one of my ultimate ends to improve the stranger's condition.[29] Since Kant does not construct a detailed account of what it means to act from sympathy, it is hard to determine the extent to which his basic concept of such action diverges from that of his critics. It seems to me, however, that Kant's account is compatible with the notion that an action from sympathy is done for another's sake.

Kant offers several arguments against the view that acting from sympathy has moral worth. According to one specification, this view is very straightforward. If an action is done from sympathy, then it has moral worth; an action's being done from this motive is a sufficient condition for its being morally good. Critics of the Kantian view have not, as a rule, held this view. One critic, for example, denies that it is morally good for a bystander to have sympathy for a corporate criminal's hiding his face from cameras as he is being led to prison. The critic would presumably also hold that it would be devoid of moral worth to act from this sympathy, for example, by trying to block the criminal from the cameras' view.[30] It is easy to generate other cases in which many of us would refuse to grant moral worth to an action done from sympathy. Two members of a band of "ethnic cleansers" are plundering the house of a hated minority. One soldier sees another struggling long and hard to open a glass display cabinet full of delicate antique dolls that the other wants to steal for his girlfriend. From sympathy for the other soldier alone, the one picks the lock. On the face of it, some actions seem to lack moral worth, even if they are done from sympathy. I will have more to say regarding examples such as this. For now, moving from the concrete to the abstract, let us examine some arguments Kant suggests against the view that all actions done from sympathy have moral worth.

First, in the *Groundwork* Kant suggests that if an action is not done from duty, it is done from inclination (section 1.6). Since actions from sympathy are not done from duty, they are done from inclination. Yet there is no guarantee that an action from any particular inclination, including sympathy, will actually be in accordance with duty (see section 5.5). Actions from inclination, including those from sympathy, sometimes conflict with duty (GMS 390, 398; Rel 30–31, English ed. 26). Since only actions in accordance with duty can have moral worth, it is not the case that if an action is done from sympathy, then it has moral worth. This argument rests on the premise that only actions in accordance with duty can have moral worth. But if I am correct, the logic of Kant's own position compels him to reject this premise (sections 6.5–7). That an action from sympathy conflicts with duty does not in itself give Kant legitimate grounds for denying it moral worth.

Kant might locate a second basis for rejecting an action's being done from sympathy as sufficient for its having moral worth in the conditional nature of actions from inclination. As we noted, in Kant's view, all of an agent's acting from sympathy is conditional on her expectation that it will have some hedonic benefit for her. But the desire for pleasure has been foisted upon us by nature. In acting from inclination, even from sympathy, we are (in part) pursuing an end that we have not set ourselves. Only in acting from duty do we manifest the independence from animality that gives our action a special worth (see sections 5.4–5). This argument rests on two very controversial premises. The first is that all acting from sympathy is conditional in the way Kant holds. Kant's critics do maintain that the sympathetic agent acts from an emotion, one component of which is an affective state (e.g., pain at the suffering of others). Yet they would probably not agree that the sympathetic agent's action is *conditional* on the expectation of a hedonic benefit to herself (e.g., the relief of her pain at the suffering of others). The sympathetic agent, the critics might say, would help even if she believed that doing so would, on balance, increase her own suffering by, for example, making her more familiar with the excruciating pain of a burn victim. Since Kant simply sets out rather than argues for his hedonistic account of all acting from inclination, and since this account does not seem to be deeply entrenched in ordinary moral psychology, he would not be on strong ground in insisting the critics are misguided. Another premise on which Kant's argument rests is that moral value accrues only to actions that manifest a greater independence from our sensuous nature than any actions from inclination (see section 5.4). Yet as Kant's critics would surely wonder, why should we consider an action's manifesting this greater independence from sensuous nature to be requisite for its having moral worth? Why place so much importance on it? It would be one thing if Kant actually held that actions from inclination were, from all perspectives, totally unfree. But he does not hold that. According to him, all actions are done on maxims the construction of which involves the spontaneity of the will (section 2.2).

More promising in my view than the first two bases is a third one Kant suggests for rejecting the notion that being done from sympathy is a sufficient condition for an action's having moral value. This basis is that an action done from duty, but not one done purely from sympathy, reflects a commitment to morality. The agent takes the action to be of a kind that is morally required. At some point, though not necessarily at the time of the action, she has judged that it is. She acts against the background of (what I have called) conscientious reflection – that is, thought regarding the moral status of her action. Even if in acting from duty she gets things wrong and violates Kant's moral law, her action expresses conscientiousness – something that cannot be said for an action done from sympathy alone. Although both an agent who acts from sympathy alone and one who acts from duty might actually act contrary to what morality dictates, the latter does so against the background of concern with the moral status of what he does. To the Kantian, such concern is a necessary ingredient in a morally valuable action.

Of course, Kant's critics might object to this use of the notion of moral commitment. Granted, actions from duty necessarily take place against the background of conscientious reflection, whereas actions from sympathy alone do not. Yet it would be question-begging simply to assume that only actions involving conscientious reflection have moral worth. Some virtue ethicists deny that conscientious reflection need play any role whatsoever in a morally valuable action.[31] Why should we not hold that what gives an action moral worth is simply its being done from sympathy?[32]

In answer I can offer only an appeal to the view (which I take to be widely shared) that some actions done from sympathy alone do not have moral worth. The ethnic cleanser's action of trying, from sympathy alone, to help a "blood brother" steal from the home of an ethnic minority is such an action. Yet in light of section 6.9 it might seem suspicious to appeal to such examples here. After all, have I not defended the view that if the ethnic cleanser's action is done from duty, then it has moral worth? Some might hold this view to be every bit as implausible as the view that done from sympathy, his action has such worth. I do not believe that it is, but I must leave it to the reader to decide.

Perhaps the following consideration can help tip the scale in favor of Kantian conscientiousness over sympathy. Given the conditions that must be met for an action to be done from Kantian duty, it seems unlikely that the ethnic cleanser's action would be. It seems not very likely, for example, that the one soldier's incentive for picking the lock on the doll case for his comrade would stem from his notion that a universally and unconditionally binding principle required this action. Yet it seems more likely that such an action would be done from sympathy. Why would it be unusual for an ethnic cleanser to think that his fellow soldier is suffering (he really wants those dolls for his girlfriend), feel distress at his suffering, and try to help him for his own sake?[33] In short, there are possible cases of acting from

duty (in Kant's precise sense) that make it difficult for us to maintain that all acting from duty has moral worth, and there are possible cases of acting from sympathy that make it hard for us to hold that all acting from sympathy has moral worth. However, I believe that possible cases of the latter sort are much more likely to be actual.

If we hold moral commitment and the conscientious reflection that goes along with it to be a necessary ingredient in morally valuable action, then we must reject not only the notion that being done from sympathy alone is a sufficient condition of an action's having moral worth, but also a second, different specification of the critics' view. On this specification, only some actions done from sympathy alone have moral value, namely the ones that are actually in accordance with what we take moral requirements (or moral virtue) to involve. On this specification, acting from sympathy alone to aid a fellow ethnic cleanser to burn down a village mosque would presumably not count as having moral value. But many other actions done from sympathy alone – for example, giving water to a thirsty old man – would count as having it. Kant would acknowledge that such an action "deserves praise and encouragement" but not "esteem" (see GMS 398). For Kantians, if an action is not done against the background of commitment to morality, then it does not have moral worth – regardless of whether it is in accordance with what morality requires.

A third version of the sympathy objection poses a greater challenge to Kant's position. According to it, an action's being done from sympathy does not itself give it moral worth. Yet if, against the background of an overriding commitment to morality, an action is done from sympathy, then it has such worth. An agent has an overriding commitment to morality just in case he acts against the background of conscientious reflection, and if after such reflection he determines that an action is contrary to what he takes to be morally required, he will for this reason refrain from performing it. A couple of points regarding this objection warrant immediate attention. First, it does not deny that actions from duty have moral worth. The objection does not embrace the conclusion that moral worth is to be found only in (some) actions from sympathy. Actually, and this is the second point, acting from sympathy and with an overriding commitment to morality will involve the possibility of reliance on the motive of duty. Suppose, for example, that someone feels sympathy for a relative in need and is inclined to help him. An overriding commitment to morality would require that if aiding him – for example, by falsely testifying to his whereabouts on the night of a robbery – would be contrary to (his understanding of) duty, he must for this reason refrain from doing so. In the absence of any inclination to refrain, he would need to rely on the motive of duty to have sufficient incentive to conform to (his understanding of) morality.

Kant himself denies moral worth to any action done from a motive other than duty. Yet does he have good grounds for denying it to actions done from

sympathy against the background of an overriding commitment to morality? The two arguments of his that we have discussed do not seem to threaten this view. Acting from sympathy against the background of such a commitment no more contingently leads to action in accordance with duty than does acting from duty. In both cases, an agent tries but might fail to conform to morality's demands. Perhaps Kant would claim that actions from sympathy fail to express the high degree of independence from sensuous drives that actions from duty do and, on that basis, deny the former moral worth. As we have noted, however, this claim rests on the premises that all actions from sympathy are conditional on the prospect of a hedonic payoff and that the lesser independence from sensuous drives expressed in such actions itself disqualifies them from having moral worth. The prospect of successfully defending either of these premises seems dim, and I will not try to do so here. Of course, one might be able to develop other Kantian arguments against the view that moral worth accrues to actions from sympathy done against the background of an overriding commitment to morality. But unless one does, it seems that Kant is left with no convincing rebuttal to this view.

I believe that many will be attracted to this view, as am I. Although the view does not here get the detailed attention it perhaps deserves, I would like to discuss one question regarding it. Suppose that someone, against the background of an overriding commitment to morality, acts simply from a desire to relax: he sees a film. At some point, the person reflected on the moral status of this sort of action. If through this reflection he had found that actions like it were morally impermissible, he would, motivated by this finding, have refrained from seeing the film. Although the agent's manner of acting in some sense reflects a good character, it would be odd and, I think, unacceptable to hold that it had moral worth. But is there a basis on which a Kantian could deny that his action had moral worth, yet affirm that some actions from sympathy have such worth, namely those done against the background of the sort of commitment we have been discussing? Of course, it would not do for a Kantian to locate the basis for this in the notion that acting from a desire to relax lacks something that acting from the motive of sympathy has, namely unconditional value. The Kantian denies (correctly, I believe) that acting from sympathy has such value. (At issue is the suggestion that perhaps the Kantian should, nevertheless, allow that an action done against the background of an overriding commitment to morality *and* [at the same time] from sympathy, has moral, and thus unconditional, worth.) There is, I think, a way for the Kantian to distinguish those actions, done against this background, which do have moral worth, from those, also done against it, which lack it. She might simply appeal to ordinary rational knowledge of morals to support the notion that different motives have different value characteristics. It is a feature of sympathy that when an agent acts from it, as well as against the background of an overriding commitment to morality, his action has moral worth. However, it is not,

for example, a feature of greed that when an agent acts from it, as well as against this background, his action has moral worth.

At this point, some further questions might come to mind. Why limit moral worth to all actions done from duty and some actions done from sympathy? Might not ordinary moral consciousness take it as a feature of other motives (e.g., love of God) that when an agent acts from them, as well as against the background of an overriding commitment to morality, his action has moral worth? I would like to acknowledge this possibility. A detailed exploration of reflective moral common sense would be necessary to develop a full list of those motives that, against the requisite background, would produce actions that have moral worth. I do not try to construct such a list here.

6.11 Summary

Kant claims that an action has moral worth if and only if it is done from duty. The logic of Kant's own position, I have found, compels him to affirm that some actions contrary to duty can be from duty, and can thus have moral worth. Against the background of this finding, I have defended one-half of Kant's claim, namely that being done from duty is a sufficient condition for an action's having moral worth. (An agent acts from duty just in case her incentive for acting stems from the notion that a principle, represented by her as a law, requires the action; this notion itself provides sufficient incentive for her acting; she acts against the background of conscientious reflection; and she does her best to realize her action's end.) The other half of Kant's claim I find far less compelling. In my view, Kant does not establish that being done from duty is a necessary condition for an action's having moral worth. Although an action's being performed against the background of a commitment to morality is requisite for its having such worth, its being done from duty might not be. Kant does not successfully rule out the view that actions from sympathy, when performed against this background, also have such worth.

Since he does not, it will not be open to us to appeal in the derivation of the Categorical Imperative to the notion that only actions from duty have moral worth. That it will not be might seem to place the derivation in jeopardy, since this notion is so entrenched as a central Kantian dictum. As I try to show in the following chapters, however, rejecting this notion lessens little if at all the force of the derivation. Key to the derivation is not the notion that only actions from duty have moral worth but that all such actions have it.

7

Eliminating Rivals to the Categorical Imperative

7.1 Aims of the Discussion

On the criterial reading, Kant's derivation of the Formula of Universal Law has three main steps. First, Kant tries to pinpoint the features that we, on reflection, believe that the supreme principle of morality must possess. Next, Kant attempts to establish that no possible rival to the Formula of Universal Law fulfills all of these criteria. Third, Kant tries to demonstrate that the Formula of Universal Law remains as a viable candidate for a principle that fulfills all of them. The third step is discussed in Chapter 8.

The current chapter concentrates on the second: how does (or might) Kant try to eliminate all possible rivals to the Formula of Universal Law? To succeed, Kant would need to prove that no possible rival possesses all of the necessary features of the supreme principle of morality that he has identified. It is doubtful that Kant could do so definitively, for it is hard to see how he could demonstrate that he had actually considered every alternative to the Formula of Universal Law. Nevertheless, if we accept Kant's view of the features that the supreme principle of morality would have to possess, then his argument by elimination has some force. For it does show that certain rivals to the Formula of Universal Law fail to be viable candidates for the supreme principle of morality.

Since in trying to eliminate rivals we will be appealing to criteria Kant develops for the supreme principle of morality, it is helpful to have the criteria in view. According to Kant's basic concept, the supreme principle of morality would have to be (i) practical, (ii) absolutely necessary, (iii) binding on all rational agents, and (iv) the supreme norm for the moral evaluation of action. Moreover, (taking into account the modifications we made to Kant's further criteria in Chapter 6) this principle must be such that: (v) every case of willing to conform to it because the principle requires it has moral worth; (vi) the moral worth of willing to conform to the principle because the principle requires it stems from its motive, not from its effects; (vii) an agent's

representing the principle as a law – that is, a universally and unconditionally binding principle – provides him with sufficient incentive to conform to it; and, finally, (viii) a plausible set of duties (relative to ordinary rational moral cognition) can be derived from the principle. Regarding criterion v, let me make explicit that, on my understanding, willing to conform to a principle because the principle requires it amounts to willing, from duty, to conform to it. It amounts to fulfilling each of the four conditions specified in Chapter 6 for acting from duty (see section 6.9).

As it happens, the rivals to the Formula of Universal Law we discuss (e.g., utilitarian principles) are also rivals to the Formula of Humanity. Even though, as I argued in Chapter 3, Kant's derivation of the Formula of Humanity is unsuccessful, we need not give up on this formula as a candidate for the supreme principle of morality. Kant conducts a criterial derivation of the Formula of Universal Law. But the same technique might be used to good effect with regard to the Formula of Humanity. In this chapter I use the term "the Categorical Imperative" loosely to refer to either formula (even though I do not hold the two to be equivalent), since no differences between the two formulas will come into play.

Among the Categorical Imperative's most pressing contemporary rivals are consequentialist principles. I try to show that some of the criteria for the supreme principle of morality that we have discussed at length serve as the basis for premises in a Kantian argument against consequentialist rivals.[1] This argument applies not only to familiar utilitarian forms of consequentialism (sections 7.3–5), but to less familiar forms of it including "Aristotelian perfectionism" (7.6) and "Kantian consequentialism" (7.7). Moreover, if we share some of Kant's views regarding the moral worth of actions, the Kantian argument against certain consequentialist principles succeeds.

Of course, not all challengers to the Categorical Imperative as the supreme principle of morality are consequentialist principles. With no pretension to exhaustiveness, I consider three that are not: a principle (somewhat similar to the Ten Commandments) in which several prescriptions are conjoined (7.8), and two variations on the Formula of Universal Law (7.9). I argue that if we accept Kant's criteria for the supreme principle of morality, then his arguments against these principles also have considerable force.

Before turning to arguments aimed at specific rivals to the Categorical Imperative, however, we consider an argument through which, with one broad stroke, Kant tries to eliminate all rivals. As I try to show in section 7.2, this sweeping argument is a failure.

7.2 A Sweeping Argument against All Rivals

Kant suggests this sweeping argument in both the *Groundwork* and the second *Critique* (GMS 444, KpV 21–41), albeit in slightly different terminology. First, he contends that all candidates for the supreme principle of morality are material, except for the Categorical Imperative, which is formal. Second, he

claims that no material practical principle could be the supreme principle of morality. Therefore, he concludes, the only viable candidate remaining is the Categorical Imperative. I begin by considering Kant's argument for the second premise, then move to the first.

We have already explored Kant's basis for the second premise of his argument, that is, for the claim that no material principle could be the supreme principle of morality. A material principle, let us recall, is one that an agent has sufficient motive to conform to only if he expects that doing so will result in the realization of some object he desires and that realizing this object will have a hedonic payoff for himself (section 1.8). A formal principle is (in one sense) a principle such that an agent's representing it as a law itself gives him sufficient motive to conform to it (section 4.7). Kant contends that a viable candidate for the supreme principle of morality must be a formal principle. To establish this, Kant must obviously show that the supreme principle of morality could not be a material principle. He tries to do so with the arguments we examined in section 5.7.

Although Kant has a (at least moderately) hedonistic view of material practical principles (section 1.8), we can summarize one of these arguments without appealing to this view. In the summary (which is based on GMS 444), we can see that another way Kant has of putting the claim that the supreme principle of morality must not be a material principle is to say that it must not be a heteronomous one.[2] The starting point of this argument is familiar – namely the notion that, according to the supreme principle of morality's basic concept, it must be unconditionally binding on all of us. It must be a categorical imperative for all rational beings who, like us but unlike perfectly rational beings such as God, do not necessarily conform to it. A material principle is a rule such that an agent has sufficient motive to adhere to it only on condition that, in her view, doing so will enable her to realize some object she desires. (For present purposes let us stop there, without invoking Kant's notion that such principles are hedonically conditioned.) In order for a material principle to be a categorical imperative, each agent must have sufficient motive available to her to abide by it. If each did not, then in some circumstances it would be impossible for her to abide by it, and, therefore, in Kant's view, it would not be binding on her. (An agent cannot have a duty to do something that it is impossible for her to do.)

But now suppose that a material principle specifies the means to realize some object that a particular agent does not desire, for example, greater perfection defined as the development of physical and rational capacities. (Maybe the agent thinks that she is mentally and physically fit enough to lead a rewarding life.) In this case, the agent might find herself without sufficient motive available to her to conform to the principle and thus unable to conform to it. If she did, then the principle would not be binding on her. But principles that, in light of a particular agent's desires, might not be binding on her (i.e., material principles) are obviously not categorical imperatives. Since they are not, they are not viable candidates for the

supreme principle of morality. In the *Groundwork*, Kant crystallizes this argument thus: "Whenever an object of the will has to be laid down as the basis for prescribing the rule that determines the will, there the rule is none other than heteronomy; the imperative is conditional, namely: *if* or *because* one wills this object, one ought to act in such or such a way; hence it can never command morally, that is categorically" (GMS 444). Material practical principles are heteronomous in the sense that their motivational force stems from something outside of the will, something that each rational being does not necessarily have: a desire for some particular object.

However compelling this argument may be, it is not enough to insure the success of Kant's attempt to sweep away rivals to the Categorical Imperative. Even if Kant establishes that the supreme principle of morality must not be material (i.e., the second premise in this attempt), he needs to convince us of the first, namely that all rivals to the Categorical Imperative are indeed material principles. Unfortunately, he does not do so.

In the second *Critique* (KpV 40) Kant sets out a table in which he categorizes rivals to the Categorical Imperative. He distinguishes "subjective" from "objective" principles, and "internal" from "external" ones. Subjective principles are empirical; their content stems from experience. Among such principles Kant mentions that "of education," which, Kant asserts, was advocated by Montaigne. Apparently, this principle derives the content of morality solely from custom. Objective principles are based on reason, or at least purported to be so. Wolff, for example, claimed to base his principle of perfection not on experience but on rational concepts alone. In Kant's scheme, some subjective principles are internal, some external. Montaigne's principle is external in that education stems from outside of the agent, while another subjective principle, that of moral feeling defended by Hutcheson, Kant classifies as internal, apparently since this feeling is internal to the agent. Kant also distinguishes between an objective principle that is external, namely that of the will of God, and one that is internal, that of perfection. In sum, Kant categorizes six rivals to the Categorical Imperative. Of the three whose content stems from within the agent (internal principles), two of these are subjective, the principles of "physical feeling" and of "moral feeling," and one objective, the principle of "perfection." Of the three whose content stems from outside the agent (external principles), two of these are subjective, the principles of "education" and "the civil constitution," and one objective, the principle of "the will of God."

Referring to this table, Kant makes several claims: "all the principles exhibited here are *material*" (KpV 41), "they include all possible material principles" (KpV 41), and "all possible cases are actually exhausted, except the one formal principle" (KpV 39), namely the Categorical Imperative. Each of these claims is controversial. Regarding the first, one might wonder whether a theologian would or need acknowledge that a principle of obeying God's will is necessarily material. Why should he accept the view that an agent has

sufficient motive to obey God's will only if she thinks that doing so will enable her to satisfy some desire she has? Might he not instead maintain that an agent's notion that God wills that she do something itself can give her sufficient motive to do it? Regarding Kant's second claim, one might wonder whether his table actually does include all possible material principles. What about a (loosely) Nietzschean principle, something such as "In order to flourish, you ought to maximize your power"? Although this principle is not among the six Kant lists, he might contend that it does fit into his general schema as, for example, a subjective and internal principle. In any case, Kant's final claim, which amounts to the first step of his sweeping attempt to eliminate all rivals, is the one most obviously vulnerable to criticism.

It is not hard to imagine rivals to the Categorical Imperative that, at least on the surface, are not material principles. Suppose someone defends the following perfectionist principle, MP': "Develop your physical and rational capacities." According to the defender, MP' commands categorically. It does not say: develop your physical and rational capacities, *if* you want or *given that* you want to perfect yourself or be happy or attain some other object. It prescribes that you develop these capacities no matter what you want. In reply, Kant might insist that though the defender does not take MP' to be a material principle, it is indeed one. For an agent could have sufficient motive to conform to MP' only on condition that he expected doing so would enable him to realize some object he desired. Yet, as far as I can tell, Kant offers no argument for this contention. Why couldn't an agent be motivated to conform to MP' simply by representing MP' to himself as an unconditionally and universally binding principle? If an agent's representing the Categorical Imperative to himself as a practical law gives him sufficient incentive to conform to this principle, why couldn't an agent's representing MP to himself as a practical law give him sufficient incentive to conform to that principle?

In the next section, we discuss in detail how Kant might try to eliminate consequentialist principles as candidates for the supreme principle of morality. But we can already see that it will not do to sweep them away on the basis that they are all material principles. Take the utilitarian principle U': "Always perform a right action, one that yields just as great a sum total of well-being as would any alternative action available to you." As far as I can tell, Kant does not establish that an advocate of this principle must acknowledge that an agent's having sufficient grounds to conform to it is conditional on her wanting to maximize well-being or to gain pleasure for herself or anything else. The possibility persists that she finds sufficient grounds for complying with U' in the notion that doing so is morally required.[3] Defenders of a variety of candidates for the supreme principle of morality might refuse to acknowledge their principles to be "material." And Kant, it seems, has no good argument with which to discredit such a refusal.

Yet perhaps we have not looked hard enough. In a chapter entitled "On the Concept of an Object of Pure Practical Reason," Kant describes his

method in the second *Critique*: "[I]nstead of the concept of the good as an object determining and making possible the moral law, it is on the contrary the moral law that first determines and makes possible the concept of the good, insofar as it deserves this name absolutely" (KpV 64). Kant claims that other philosophers failed to adopt this method, a failure that led them into error regarding the supreme principle of morality. They began with an object that they considered to be good (e.g., the perfection of our capacities) and tried to derive a practical principle from that object (e.g., the principle that we are required to perfect our capacities). But by beginning with a concept of the good and then trying to derive a practical principle from it, they condemned themselves to advancing material practical principles – ones that are not suited to be the supreme principle of morality. In Kant's words, other philosophers

> sought an object of the will in order to make it into the matter and the ground of a law (which was thus to be the determining ground of the will not immediately but rather by means of that object referred to the feeling of pleasure or displeasure), whereas they should first have searched for a law that determined the will a priori and immediately, and only then determined the object conformable to the will. Now, whether they placed this object of pleasure, which was to yield the supreme concept of good, in happiness, in perfection, in moral feeling, or in the will of God, their principle was in every case heteronomy and they had to come unavoidably upon empirical conditions for a moral law, since they could call their object, as the immediate determining ground of the will, good or evil only by its immediate relation to feeling, which is always empirical. (KpV 64)

Kant suggests the following claim: if we begin in ethics with a concept of the good and then construct a moral principle that requires the promotion of this good, we must acknowledge that an agent will have sufficient grounds to conform to the principle only if she expects doing so will have some hedonic benefit. In effect, we must acknowledge that the principle is material. If Kant successfully defended this claim, then he might indeed have good grounds for asserting that the rival principles we mentioned earlier were material. An advocate of MP' or U' would likely begin his moral theorizing with the concept of an object as the good: in the former case, perfection; in the latter, the general happiness. However, as the cited passage illustrates, Kant does not really argue for the claim in question. He leaves it strikingly unclear why a principle based on the concept of some object as the good must be such that an agent could have sufficient grounds for conforming to it only if he expected a hedonic payoff from doing so. To insist that such a principle must have this feature seems unfounded.

Perhaps Kant is correct that no material principle could cohere with his basic concept of the supreme principle of morality. However, this claim does not give him a quick route to the elimination of all rivals to the Categorical Imperative. For Kant does not show that all rivals actually are material practical principles.

7.3 The Structure of Act Utilitarianism

In light of the shortcomings of Kant's sweeping attempt to dismiss all rivals to the Categorical Imperative on the basis that they are material practical principles, it makes sense to look for other arguments he might offer against particular competitors.

Let us begin with consequentialist principles, specifically utilitarian ones. In his critical writings in ethics, Kant does not explicitly consider utilitarianism. He mentions "the principle of sympathy for the happiness of others," (GMS 442, note), which he attributes to Hutcheson. And he discusses briefly the possibility that the happiness of others is the object of the will of a rational being (KpV 34). So Kant does seem to entertain the notion that the supreme principle of morality is one that requires us to promote the happiness of others. Against this quasi-utilitarian notion, however, Kant employs the suspicious argument we discussed in section 7.2, one according to which such a principle must be material.[4]

Kant might argue against a utilitarian principle with the help of an appeal to his view that only the good will is unconditionally good. Whatever the supreme principle of morality is, he might claim, it must have something unconditionally good as its "ground." The utilitarian would have to take everyone's being happy as the unconditionally good ground of her principle. But everyone's being happy is not unconditionally good. Since it is not, Kant might conclude, the utilitarian principle could not be the supreme principle of morality. This argument does not seem promising, for Kant fails to establish that everyone's being happy is not unconditionally good (see section 3.7). Does he have any better argument available to him with which to eliminate utilitarianism as a rival?

To answer this question, it is helpful to have a particular utilitarian principle in view.

> U: An action is right if and only if it yields as great a sum total of individual well-being as would any alternative action available.

Amartya Sen has shown that U follows from two separate views:, one is an account of goodness; the other, an account of the connection between goodness and rightness. According to "Outcome Utilitarianism," the goodness of a state of affairs is solely a function of the sum total of individual well-being in it. More precisely, any state of affairs is at least as good as an alternative state of affairs if and only if the sum total of individual well-being in the one is at least as large as the sum total of individual well-being in the other.[5] According to "Act Consequentialism," the rightness of an action is solely a function of the goodness of its consequences. More precisely, an action is right if and only if the state of affairs resulting from the action is at least as good as each of the alternative states of affairs that would have resulted respectively from the alternative feasible acts.[6] In discussing U, we assume that its defender grounds it in Act Consequentialism and Outcome Utilitarianism.

Against the possibility that U could be the supreme principle of morality, Kant has available to him a simple, straightforward argument. U runs afoul of Kant's basic concept of this principle. On this concept, the supreme principle of morality would manifest itself to us (human rational agents) as a categorical imperative. It would be absolutely necessary, prescribing that we ought to act in a certain way, no matter what our particular inclinations might be. However, U just tells us which actions are right. It does not prescribe to us that we ought to do right actions. Strictly speaking, it does not prescribe how we ought to act at all. U does not have the form of a categorical imperative. Of course, U's advocate might simply reject Kant's basic concept of the supreme principle of morality. She might insist that such a principle need not take the form of a categorical imperative, or even that it need not be practical (i.e., something on account of which we can act). The utilitarian might conceive of her principle as a fundamental description of right action and nothing more. Doing this would, however, not threaten the Kantian claim we are considering – namely, that if there is a supreme principle of morality, *in the basic sense of such a principle that Kant employs*, then it is the Categorical Imperative.

Rather than responding to Kant's argument against U by rejecting his basic concept of the supreme principle of morality, the utilitarian can simply give U the form of a categorical imperative:

> U': Always perform a right action, one that yields just as great a sum total of well-being as would any alternative action available to you.

Here the utilitarian has added a further principle to the two from which U was constructed – namely, what we might call the principle of imperative rightness: Always act rightly. The resulting principle U' appears to conform to Kant's basic concept of the supreme principle of morality. It could be a practical, absolutely necessary, universally binding, fundamental norm for moral evaluation of action. How might Kant exclude the possibility that it is the supreme principle of morality?

7.4 Against Act Utilitarianism

To eliminate rivals to the Categorical Imperative, Kant has at his disposal not only the criteria contained in his basic concept of the supreme principle of morality but also the further ones he develops in *Groundwork* I. Using some of these further criteria, it is fairly simple to construct an argument to block the possibility that U' is the supreme principle of morality:

1. Whatever the supreme principle of morality is, your willing from duty to conform to it has moral worth.
2. This moral worth does not stem from any effect of what you do, but rather solely from your willing from duty to conform to this principle.

3. Suppose you held that U' were the supreme principle of morality.
4. You would then have to hold that whether your willing from duty to conform to U' had moral worth depended solely on its effects.

5. Since (according to 1 and 2) the moral worth of your willing from duty to conform to the supreme principle of morality does not stem from its effects, you must conclude that U' cannot be this principle.

Although Kant does not make this argument explicitly, its steps are familiar to us from our exploration of his criteria for the supreme principle of morality.

The argument's first step stems, of course, from a criterion Kant develops in *Groundwork* I in his "first proposition." This principle, he claims, must be such that all and only actions conforming to it because the principle requires it (i.e., all and only actions done from duty) have moral worth. However, step 1 differs from Kant's criterion in two significant ways. Obviously, it invokes not at all Kant's view that *only* actions from duty have moral worth. Moreover, embracing the premise involves rejecting the idea that only actions in conformity with duty can have moral worth. According to step 1, whatever the supreme principle of morality is, your *willing*, from duty, to conform to it has moral worth – even if, as it turns out, you fail to conform to it.

The second step is closely related to a further criterion for the supreme principle of morality that Kant develops in *Groundwork* I (in his "second proposition"). It follows from this criterion that we cannot affirm a principle to be the supreme principle of morality unless we can hold that the moral worth of any actions conforming to it from duty does not stem from the actions' effects. The main thrust of step 2 is the same as that of this criterion, namely that the moral worth of an action does not stem from its effects or results. However, in line with 1, step 2 does not restrict moral goodness to actions that actually conform to the supreme principle of morality. It (implicitly) grants that attempts to conform to the supreme principle of morality, even if they fail, can have moral worth.

Step 4 also requires attention. In considering 4 it is important to put ourselves in the position of someone who has, as 3 specifies, accepted U' as the supreme principle of morality. We are assuming, let us recall, that a person who accepts U' as the supreme principle of morality grounds it in Act Consequentialism and Outcome Utilitarianism. Such a person holds that the goodness of a state of the universe is solely a function of the sum total of individual well-being in it: the greater the sum, the better the state of the universe. Now the question arises: according to a defender of U', when would an action done on account of U' have moral value? A defender of U' sees all value (goodness) in terms of individual well-being. Therefore, he must see the value of an action in terms of its effects on individual well-being. It is unclear precisely where he will draw the line between an action

that has a positive value and one that does not. He might, for example, claim that an action has positive value if, on balance, it raises the sum total of individual well-being (rather than diminishing it or having no impact on it). Or he might claim that to have positive value an action must be right – that is, produce just as much well-being as any alternative action. Whatever his particular view might be, for him the value (and thus moral value) of an action is solely a function of its effects.

The proponent of U' might respond that in his view any (positive) value of an agent's conforming to U' "from duty" stems not from the effects the action actually has but from the agent's motive in conforming to U'. The value derives from her willing to conform to U' because, she believes, conforming to it is morally required. So, for example, suppose someone tries to save a stranger who is choking because she believes that morality (in the guise of U') demands it. The value of this action, the proponent of U' might say, is just a function of her motive in doing it, not its effects, for example, not whether she indeed succeeds.

But this response lacks force. Granted, the proponent of U' is not committed to holding that aiding a choking victim has moral value only if it results in the victim's being saved. (Although the victim might die, the example provided to others by the attempt to save him might inspire others to actions of the same sort, and thereby increase the sum of individual well-being.) Yet it is not open to the proponent to derive the value of someone's willing to save the victim solely from his being motivated by U' to do so. The proponent has defined the good in terms of well-being. Given that he has, any value possessed by acting on U' as a motive would stem from its (somehow) promoting the general welfare. It would stem ultimately from its effects.

7.5 Against Expectabilist Utilitarianism

Of course, even if the Kantian argument I have sketched is effective against an act utilitarian principle such as U', it might not work against other varieties of utilitarianism. For example, what about an expectabilist principle? This kind of utilitarian principle some might find most plausible. Would Kant's argument, if we assume that all actions from duty have moral worth, eliminate such a principle as a candidate for the supreme principle of morality? Consider

> EU: Always perform a right action: one that you expect will yield as great a sum total of well-being as would any alternative action available to you.

Now let us suppose that a defender of this principle embraces it partly because, like the defender of U', he endorses Outcome Utilitarianism: he holds that the goodness of a state of affairs is solely a function of the amount

of individual well-being in it. At first it might seem that a defender of EU could easily escape the Kantian argument. For unlike U', it might appear that EU would not run afoul of step 2. The defender of EU would, it seems, not be committed to the view that the value of conforming to this principle depended on its effects. He might coherently claim, for example, that an agent's action has moral value just in case she does what she expects would maximize well-being because she takes that to be the right thing to do. To have moral value, her action need not have the result of actually maximizing, or even promoting, well-being. Therefore, it seems, the defender of EU is not forced to reject step 2 and thus does not fall prey to the argument.

In response to this challenge, I want to argue that, actually, this defender of EU cannot coherently hold that the moral value of conforming to EU from duty does not depend on its effects. The most efficient way to make this argument is with the help of a thought experiment. Imagine an agent who has always conformed to EU because he has taken it to be morally required that he do so. Nevertheless, each of the agent's actions has diminished the sum of individual well-being, even though there have always been actions available to him that would have promoted it. Various factors are responsible for this phenomenon. Sometimes, his best efforts notwithstanding, the agent, who is no expert in psychology, economics, or probability theory, developed irrational expectations of the effects of a proposed course of action on the general welfare, and, as luck would have it, things went just as an expert would predict. At other times, the agent's expectations corresponded with those of the experts, but the world simply failed to cooperate. He expected that praising his colleague would make him feel better, but it actually plunged the colleague deeper into depression. He expected that his giving to a famine relief fund would reduce the suffering caused by starvation, but it actually ended up providing food for a paramilitary unit who ransacked a peaceful village. Not even the example the agent set for others by his unwavering conformity to EU had a positive effect on the sum of individual well-being. Taken individually and taken as a whole, his actions neither directly nor indirectly increased the sum total of individual well-being but actually decreased it (though actions available to him would have increased it).

Our defender of EU as the supreme principle of morality would not be justified in holding that the agent's actions had moral value. The defender embraces Outcome Utilitarianism. He holds that the goodness of a state of affairs is solely a function of the sum of individual well-being in it. Ultimately, his only basis for judging that an action is good is that it have a positive effect on this sum (perhaps relative to other available actions). However, the agent's actions do not have a positive effect on this sum (even relative to other available actions). Since they do not, the defender has no basis for saying that they are good. Despite initial appearances, the defender of EU is committed to the view that, contrary to step 2, the value (including the moral value) of an agent's actions does depend on their effects. Therefore,

as the thought experiment illustrates, the defender is committed to the view that an agent's action of conforming to EU because he thinks it to be the right thing to do can fail to have moral value, thus contradicting step 1.

Some philosophers will not be satisfied with this response, insisting that it neglects an important distinction between evaluation of actions and that of states of affairs. While the defender's embracing of Outcome Utilitarianism requires him to judge the goodness of a state of affairs solely in terms of the sum of individual well-being in it, his embracing of it does not require him to judge actions solely in these terms, they will say. He is free to judge the goodness of actions independently of their effects, in agreement with step 2 and thus ultimately with step 1. That the defender defines the goodness of a state of affairs solely in terms of well-being does not entail that he must define the goodness of actions simply in terms of their production of well-being.

This reply does not seem convincing, as a further thought experiment may show. Imagine two worlds. World I is that of our unfortunate agent from the previous example – the one who, "from duty," always conforms to EU, but whose actions never have a positive effect on the sum of individual well-being (though actions available to him would have such an effect). Let us suppose that in this world at a particular time (t), the sum of individual well-being is ten units. In World II, the sum of individual well-being at t is also ten units. World II is just like World I except that in it our agent's motive for conforming to EU has never been that he takes it to be morally required to do so.

Now let us return to the defender of EU as the supreme principle of morality. The proposal on the table is that the defender hold the following. Although the value of a state of affairs is solely a function of the sum total of individual well-being in it (Outcome Utilitarianism), the value of actions is not. But I do not see how the defender can coherently hold this. On a straightforward understanding, a state of affairs is simply a state of the universe at some particular time. The defender would have to hold that the state of affairs (World I at t) has greater value than the state of affairs (World II at t), since more good actions have been performed in the former than in the latter. But in holding this, he would be betraying his commitment to Outcome Utilitarianism. For, according to this doctrine, the value of a state of affairs is solely a function of the sum total of individual well-being in it. Therefore, according to Outcome Utilitarianism, the value of World I and World II would be identical. This reply depends on the observation that the actions that have been performed at t constitute a part of the state of the universe at t.[7] Actions that have been performed are an element in a state of the universe. In order to rebut my reply, a philosopher would have to deny this – in my view, very plausible – account of states of affairs.

With help from two of the criteria he develops for the supreme principle of morality, Kant can construct a strong argument against one version of expectabilist utilitarianism. But another version of expectabilism seems not to be vulnerable to this argument.[8] Suppose someone defends the principle

EU as the supreme principle of morality but does not embrace Outcome Utilitarianism. She holds that an agent performs a right action just in case he does something that he expects will yield as great a sum total of well-being as would any alternative action available to him. Moreover, the defender affirms that an action is good if and only if it is right. (She acknowledges, of course, that under certain circumstances an action that she calls good diminishes well-being relative to other available actions.) The defender can coherently claim that the moral worth of an action does not depend on its (actual) effects. She can also coherently claim that each instance of willing, from duty, to conform to EU has moral worth. In her view, an action has moral worth just in case it conforms to EU (or, equivalently, just in case it is right). And presumably every case of willing from duty to conform to EU will be a case of conforming to it.[9]

An argument advanced in Chapter 6 supplies the basis for a Kantian response to this version of expectabilist utilitarianism. This response, which I merely sketch, emerges from discussion of, but does not appeal to, Kant's criteria for the supreme principle of morality.[10] In Chapter 6, I defended the view that Kant should acknowledge that some actions contrary to the Categorical Imperative have moral worth. Suppose an agent has done his best to figure out what the supreme principle of morality is but has become convinced that it is something other than the Categorical Imperative. If, from duty, he wills to conform to this other principle but violates the Categorical Imperative, Kant should nevertheless acknowledge that his action has moral worth. He should acknowledge this (roughly) because, intuitively speaking, the factors that are requisite for moral worth are present. The agent's incentive for the action stems from the notion that it is required by an unconditionally and universally binding principle; he holds that the action's being morally required itself gives him sufficient incentive for the action, and so forth. The same sort of argument applies to the version of expectabilism in question. Its defender is committed to the following view. The *only* actions that are good (and thus the only ones that have moral worth) are those that conform to EU. But having done his best to discover the supreme principle of morality, someone might conclude that it is something other than EU. If, from duty, this person wills to conform to this other principle but violates EU, then the defender of this version of expectabilism must hold that the person's action is devoid of moral worth. But, intuitively, I think we would want to attribute moral worth to his action. And that is a reason for rejecting this version of expectabilist utilitarianism. The supreme principle of morality must be such that its defender can coherently claim that all instances of willing from duty to conform to it have moral worth, suggests Kant. The defender of this version of expectabilist utilitarianism *can* coherently claim this. However, she cannot hold something that many of us take to be intuitively clear – namely, that some actions done from duty do not conform to EU, and that *these* actions have moral worth.

It is, however, possible to conceive of moral theories, which some might call utilitarian, that elude even this argument. For example, a philosopher might defend a principle discussed earlier, U', yet not advocate it even partly on the basis of Outcome Utilitarianism (the doctrine according to which goodness is solely a function of well-being). The philosopher might hold that each agent ought always to perform a right action: one that yields just as great a sum total of well-being as would any alternative action available to him. Yet she might divorce the question of an action's rightness from its goodness. She might hold that an action has moral worth just in case an agent does it solely because he takes it to be morally required, regardless of whether his action is right. To rebut this sort of theory a Kantian could, of course, claim that U' fails to fulfill criterion viii for the supreme principle of morality – that it fails to generate a set of moral prescriptions that coheres with ordinary moral thinking.[11] But I do not defend this claim here.[12]

7.6 Against Perfectionism

The two preceding sections focused largely on a type of argument that appeals to Kant's notions (roughly) that all actions from duty have moral worth and that this worth does not depend on the actions' effects. A shortcoming of this type of argument is that, as we just noted, it fails to apply to some forms of utilitarianism (though I think Kant does have other recourse against these forms). A strength of this type of argument is that it applies to some nonutilitarian principles. Recall MP', "Develop your physical and rational capacities." This is a principle of what Thomas Hurka calls "Aristotelian perfectionism."[13] A proponent of MP' as the supreme principle of morality identifies human perfection as the good. He embraces what we might call "Outcome Perfectionism," the view that the goodness of a state of affairs is solely a function of the sum total of individual perfection in it. To will from duty to conform to MP' would presumably involve trying one's best to develop one's physical and rational capacities. It is easy to see how the argument we deployed against utilitarian principles would apply to MP':

1. Whatever the supreme principle of morality is, your willing from duty to conform to it has moral worth.
2. This moral worth does not stem from any effect of what you do but solely from your willing from duty to conform to this principle.
3. Suppose you held that MP' was the supreme principle of morality.
4. You would then have to hold that whether your willing from duty to conform to MP' had moral worth depended solely on its effects.

5. Since (according to 1 and 2) the moral worth of your willing, from duty, to conform to the supreme principle of morality does not stem from its effects, you must conclude that MP' cannot be this principle.

In light of our exploration of Kant's argument against utilitarian principles, the only step we need consider here is 4. Since a proponent of MP' as the supreme principle of morality identifies the good (including the moral good) with human perfection, he must judge an action's moral worth to be a function of its effects on human perfection. Yet the effects of an action (e.g., an increase in an agent's physical perfection) are obviously not identical with the action itself (e.g., an agent's willing to develop his physical capacities).[14] Although a person might will, through a strenuous exercise regimen, to develop his physical capacities, he might injure himself in the process. His good faith attempt to get himself in shape might do nothing but diminish his health and vigor. In this case, a proponent of MP' as the supreme principle of morality would be committed to denying moral worth to the agent's attempt. MP' falls prey to the same sort of Kantian argument that applies to some utilitarian principles.

If we conceive of a consequentialist moral principle as one according to which the moral value of an action depends on its effects, then both U' and MP' count as consequentialist. Moreover, it is evident how Kant might appeal to his account of ordinary moral reasoning to argue against any such consequentialist principle's being the supreme principle of morality. He would simply invoke steps 1 and 2 as they appear in the arguments against these two principles. I call this kind of argument "valuational," since it turns on the question of an action's moral value or worth.

7.7 Kantian Consequentialism?

According to David Cummiskey, Kant's first and second propositions do not conflict with consequentialism.[15] The argument of *Groundwork* I does not really threaten the notion that the supreme principle of morality is consequentialist. Contrary to Cummiskey, I have found in these propositions the basis for a Kantian argument against three forms of consequentialism: act utilitarianism, expectabilist utilitarianism (in one version), and perfectionism. Cummiskey proposes a different "Kantian" form of consequentialism. In this section, I try to show that the Kantian argument also applies to Cummiskey's Kantian consequentialist candidate for the supreme principle of morality.

Cummiskey's detailed statement of his candidate is very lengthy. In the end, though, he suggests that the candidate amounts to (roughly) the following:

> KC: Maximally promote two tiers of value: rational nature and happiness, where rational nature is lexically prior to happiness.[16]

In a nutshell, KC enjoins first that we must maximally promote the conditions necessary for the rational choice of ends, conditions such as liberty and life.[17] The principle thus entails that if we find that the only way to save

two rational agents is to kill one innocent agent, then we are required to kill him.[18] Second, KC enjoins that we must maximally promote the effective realization of rationally chosen ends.[19] The first requirement is lexically prior to the second requirement in the Rawlsian sense that we are to fulfill the second only if we have completely fulfilled the first: we are to promote happiness maximally only if we have done all we can to promote rational nature.[20] The principle thus entails that we must not kill one person in order to make others happy.

I do not offer a thorough discussion of KC but rather focus only on features of it that are directly relevant to my present aim. First, Cummiskey presents KC as a categorical imperative with a scope extending to all rational agents.[21] It requires all rational agents, regardless of their particular inclinations or desires, to promote maximally two tiers of value. Second, Cummiskey holds KC to be a consequentialist principle in the following sense. It sets out a requirement to promote the good (the two tiers of value), and it does not set limits on the acceptable means that an agent may employ to promote the good.[22] It does not, for example, specify a duty not to sacrifice one innocent person to save two others. Third, Cummiskey holds that KC has a Kantian foundation. A proponent of KC as the supreme principle of morality would, he suggests, defend it *in part* by arguing as follows. If an agent holds there to be a categorical imperative, then he must hold there to be something unconditionally valuable. Upon reflection, he must find that this unconditionally valuable thing is rational nature (humanity). For he must hold rational nature to be the source (i.e., the unconditioned condition) of value, and thus to be unconditionally valuable.[23] We examined (and criticized) this argument in Chapter 3.[24] Whereas Korsgaard and, presumably, other Kantians hold that this argument supports the Formula of Humanity (interpreted as a nonconsequentialist principle), Cummiskey claims that it is better suited to supporting a consequentialist principle such as KC. I do not address the issue of whether, when taken in isolation, the argument is better suited to supporting KC. However, I defend the view that there are Kantian grounds, manifest in *Groundwork* I, for rejecting KC as a candidate for the supreme principle of morality.

KC is subject to basically the same valuational argument as the other principles we have examined. It runs afoul of the argument's first step, according to which every case of willing from duty to conform to the supreme principle of morality (whatever it turns out to be) has moral worth. A defender of KC as the supreme principle of morality has adopted a two-tiered conception of the good. On the higher tier is rational nature; on the lower is happiness. There is, says Cummiskey, "a normative hierarchy in the theory of the good."[25] The goodness of rational nature is such that we are never to refrain from maximally promoting it for the sake of promoting happiness. Imagine a scenario in which a defender of KC as the supreme principle of morality reasonably believes that she has not done all she can to promote rational

nature: to promote the conditions necessary for rational agency. KC, which she considers to be the supreme principle of morality, requires that she maximally promote these conditions. One of the conditions necessary for rational agency is life. After much reflection, the defender concludes that the only way to save two people is to kill one innocent person. From duty, the defender conforms to KC and kills the innocent person. Despite the defender's efforts, however, the other two are also killed. And there are no other morally relevant effects – for example, no one, not even the defender herself, is inspired by her action to strengthen a commitment to conforming to KC. In this scenario, the defender would have to deny that her own action had positive moral value (moral worth), thereby contradicting step 1. For the action did not at all succeed in promoting the good; it did not secure the conditions necessary for rational agency.

In response, Cummiskey would, perhaps, insist that the defender may claim her action to have moral worth even if it does not actually secure the conditions necessary for rational agency. She may claim that moral worth is intrinsic to the action. But I do not see how she may do so coherently. She has identified the good first with rational nature and second with the realization of the objects of rational nature (i.e., happiness). Her action – her killing from duty one innocent to save others – has succeeded not at all in promoting the good in either sense. It has not helped to secure the conditions necessary for rational agency, and it has not helped to secure the realization of rationally chosen ends. So the defender finds herself with no basis on which to conclude her action to have been good.

7.8 Against a Principle Akin to the Ten Commandments

In the preceding sections, we have explored an argument Kant might employ to eliminate consequentialist candidates for the supreme principle of morality. Yet not all rivals to his principle are consequentialist. In this section and the next, we examine how he might eliminate some nonconsequentialist candidates. Kant's successfully excluding these candidates would not itself give him warrant to conclude that no nonconsequentialist rival to the Categorical Imperative remains. In my view, Kant offers no plausible way of guaranteeing that his arguments would be effective against all possible nonconsequentialist principles.

Let us begin with a principle somewhat akin to the Ten Commandments. Why, in Kant's view, couldn't the following conjunctive principle be the supreme principle of morality?

> TC: You ought to honor your father and mother; you ought not to kill; you ought not to commit adultery; you ought not to steal; you ought not to bear false witness; you ought not to covet anything that is your neighbor's.

To simplify matters, let us view TC in a detheologized way, as a conjunctive prescription "legislated" by individuals to themselves. Let us further suppose that TC is a categorical imperative in the sense of a principle that sets out a prescription to all rational agents regarding what they ought to do, regardless of what they might be inclined to do.[26] Obviously, a proponent of TC might conceive of morally permissible actions as ones that conform to TC and morally impermissible actions as ones that do not. Moreover, a proponent of TC might hold that an agent acts from duty when he wills to conform to TC just because, in his view, TC requires that he do so. She might further hold that any action done from duty has moral worth, regardless of its effects. (If from duty someone does his best to honor his parents, his action has moral worth – even if, through some unforeseen chain of events, he ends up dishonoring them.) Mirroring Kant, the proponent of TC might conceive of willing from duty to obey TC (in her view, a good will) to be good without qualification. This principle seems to fulfill much of Kant's basic concept of the supreme principle of morality. TC (or a principle quite like it) could be practical, absolutely necessary, and binding on all rational agents.

Moreover, several of the further criteria Kant develops for the supreme principle of morality do not seem to serve as a basis for rejecting TC. The valuational argument can be successful against a particular principle only if the principle's advocate must hold that the moral value of willing from duty to conform to it depends on the effects of doing so. But an advocate of TC need not hold this. Equally unpromising as a response to TC would be to insist that it is a material principle, and, therefore, it could not be the supreme principle of morality. For Kant has given us no good reason to think that TC is such a principle – that an agent has sufficient grounds to conform to it only if he expects doing so will enable him to realize some object he desires (and/or have a hedonic payoff). One might appeal to Kant's notion that the supreme principle of morality must generate a set of duties endorsed by ordinary moral reason, arguing that TC leaves some important ones out, for example, that to promote others' welfare. But this tactic would be ineffective since the list of duties in TC could simply be expanded. TC and principles like it seem to pose a particular challenge to the possibility of a successful derivation of the Categorical Imperative.

Kant does, however, have at his disposal grounds for rejecting TC. It belongs to Kant's basic concept of the supreme principle of morality that it serve as the justificatory basis for all duties (section 1.2). The principle must, therefore, exhibit the reason why we have a duty to do a certain thing, yet we do not have a duty to do something else. To some extent, TC accomplishes this task. To the question of why x (e.g., telling the truth) is a duty, a proponent of TC can respond: because x is among the prescriptions conjoined in the supreme principle of morality. Yet Kant would, I think, insist that this is not enough. How, he would ask, would the proponent answer the question of why this particular prescription is incorporated into the principle but

another one (e.g., worship Amon) is not? Of course, from Kant's perspective it wouldn't suffice for her to say that the list is a product of present social conditions – merely the reflection of the values of a particular place and time. To say this would be to offer an explanation but not a justification of particular duties contained in TC. Kant implies that the supreme principle of morality must provide a justificatory rationale for the duties derived from it. As we will see in the next chapter, Kant's own candidates do (though one might disagree with this rationale). TC, it appears, does not really provide a justificatory rationale for the duties derived from it. Therefore, TC cannot be the supreme principle of morality.

Of course, this response would not give an answer to someone who rejected the notion that the supreme principle of morality need provide a principled method (in Kant's sense) of enumerating duties. One brand of rational intuitionism, for example, might hold that we immediately grasp TC, incorporating just these duties, and that is all there is to it.

There is another reason Kant might give for rejecting TC as a viable candidate for the supreme principle. For Kant if one can rule out the possibility that a candidate for the supreme principle is knowable a priori, then this candidate is not viable. As we saw earlier (section 4.10), Kant holds that only if we can plausibly hold that a candidate is justifiable a priori could we have good reason to hold that it conforms to the basic concept of the supreme principle, according to which this principle must be absolutely necessary. Kant maintains that it is plausible to hold that the Categorical Imperative is justifiable a priori. In *Groundwork* III, he attempts to provide an a priori justification of the Formula of Universal Law (or something quite like it), appealing to the essential character of freedom, causality, rational willing, and so forth, rather than to the experiences of particular individuals or cultures. Kant might claim that it would be hopeless from the outset to attempt to provide an a priori justification of TC. How, he might ask, could one make a sincere attempt to justify TC as the supreme principle of morality without appealing to the notion that, in the past, human beings have found committing adultery to be wrong, honoring their parents to be required, and so forth?

It would, I think, be misguided to react to this question by making the following claim: "Kant's situation is no better, for he also relies on experience to justify the Categorical Imperative, since for him a condition of success for this principle's derivation is that the duties the principle generates cohere with ordinary moral reason." For this claim neglects the distinction between the derivation of the Categorical Imperative and its deduction. As I suggested earlier (section 4.10), that a successful derivation of the Categorical Imperative must be grounded in experience does not entail that a deduction of it must be as well.

Once again, though, an opponent might respond to Kant's question by asserting that we know TC through rational intuition. Our reason enables

us to recognize immediately that TC is the supreme principle of morality, she might say. To me this response seems implausible, but I do not think that Kant demonstrates it to be indefensible.

7.9 Further Nonconsequentialist Rivals

Chapter 2 focused on Henry Allison's claim that if we grant Kant the assumption that rational agents have transcendental freedom, then Kant can offer a successful derivation of the Formula of Universal Law. Allison reconstructs a derivation of this formula that in his view achieves its aim – that is, establishes that if there is a supreme principle of morality, then it is this formula. Key to Allison's reconstruction is the notion that only the Formula of Universal Law (or, presumably, equivalent principles) is capable of justifying the maxims of transcendentally free rational agents. I challenged this notion, arguing that Kant fails to eliminate the possibility that some other principle plays this role. In effect, I contended that Allison's reconstructed derivation does not eliminate certain rival candidates for the supreme principle of morality. One rival was the "bizarre principle" BP: "Act only on that maxim that you *cannot*, at the same time, will that it become a universal law"; the other rival was WU: "Act only on that maxim which, when generalized, could be a universal law." On the criterial reading I have advocated, does Kant have the resources to eliminate these candidates?

According to criterion viii, the supreme principle of morality must be such that a plausible set of duties (relative to ordinary rational moral cognition) can be derived from it. BP clearly fails to fulfill this criterion. According to it, an agent's acting on the following maxim would be morally *impermissible*: "From self-love, during my free time, I will exercise in order to stay in shape." According to BP, let us specify, willing the universalization of a maxim amounts to willing a world in which each agent adopts the maxim and if the circumstances described in the maxim arise, he or she acts on it. It *is* (rationally speaking) possible for an agent to act on this maxim and will its universalization. First, there is nothing incoherent in the agent's imagining the world of the universalized maxim, so it is not the case that it is irrational to will it on the grounds that it is irrational to will the impossible. Second, in willing that each agent adopt the maxim, and if he has any free time, acts on it, the agent would not be exhibiting the practical irrationality of undermining her own capacity to attain her end of staying in shape. All others' exercising during their free time to stay in shape would not preclude her from exercising in her free time and thereby staying in shape herself. BP entails not only that we must not act on this (apparently innocuous) exercising maxim, but that we are forbidden from acting on a maxim such as this: "From duty, unless I am incapacitated I will devote time and/or money to charity work in order to better the condition of fellow human beings." For

this maxim also fails the test implicit in BP. Clearly, BP does not generate a set of duties amenable to ordinary moral reason.

Although WU does not have quite the counterintuitive implications of BP, it also seems to fail to generate a set of duties acceptable to ordinary moral reason. WU commands that we act only on maxims that, when generalized, can be a universal law. Consider the maxim: "In order to promote my own happiness, I will never help a stranger or mere acquaintance in need." This maxim would be generalized, let us specify, if in order to promote his or her own happiness, each agent never helped a stranger or mere acquaintance in need. But it could be a universal law that this occur. There is nothing incoherent or self-contradictory in imagining it. Therefore, according to WU, it would be morally permissible to act on the maxim in question. Yet this result seems to clash with our ordinary moral consciousness, which embraces at least a minimal duty of beneficence.

Of course, we cannot take it for granted that the Categorical Imperative itself satisfies Kant's eighth criterion for the supreme principle of morality. As we will see in the next chapter, it is doubtful whether the Formula of Universal Law generates a set of duties acceptable to commonsense morality.

7.10 Summary

Kant has the materials at hand to argue (plausibly, in my view) that certain rivals to the Categorical Imperative are not viable candidates for the supreme principle of morality. Each of these rivals, he can show, fails to fulfill the criteria he has developed for the supreme principle. Nevertheless, the derivation remains incomplete. First, a full derivation would require Kant to eliminate all (possible) rivals to the Categorical Imperative. But as far as I can tell, Kant does not provide us with an effective method for insuring that we have *considered* all rivals. (As we have seen, the method Kant suggests, namely that of categorizing all possible rivals as material principles, is not promising.) Therefore, I do not see how even those very well disposed to Kant's arguments could claim that he had actually proved there to be no rival to the Categorical Imperative that could fulfill each of the criteria he develops for the supreme principle.

However, it would be no small achievement for Kant to show that, unlike the rivals we have discussed in this chapter, the Categorical Imperative (either in the Formula of Humanity or the Formula of Universal Law) remains as a viable candidate for the supreme principle of morality. To show this, Kant must demonstrate that his candidate could fulfill all of his criteria for this principle. His attempt to do this is the topic of Chapter 8.

8

Conclusion: Kant's Candidates for the Supreme Principle of Morality

8.1 Kant's Candidates and Criteria for the Supreme Principle of Morality

Kant's derivation as I have interpreted it is in the first instance a derivation of the Formula of Universal Law. Yet it is open to Kant to offer a derivation of the Formula of Humanity using the same basic steps. After all, the rivals to the latter formula (e.g., utilitarian principles) are also rivals to the former. If, based on an appeal to criteria he develops for the supreme principle, Kant succeeds in disqualifying the rivals we discussed in Chapter 7 to the Formula of Universal Law, then, in effect, he also succeeds in eliminating rivals to the Formula of Humanity.

Now an opponent might grant that Kant, through appeals to his criteria, eliminates many rivals to his candidates for the supreme principle of morality. But, the opponent might claim, this is a Pyrrhic victory; appeals to Kant's criteria would also dispose of Kant's own formulas. Does Kant have the resources to rebut this claim? Does each of his formulas remain a viable candidate for the supreme principle of morality? This is the question that this chapter addresses, although it does not attempt to answer it thoroughly.

At the outset, it is once again helpful to have in view the criteria Kant embraces for the supreme principle of morality. There are eight main ones; four Kant incorporates into his basic concept of the supreme principle, the other four he develops through analysis of ordinary moral thinking. According to Kant's basic concept, the supreme principle of morality would have to be (i) practical, (ii) absolutely necessary, (iii) binding on all rational agents, and (iv) the supreme norm for the moral evaluation of action. Moreover, this principle must be such that: (v) every case of willing to conform to it because the principle requires it has moral worth; (vi) the moral worth of willing to conform to the principle because the principle requires it stems from its motive, not from its effects; (vii) an agent's representing the principle as a

law – that is, a universally and unconditionally binding principle – provides him with sufficient incentive to conform to it; and, finally, (viii) a plausible set of duties (relative to ordinary rational moral cognition) can be derived from the principle.

The main issue before us is whether either the Formula of Universal Law or the Formula of Humanity remains as a *viable candidate* for a principle that fulfills the full set of criteria. A *derivation* of a principle does not aim to show that it actually fulfills the entire set of criteria. For showing this would require proving that the principle fulfills criteria ii and iii – that it is absolutely necessary and binding on all rational agents. It would involve giving a *deduction* of the principle, an endeavor that is not our concern here.

Here is how the chapter unfolds. In section 8.2, I argue that each formula remains a viable candidate for fulfilling criteria i–iii, and that, if we are willing to modify iv slightly, each one also remains a viable candidate for fulfilling it. Neither of the formulas fails as a candidate for the supreme principle of morality on the grounds that it could not satisfy Kant's basic concept of the supreme principle of morality (if we modify this concept a bit). The next section (8.3) attempts to show that criteria v–vii are also unproblematic. The bulk of the chapter concerns criterion viii. Is either the Formula of Universal Law or the Formula of Humanity such that, if it was actually binding, from it would stem duties acceptable to ordinary moral consciousness? Sections 8.4–6 focus mainly on the Formula of Universal Law, 8.7–9 on the Formula of Humanity. A lengthy book could easily be devoted to the question of whether, if valid, these formulas would generate duties that square with those we take ourselves to have. As students of Kant are well aware, each formula presents thorny difficulties of interpretation. So I am not able here to answer this question thoroughly. I argue, however, that we have good reason to doubt whether the Formula of Universal Law fulfills criterion viii. Therefore, we have good reason to doubt whether this formula remains as a viable candidate for the supreme principle of morality. The Formula of Humanity, I suggest, seems more promising regarding criterion viii, although it leaves us with some troubling concerns.

The two formulas, claims Kant, are representations of "the very same law" (GMS 436). If, as it seems reasonable to assume, this claim implies that the two would give rise to the same moral requirements, then this chapter offers some evidence that it is incorrect. Unfortunately for its defenders, the Formula of Universal Law does not seem to forbid acts of violence committed for revenge, whereas the Formula of Humanity does. It is open to a champion of the Formula of Humanity to take the Formula of Universal Law as a rival and to try to eliminate it as a candidate for the supreme principle of morality on the grounds that it would clearly fail to yield duties acceptable to ordinary moral thinking.

8.2 Two Formulas and the Basic Concept of the Supreme Principle of Morality

Neither the Formula of Universal Law nor the Formula of Humanity should be eliminated as a candidate for the supreme principle of morality on the basis of a discernible inability to fulfill criteria i–iii. We could act on account of each one – each is practical – though, as we have noted regarding the Formula of Universal Law, it is harder than Kant acknowledges to determine which actions the principles require. Each formula could also be absolutely necessary, that is, binding on all the agents within its scope, regardless of the agents' particular inclinations. Moreover, the scope of each could extend to all rational agents. Despite its use of the term "humanity," the Formula of Humanity is not limited in scope to human beings. For, as we have noted, "humanity" there refers to rational nature, that is, the capacity for rational choice, a capacity inherent in all rational agents.

There is, however, a difficulty that arises in connection with criterion ii. For the sake of simplicity, in explaining it I focus on us, human agents, bracketing other rational agents, and I use the generic term "Categorical Imperative" to refer to both the Formula of Universal Law and the Formula of Humanity, since no differences between the two formulas will come into play. To say that the supreme principle of morality is absolutely necessary is to say that without possible exception we ought to conform to it. And it indeed does seem that the Categorical Imperative could be such that without possible exception we ought to conform to it. The difficulty is not with the possibility of the Categorical Imperative's fulfilling ii, but with something that would result if it did fulfill ii.

The difficulty arises against the background of Chapter 6. There I argued that Kant needs to acknowledge that even with the best intentions and effort an agent might not only fail to apply the Categorical Imperative correctly but might even embrace a rival as the supreme principle of morality. Take someone who has done the latter, Stram the utilitarian from section 6.6. For the Categorical Imperative to be a viable candidate for the supreme principle of morality, it must be at least possible that Stram ought always to abide by it, even though he often fails to do so. Yet, and here the difficulty emerges, it seems that Kant must deny this. Kant embraces as an axiom that ought implies can. According to him, I believe, this axiom entails that if an agent is obligated to conform to a principle, then he must have an incentive to conform to it. If an agent does not have an incentive to conform to a principle, then she will not be able to do so, since for Kant all action requires an incentive (Rel 35, English ed. 30).

However, it appears that at some points Stram might not have an incentive to conform to the Categorical Imperative. He has done his best to determine what his duty is, yet has adopted a principle that in particular cases in his life clashes with the Categorical Imperative. In accordance with a Kantian

theory of agency, let us suppose that Stram always has an incentive to do what *he takes* to be morally required. (Of course, his having an incentive to do something does not, on this Kantian theory, entail that he will do it. For no incentive can determine an agent's will unless he has incorporated it into his maxim. And instead of the moral incentive, he might choose to incorporate into his maxim some inclination.) Even if we suppose that Stram always has this incentive, he would, it seems, sometimes fail to have an incentive to do what the Categorical Imperative requires, since doing what this imperative requires would sometimes amount to doing just the opposite of what he thinks he is morally obligated to do. For example, he takes himself to be required to lie in certain circumstances, but the Categorical Imperative entails that he has a duty not to lie in these circumstances. If we want to maintain that Stram nevertheless ought to (has a duty to) abide by the Categorical Imperative, then we must deny Kant's notion that ought implies can. In light of Chapter 6, it seems that for Kant maintaining that the Categorical Imperative is absolutely necessary would require him to abandon a notion he holds near and dear. What might make matters seem even worse is that, on my reading, Kant appeals to this very notion in his defense of criterion vii (see section 5.7).

In response, I do not see any great harm in Kant's abandoning the notion that if an agent has a moral duty to do something, he must have an incentive to do it. First, it does not strike me as implausible to maintain the following. A morally reflective agent who, since he did not believe it to be his duty to do something, did not have an incentive to do it nevertheless morally ought to have done it. (It might, however, be implausible to blame the person for failing to do what he ought to have done.) Second, though it is true that one argument for criterion vii appeals to the "ought implies can" notion in question, Kant has another argument at his disposal that does not (see section 5.7).[1] Third, Kant's moving away from the notion in question would actually be a far less radical departure for him than it might seem. In Chapter 6, we came across the following passage:

> [W]hile I can indeed be mistaken at times in my objective judgment as to whether something is a duty or not, I cannot be mistaken in my subjective judgment as to whether I have submitted it to my practical reason (here in its role as judge) for such a judgment. . . . [I]f someone is aware that he has acted in accordance with his conscience, then as far as guilt or innocence is concerned nothing more can be required of him. It is incumbent on him only to enlighten his *understanding* in the matter of what is or is not duty. (MS 401)

Here Kant seems to acknowledge that without being led astray by her inclinations, an agent can make an error in determining whether she has a duty to do something. But in order for the notion of her making such a mistake to make sense, there must be a correct answer to the question of what her duties are. It must be possible that though an agent does not believe

she has a duty to do something, she actually does. Let us suppose that the following is such a case. An agent is convinced that being truthful to the police and telling them the whereabouts of an innocent person they intend to jail is morally forbidden, when it is actually morally required. Kant must admit that the agent might not have an incentive to abide by what is morally required, that is, to abide by her duty. After all, why should she have one? She thinks that in this case being truthful to the police is morally forbidden. If Kant here invoked the notion that an agent does not have a duty to do something unless she has an incentive to do it, he would have to conclude that, actually, the agent does not have a duty to be truthful to the police. But this would contradict the assumption with which we began, namely that, as a matter of fact, she does have such a duty. In short, it would be difficult for Kant to cling to the notion that if an agent has a duty to do something, he must have an incentive to do it, all the while acknowledging, as he seems to in the *Metaphysics of Morals*, that an agent can be mistaken about what her duties really are. It appears that in making this acknowledgment, Kant himself is, at least implicitly, moving away from the view that ought implies can (interpreted in the particular way in question).

At any rate, returning to the essential point at hand, the Formula of Universal Law and the Formula of Humanity remain as viable candidates for principles that fulfill criteria i–iii. However, the last criterion in Kant's basic concept of the supreme principle of morality, iv, poses a problem. For a principle to be in Kant's sense the supreme norm for the moral evaluation of action, every action's moral permissibility, moral requiredness, and moral worth must be defined in terms of it. It is possible to define the moral permissibility or moral requiredness of any action ultimately in terms of the Formula of Humanity or the Formula of Universal Law. To focus on the former, if in performing an action an agent treats humanity in herself and others as an end, the action is morally permissible. If in refraining from performing an available action the agent would not be treating humanity in herself and others as an end, then the action is morally required. There seem to be no actions the moral permissibility or requiredness of which could not be "covered" by either one of these principles, although it might not be a simple matter to determine how the action is covered – that is, whether it is permissible or required.

But what about moral worth? Granted, there is nothing within either principle itself that would preclude our defining moral worth with reference to it. We might, for example, hold that a necessary condition for an action's having moral worth is that it not conflict with what the Formula of Universal Law requires. In effect, this seems to be Kant's own position. If the argument of Chapter 6 has been successful, however, we can see that this is a problematic view for Kant to hold. He needs to acknowledge that some actions not in accordance with this formula, namely those done from duty, have moral worth. In my view, neither the Formula of Universal Law nor the

Formula of Humanity, nor, for that matter any other principle, is suited to be the (sole) principle in reference to which all morally valuable action is defined. In acting from duty (and thereby fulfilling the four Kantian conditions on such action specified in section 6.9), agents can be acting on various different principles, ones that clash with Kant's as well as with one another. Nevertheless, all of these actions have moral worth, or so Kant should grant. If this is correct, then Kant must either conclude that neither of his own candidates for the supreme principle of morality satisfies his basic concept of this principle, or alter his basic concept. The latter course clearly seems preferable. From now on, we will understand Kant to hold that the supreme principle of morality must be the supreme norm for the evaluation of the moral permissibility and requiredness of an action, but not of its moral value. Kant's Formula of Universal Law and Formula of Humanity remain viable candidates for principles that satisfy criterion iv understood in this way.

Of course, we are in no position to contend that Kant's formulas actually do meet his basic concept of the supreme principle of morality. For we have not shown (nor will we show) that either one is binding on all rational agents. However, we can see that Kant's formulas remain as viable candidates for principles that realize this basic concept (if we modify the concept slightly).

8.3 Two Formulas and Further Criteria

Do Kant's formulas also stand as ones that we can reasonably maintain might fulfill the other four criteria he develops?

According to criteria v and vi, the supreme principle of morality must be such that each case of willing from duty to conform to it has moral worth – worth that does not stem from the willing's results. The Formula of Universal Law stands as a viable candidate for fulfilling these criteria, since its defender can coherently claim that each case of willing from duty to conform to it has moral worth, regardless of its results. For the defender of this formula obviously need not identify the good with anything external to willing, such as the general happiness, and thus need not hold the value of willing to depend on anything external, such as its effects on the general happiness. In Chapter 6 I defended the claim that *all* acting from duty has moral worth – worth that does not stem from the willing's effects. So in my view, Kant's Formula of Universal Law actually does fulfill criteria v and vi.

Whether the Formula of Humanity stands as a viable candidate for fulfilling (let alone fulfills) v and vi might seem to be more questionable. Recall that this formula reads: "So act that you treat humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" (GMS 429, emphasis omitted). As we will discuss, an advocate of this principle holds rational nature to be unconditionally and incomparably valuable. He judges the moral permissibility of an agent's

action in terms of whether she treats rational nature as such. One might think that he is thereby committed to denying v and vi. After all, would he not be required to acknowledge that some actions from duty would fail to have moral worth, specifically those that had the effect of harming rational nature?

I do not see why the advocate of the Formula of Humanity would be required to acknowledge this. I have defended the view that every case of acting from duty has moral worth. If an agent's (in itself sufficient) incentive for acting stems from the notion that the Formula of Humanity requires the action (and she meets the other requirements for acting from duty discussed in section 6.9), her action has moral worth. That worth is not at all a function of her action's results or even of its actually conforming to what the Formula of Humanity requires. There is nothing incoherent in holding this and, at the same time, holding humanity to be unconditionally and incomparably valuable. That one takes humanity to be unconditionally and incomparably good does not rationally compel him to take it to be the only thing that is unconditionally good. An advocate of the Formula of Humanity can consistently maintain, in accordance with criteria v and vi, that every case of willing, from duty, to conform to this principle has moral worth, regardless of what results from it.

(In fairness to [potential] opponents of Kant, I should remark that a parallel point could be made with regard to maintaining everyone's happiness to be unconditionally valuable. In Chapter 7 we discussed the principle U': Always perform a right action, one that yields just as great a sum total of well-being as would any alternative action available to you. Now suppose that an advocate of U' holds everyone's happiness to be unconditionally valuable. That he holds this does not itself entail that he must hold it to be the only thing that is unconditionally valuable. Without contradiction, he can also claim that acting from duty is unconditionally valuable. It is possible that an advocate of U' could coherently maintain, in accordance with v and vi, that every case of willing, from duty, to conform to U' has moral worth, regardless of what results from it. What I tried to show in section 7.4 is that a typical advocate of U' cannot coherently maintain this; for a typical advocate embraces Outcome Utilitarianism, the view [roughly] that goodness is solely a function of well being.)

According to criterion vii, the supreme principle of morality must be such that an agent's representing it as a law provides him with sufficient incentive to conform to it. Should the Formula of Universal Law and the Formula of Humanity be disqualified based on an inability to fulfill this criterion? Opponents of Kant (e.g., Humeans) would be quick to suggest that no principle could fulfill it, since the criterion itself rests on a mistaken theory of agency. Sensuously based desire is a necessary ingredient in any incentive for action, the opponents might say. And your representing a principle as a law is not itself going to generate any desire to conform to the principle.

In this book, however, we have in effect assumed that the Kantian view of agency is the correct one. This assumption would be question-begging if it had been employed to eliminate non-Kantian candidates for the supreme principle of morality. Yet it has not been used in this way. Not one rival candidate has been dismissed on the basis of its failure to meet this criterion. Actually, I have argued against Kant's claim that all rival principles must be understood to be material, and thus unable to fulfill (vii) (see section 7.2). Nevertheless, the question remains as to whether Kant's candidates could fulfill this criterion. Assuming that sensuously based desire is not a necessary ingredient in all incentives for action, I find no good reason to suppose that Kant's principles could not do so.[2]

The final criterion for the supreme principle of morality might pose the greatest difficulty for Kant's candidates. According to criterion viii, the supreme principle must be such that, if it were binding on us, a plausible set of duties would stem from it, where "plausible" means in accord with reflective moral common sense. Unless Kant's formulas meet this criterion, we must eliminate them as candidates for the supreme principle of morality. Much of the rest of the chapter is devoted to the question of whether they do.

8.4 Two Formulas and Ordinary Moral Consciousness

A couple of observations will be helpful before we examine whether the Formula of Universal Law and the Formula of Humanity would generate prescriptions in accord with reflective moral common sense.

First, the project of examining whether these formulas fulfill criterion viii faces a difficulty from the start. Contrary to what Kant implies, ordinary moral thinking is not of a piece. Sincere, reflective people disagree about what a person morally ought to do – for example, when a gravely ill person's committing suicide would end her suffering and diminish that of her family. My criticism of Kant's notion that all actions from duty conform to it turns on there being such disagreement. However, there seems to be widespread agreement on some issues. The commonsense view, for example, seems to be that making a false promise from the motive of financial gain is morally wrong – we have a duty to refrain from doing so. It is when a principle would, if valid, fail to generate duties of this sort, ones that ordinary rational moral cognition seems clearly to endorse, that we should reject the principle, or at least that is how I interpret criterion viii.

Second, thorough assessment of whether Kant's formulas would generate duties that accord with reflective moral common sense would require thorough examination of precisely how best to interpret the formulas. The latter task alone might call for a book-length treatment. For as anyone who has taught the *Groundwork* is all too aware, Kant himself suggests various different readings of these formulas, especially of the Formula of Universal

Law. In offering a brief assessment of the formulas, which does not aim to be definitive, I often rely on the interpretive work of others.

8.5 Formula of Universal Law: Practical Contradiction Interpretation

Does the Formula of Universal Law generate prescriptions acceptable to common sense? To begin, let us suppose, as Kant quite reasonably does, that according to common sense an agent ought not to make false promises for his own financial gain; it is morally impermissible to do so. Since Kant holds that all acting is acting on a maxim, if the Formula of Universal Law is to yield results consistent with common sense, a maxim of false promising for one's own financial gain must fail the test contained in this formula. Kant, of course, holds that such a maxim does fail, and thus that we have a duty not to act on it (GMS 422).[3] Philosophers have offered various accounts of precisely how the maxim fails the test, but I explore only two of them here. (As I indicated earlier, I simply assume that each of these two accounts is permitted by Kant's texts.)

According to Korsgaard, on the most philosophically plausible reading, an agent cannot act on the sort of false promising maxim in question and at the same time will that it become a universal law because doing so would generate a "practical contradiction."[4] To see how it would, we need first to note that, as Kant indicates (GMS 422), the maxim of the action would be something like FPM, "From self-love, when I believe myself to be in need of money I shall borrow money on a promise to repay it, even though I know that this will never happen." How would we describe a world in which this maxim would be a universal law, that is, the maxim's "universalization"? On the interpretation Korsgaard advocates, the Practical Contradiction Interpretation, we would say that in this world the following obtains: from self love, when anyone believes himself to be in need of money, he tries to borrow money on a promise to repay it, even though he knows that this will never happen.[5] What would be contradictory in the agent's acting on FPM and, at the same time, willing the world in question? Imagine that she is doing this. First, since she is acting on FPM, the agent is trying, through the means of making a false promise, to attain her end of getting money. Second, in willing the world of FPM's universalization, she is willing a world in which taking these means will not enable her to attain her end. For if each person in financial need tries to get money on a promise of repayment (and, if she succeeds, does not in fact repay), then potential lenders will not lend money simply on a promise to repay. It will not be possible using a promise alone – in contrast, for example, to some kind of written contract – for a person in financial need to get money.[6] So the agent is trying through a particular means to attain an end and at the same time willing a situation in which it is impossible through this means to attain the end. In effect, the agent is willing that he be thwarted in attaining the end he is pursuing. Therein

lies a practical contradiction.[7] Therefore, insofar as he is rational, an agent cannot act on FPM and at the same time will that it become a universal law.

Kant holds that in acting on FPM an agent would be violating the Formula of Universal Law and that, therefore, we have a duty not to act on this maxim. The Practical Contradiction Interpretation offers a way of understanding how, precisely, acting on FPM would violate this formula. In short, the maxim would fail because the agent's attaining the end it specifies (getting money) through the means it specifies (making a false promise) depends on most agents' not taking this means to the end.[8] The maxim's effectiveness would be a function of its being exceptional. The Practical Contradiction Interpretation allows us to see that as far as a maxim such as FPM is concerned, the Formula of Universal Law generates results that cohere with ordinary moral thinking.

This interpretation, however, also generates results that clash with ordinary moral thinking. Following Barbara Herman, suppose that an agent acts on the maxim: "From self-love, I will shop in this year's after-Christmas discounts for next year's Christmas presents in order to save money."[9] If everyone acted on this maxim (i.e., if it were universalized), then after-Christmas discounts would disappear – they would be too damaging to pre-Christmas income. In willing the world of his universalized maxim, the agent would be willing a world in which it would not be possible to save money by means of shopping in this year's after-Christmas discounts for next year's Christmas presents. It would be irrational for the agent to will this world and at the same time act on his maxim of saving money through taking this very means. For in willing this world, he would be willing to be thwarted in his pursuit of his end. If the maxim of false promising generates a practical contradiction, then so does this maxim of economical shopping. The effectiveness of acting on either of them is a function of its being exceptional that people do so. Although common sense would condemn as morally impermissible acting on the maxim of false promising, it would not condemn as such acting on the maxim of economical shopping. There just does not seem to be anything morally wrong with taking advantage of after-Christmas sales in a way that is effective only against the background of others not trying to take advantage of them in this way. If the Formula of Universal Law says otherwise, then so much the worse for it – in particular for its prospects of fulfilling criterion viii for the supreme principle of morality.

The discussion of maxims in section 1.3 laid the groundwork for a possible reply to this objection. On the view I adopted, *the* maxim of an agent's action is the most general rule of the proper form on which he acts. If this view is correct, it seems unlikely that the agent's rule of economical shopping really counts as his maxim; for this rule seems too specific. Is it not likely that the rule is ancillary to (i.e., serves as a means of executing) a maxim such as "From self-love, I will shop at sales in order to save money"? If so, then (arguably) no practical contradiction is generated by the agent's acting on

his maxim and at the same time willing that it become a universal law. In willing the universalization of his maxim, the agent would not (arguably) be preventing himself from attaining his end through the means specified in his maxim.

Is this response effective? In acting on the very specific rule of shopping at this year's after-Christmas sales for next year's gifts, the agent might not actually be implementing a more general rule, which would count as her maxim. The very specific rule might just be her maxim.[10] It would be strange if it were since it is hard to see how someone would think to adopt the rule if not in the context of carrying out some general policy of trying to save money in her shopping. But the strange is far from the impossible. If this very specific rule were her maxim, it would, I think, be contrary to ordinary moral consciousness to claim it to be morally impermissible for the agent to act on it. Yet that is what a defender of the Formula of Universal Law would have to do, at least on the interpretation of it that we have been employing.

Moreover, not all rules that, contrary to ordinary moral reason, fail the Formula of Universal Law test on this interpretation are so specific that we would question them as examples of maxims. Suppose that Jack, the son of dock workers, acts on the following rule: "In order to earn a comfortable living, I will become a professor, rather than do physical labor." (For Jack making a comfortable living amounts to making enough to have his own house, car, computer, and so forth.) There would be nothing odd if, in Jack's case, this rule were not ancillary to a more general one. Let us, then, assume that the rule is his maxim.[11] On the Practical Contradiction Interpretation of the Formula of Universal Law, this maxim turns out to be morally impermissible. In willing a world in which everyone acted on his maxim, Jack would be willing the ineffectiveness of the means he takes (becoming a professor) to his end (earning a comfortable living). For in this world the institutional framework for salaried professors would, in all likelihood, not be in place. Universities do not function without support from people who earn a living through physical labor.[12] Some maxims – for example, Jack's as well as the false-promising maxim (FPM) – specify a means that is effective for attaining their end only in a context in which it is exceptional for agents to take this means to the end. These maxims take advantage of predictable regularities in agents' behavior. On the Practical Contradiction Interpretation, the Formula of Universal Law condemns as morally impermissible all acting on such maxims. But this condemnation clashes with commonsense morality, according to which acting on some of these maxims (e.g., Jack's) is not contrary to duty.

Of course, the Practical Contradiction Interpretation is not the only reading of how Kant envisages (or might envisage) that a maxim of false promising would fail the Formula of Universal Law test. Perhaps there is an

alternative reading according to which FPM would fail, but other maxims, ones such as Jack's which we take to be morally permissible, would pass.

8.6 Formula of Universal Law: Universal Availability Interpretation

Thomas Pogge has presented an alternative that might seem to secure these results.[13] The main feature that distinguishes Pogge's reading from the one we have already considered is its account of how a maxim is to be universalized. On the Practical Contradiction Interpretation, imagining a maxim to be a universal law amounts to imagining a world just like ours except that everyone has actually adopted the maxim (and acts on it when occasions arise). On Pogge's reading, imagining a maxim to be a universal law amounts to imagining a world just like ours except that everyone believes himself to be permitted (i.e., "morally" free) to adopt the maxim, and those who are inclined to adopt it do so (and act on it when occasions arise).[14] In light of this difference between the Practical Contradiction Interpretation and Pogge's reading, I call the latter the Universal Availability Interpretation. According to the Universal Availability Interpretation of the Formula of Universal Law, an agent is to ask herself whether she can act on a maxim and at the same time will (in short) that everyone hold the maxim to be available, in the sense of morally acceptable.

The Universal Availability Interpretation saves the Formula of Universal Law from yielding the result so unwelcome to common sense that it is morally impermissible to act on the maxim of earning a comfortable living by becoming a professor or that of economizing by shopping at this year's sales for next year's gifts. For purposes of illustration, I will just consider once again the former maxim, held by Jack: "In order to earn a comfortable living, I will become a professor, rather than do physical labor." Jack could act on this maxim and at the same time will that everyone feel (morally) free to act on it. In willing a world in which everyone did feel this way, Jack would not be rendering ineffective the means specified in his maxim for attaining his end of earning a comfortable living. If the moral availability of this maxim resulted in a mass rush to graduate school of those aiming at a comfortable living, then he would be thwarting this means. But surely such a rush would not occur. For it is not any moral qualms about Jack's maxim that stand in the way of masses of people adopting it but rather things like inclinations to take different means – for example, ones that require less time in the library or laboratory – to the end of earning a comfortable living. The Universal Availability Interpretation has the advantage over the Practical Contradiction Interpretation of allowing the Formula of Universal Law to grant the moral permissibility of some maxims that, though they depend for their effectiveness on being exceptional, are not condemned by ordinary moral reason.

Perhaps, however, this advantage comes at too high a price. For it is questionable whether, on the Universal Availability Interpretation, false-promising maxims turn out to be morally impermissible.

Pogge suggests that his interpretation does generate the desired results regarding such maxims. The example of a false-promising maxim Pogge considers is a bit more general than the one we have thus far discussed; it is "When in need, I will make deceitful promises so as to alleviate my difficulties."[15] On his reading, in the world of the universalized maxim everyone would feel (morally) free, when in need, to make deceitful promises so as to alleviate his difficulties.[16] According to Pogge, the false-promising maxim would be "pointless" in this world; for acting on it would not alleviate one's difficulties.[17] That it would be pointless, he continues, "leads to the rejection of that maxim, because . . . its universal availability would block the agent's attainment of the material end of his conduct under the maxim. And with the objective out of reach, the agent *cannot* will the maxim: If it cannot satisfy his interest in its material end, the agent loses his only possible (heteronomous) motive for adopting it."[18] If his acting on a maxim is to pass the Formula of Universal Law test, an agent must (rationally speaking) be able to act on it in the world in which the maxim has been universalized, suggests Pogge. But in the world of the universalized false-promising maxim, the agent could not act on his maxim. For, as the agent would realize, acting on it would do nothing to enable him to secure his end of getting out of difficulties. Therefore, the agent (insofar as he was rational) would find himself with no motive to adopt his maxim. In this sense, he could not act on it. So the false-promising maxim turns out to be morally impermissible.

On the Universal Availability Interpretation, the maxim's turning out this way depends on its being the case that in a world where everyone felt (morally) free to act on the maxim, it would be "pointless" for a particular individual to act on it. But is this really the case? According to Pogge, "people in need would (be known to) have no reason not to make deceitful promises" and "potential promisees would (be known to) have good reason to reject promises made by persons in need."[19] I think the first point is questionable. Granted, in the world in which everyone feels morally free to act on the maxim of false promising, people in need would (be known to) have no *moral* reason not to make deceitful promises. However, this does not entail that they would (be known to) have no reason at all not to make such promises. For there are *prudential* reasons not to make deceitful promises, even when one is in difficulties. For example, an agent might judge that the sanctions he would incur if it were to become known that he made a deceitful promise would be worse than his present difficulties and that the chances of its becoming known are great enough to render it not worth the risk for him to make the deceitful promise. The penalties in question might range from prison time if, for example, the deceitful promise was that his home remedy would cure cancer, to an inability without collateral

ever to again obtain money from his family if, for example, the deceitful promise was that he would pay back a loan from his uncle. (The notion that, if found out, the agent would incur these penalties is compatible with no one's holding it to be morally wrong to make deceitful promises in order to alleviate one's difficulties. Members of the agent's family, for example, might refuse to lend him any more money [in the absence of collateral] not at all on the basis that his behavior was morally bankrupt but simply because they do not want to lose any more money.) Contrary to Pogge's first point, in the world of the universalized maxim, people in need would sometimes have reason to refrain from making deceitful promises; it would be prudential rather than moral reason.

If the first point is questionable, then so is the second. In the world of the universalized maxim, would potential promisees always (or even the great majority of the time) have good reason to reject promises made by persons in need? Let's say that in the imagined world someone asks you to loan him money on the basis of a promise that he will pay it back. You would have good reason to reject his proposal if you (reasonably) believed that, in his view, his breaking the promise would not result in any significant penalty for him. (You might reasonably believe this if he is a stranger who probably does not think he will ever see you again). But you would have good reason to accept it if you (reasonably) thought that in his view his breaking the promise would hurt him a great deal. (You might reasonably believe this if he were a young business associate who depended on you for his climb up the corporate latter.) In the imagined world, it is not clear that in acting on the maxim of deceitful promising, an agent would be employing an obviously ineffective means to an end. Whether he would be depends (among other things) on others' *perceptions* of his prudential reasons for keeping his promise. In short, it seems that sometimes an agent acting on the deceitful-promising maxim could attain his end of alleviating his difficulties in a world in which everyone felt morally free to act on this very maxim. So it is questionable whether on the Universal Availability Interpretation the false-promising maxim actually turns out to be morally impermissible.

There is another problem with the Universal Availability Interpretation. Ordinary moral reason would, I venture, condemn as contrary to duty acting on the maxim "If anyone commits adultery with my spouse, I will kill the person in order to get revenge." On the reading in question, however, the Formula of Universal Law would not. An agent (insofar as she was rational) could act on this maxim in a world in which everyone felt morally permitted to do so as well. In this imagined world, perhaps those who did commit adultery would take greater precautions than they do now to avoid contact with the betrayed husband or wife. But that would not, for example, preclude an agent in the imagined world from attaining her goal of getting revenge through killing the woman who seduced her spouse; it would just make the killing more difficult.[20] Perhaps it is partly because (on his interpretation)

the Formula of Universal Law licenses such maxims that Pogge does not believe that on its own it constitutes a viable candidate for the supreme principle of morality.[21]

In sum, on neither the Universal Availability Interpretation nor the Practical Contradiction Interpretation would the Formula of Universal Law, if it were binding on us, generate duties that cohere with the dictates of ordinary moral thinking. On the former interpretation, it would turn out that, contrary to ordinary conviction, we have no duty to refrain from acting on (certain) maxims of false promising and violence. On the latter, it would turn out that, contrary to ordinary conviction, we have a duty to refrain from acting on (certain) maxims of taking advantage of predictable regularities in others' behavior, maxims such as that of earning a comfortable living by becoming a professor rather than by doing physical labor. At least on two readings, the Formula of Universal Law does not fulfill criterion viii for the supreme principle of morality.

It would, of course, be unwarranted to take this to show that the Formula of Universal Law fails to fulfill criterion viii. Our discussion has not been thorough enough to establish this conclusion. However, I do think that it helps to confirm a suspicion expressed recently by several Kantians that despite some ingenious efforts, no one has been able to make this formula work.[22] Perhaps someone will, but as Herman says, "past experience suggests a permanent fix-it situation: the correction of one difficulty or apparent oversight creates space for new problems to emerge."[23]

8.7 Fundamentals of the Formula of Humanity

The prospects for the Formula of Universal Law's generating a set of duties acceptable to ordinary moral reason do not appear to be good. Are the prospects for the Formula of Humanity any better? Kant himself seems to favor the Formula of Humanity as a basis on which to derive duties. For in the *Metaphysics of Morals*, Kant relies (at least implicitly) on this formula to derive the vast majority of the ethical duties he sets out.[24] But given Kant's suggestion that the two formulas are equivalent (GMS 436), perhaps he favors the Formula of Humanity simply because in his view it is less cumbersome to work with than the other formula. At any rate, I do not offer anything approaching an exhaustive treatment of the issue of whether the Formula of Humanity would generate a plausible set of duties relative to ordinary moral thinking. However, I do hope to say enough to suggest that, although the Formula of Humanity holds significant promise, defenders of it must confront some troubling issues.

Before we can discuss the question of which duties would stem from this formula (if it was valid), we need to understand the terms it employs. Unfortunately, like the Formula of Universal Law, it is not easy to interpret. The Formula of Humanity commands: "So act that you treat humanity, whether

in your own person or in the person of any other, always at the same time as an end, never merely as a means." An agent's acting so that she treats humanity (in herself or any other) as an end is a necessary and sufficient condition for her conforming to the formula. It is a necessary condition since the formula commands that an agent so act that she *always* treat humanity as an end. It is a sufficient condition since even if the agent acts so that she treats humanity in herself or another as a means, as long as she at the same time acts so that she treats it as an end, she has conformed to the formula.[25] So, at bottom, the Formula of Humanity amounts to a command so to act that we always treat humanity as an end.[26] "Humanity," let us recall, does not refer to the class of human beings but rather to a set of capacities: the capacities to set oneself ends and to adopt and act on rules, including rules of prudence (hypothetical imperatives) and rules of morality (categorical imperatives), often in pursuit of these ends. Would duties acceptable to ordinary moral reason stem from the command always to treat humanity so understood as an end?

An initial step toward answering this question is to examine the sense of "end" or, equivalently, "end in itself" at work in the Formula of Humanity.[27] Kant holds that humanity exists as an end in itself. But what does it mean for humanity to exist in this way? First, as we know from our discussion of Kant's derivation of this formula, an end in itself is something that has absolute or unconditional worth (GMS 428). It would be judged by an impartial rational spectator to be good in every possible context, even in ones in which it brought about undesired results. For Kant that an end in itself has absolute worth implies that all rational agents must (are rationally compelled to) value it and to act in ways that express their valuing it, regardless of whether they are inclined to do so (section 3.2).

Second, to say that humanity exists as an end in itself is to say that it has dignity (GMS 435; MS 434–435, 462). To have dignity, Kant suggests, is to have "unconditional and incomparable worth" (GMS 436). We have just noted what it means to have the first aspect of dignity, namely unconditional worth. Kant explains the second aspect of dignity, namely incomparable worth, by contrasting it with price: "What has a price can be replaced by something else as its *equivalent*; what on the other hand is raised above all price and therefore admits of no equivalent has a dignity" (GMS 434; see also MS 462). The value of something with dignity, then, is incomparable in the sense that it has no equivalent for which it can be exchanged. As Thomas Hill has argued, that it is seems to have two implications.[28] First (and quite clearly), something with dignity can never be legitimately sacrificed for or replaced by something with price. Not even all the gold in Fort Knox would compensate for the killing of one rational agent. Second (and not quite so clearly), something with dignity cannot even be legitimately sacrificed for or replaced by something else with dignity. Beings with dignity, says Kant, admit of "no equivalent." If, therefore, it is ever legitimate to kill one being

with dignity, thereby saving several other such beings, it will not, it seems, be because it is legitimate to make an exchange of the (lesser) value inherent in the former with the (greater) value inherent in the latter. An end in itself (and thus humanity) has dignity in that it has an unconditional value that admits of no equivalent, not in terms of price, nor, it appears, even in terms of other beings with dignity.

We have found that to say that humanity is an end in itself is to imply that it is something that has an absolute and incomparable worth. But presumably if Kant calls humanity an end in itself, then in some sense he thinks of it as an end. In what sense, precisely? This question is puzzling if one takes as a point of departure Kant's definition of an end in the *Metaphysics of Morals*. "An *end*," he says, "is an object of the will [*Willkür*] (of a rational being), through the representation of which the will is determined to an action to bring this object about" (MS 381; see also MS 384–385). In other words, an end is a state of affairs or event such that an agent, through her idea of it, is determined to will to realize it. An agent might, for example, have as an end to maintain his weight under two hundred pounds for the next six months or to win a tennis tournament. An end on this account is a goal, aim, or target – an object to be produced.[29] Yet Kant suggests that humanity is not an "end to be *effected*," but rather an "*independently existing* [*selbstständiger*] end" (GMS 437; see also MS 442). So in calling humanity an end in itself, he must have a broader notion of an end in view. Indeed, in the *Groundwork*, immediately before his derivation of the Formula of Humanity, Kant says that "an end is what serves the will as the objective ground of its self-determination" (GMS 427). An end is an objective ground of an agent's determining his will to an action. An end is a ground in that it is a reason that an agent has (or at least ought to have) for acting; an end is an *objective* ground in that it is an *object* such that, through representing it to himself, the agent gives himself (or at least ought to give himself) a reason for acting.[30] On this broader conception, ends are not limited to objects to be produced. They include any object the idea of which does (or ought to) give an agent a reason to act in a certain way. Someone's aim or goal of winning a tennis tournament might count as such an object, but so might an existent object such as her humanity. And Kant, of course, thinks that humanity, wherever and whenever it manifests itself, counts as an end in this broad sense. It does so by virtue of its being absolutely and incomparably valuable.

For Kant humanity exists as an end in itself, an unconditionally and incomparably valuable object the idea of which gives (or at least ought to give) all rational agents a reason for acting. What does it mean to act so that one always treats humanity as an end in itself, as the Formula of Humanity commands? Presumably one acts so that one treats humanity as an end in itself just in case what one wills to do is consistent with holding humanity to be of absolute and incomparable value. In the *Groundwork*, Kant calls rational nature (i.e., humanity) an "object of respect." In the *Metaphysics of Morals*,

he suggests that any being with humanity must not only respect himself but "exacts *respect* for himself from all other rational beings in the world" (MS 435; see also MS 462). These comments suggest that an agent's action is consistent with his holding humanity to be of absolute and incomparable value only if it manifests respect for humanity. Humanity is not an end to be effected (produced), but it is an end to be respected. However, the question remains: which actions are consistent with an agent's holding humanity to be an end in itself? Might the Formula of Humanity be more effective than the Formula of Universal Law at generating duties acceptable to ordinary moral reason?

8.8 Deriving Duties from the Formula of Humanity

The Formula of Humanity would (if it was binding on us) seem to be effective at engendering certain duties we take ourselves to have (e.g., a duty not to kill for revenge). From the Formula of Humanity (unlike from the Formula of Universal Law) it clearly follows that one must not act on a maxim of killing an adulterer to get revenge; for murdering an adulterer, and thus destroying his humanity, is obviously not consistent with respecting it as something absolutely and incomparably valuable. In general, destroying humanity would rarely, if ever, seem to express respect for it. And that is one reason why philosophers find puzzling Kant's strong advocacy of capital punishment (see, e.g., MS 334). In any case, that killing to get revenge is morally impermissible according to the Formula of Humanity, but not according to the Formula of Universal Law (at least on the readings of it we have discussed), suggests that, contrary to Kant's view, the one formula is not equivalent to the other.

A duty of beneficence would also seem to follow from the Formula of Humanity, although its derivation is not without difficulties. How does Kant arrive at this duty in the *Groundwork* (GMS 430)?[31] Humanity as Kant understands it is a set of capacities, including the capacity to set ends and pursue them. According to Kant, one end that each of us (i.e., each human agent) sets and pursues is that of his own happiness.[32] So, in concrete terms, valuing our humanity (as opposed to that of other rational agents, such as angels, who might not have their own happiness as an end) involves valuing our capacity to pursue happiness. To conform to the Formula of Humanity, then, an agent's actions must be consistent with his valuing this capacity.

But it seems that one does not really value a capacity unless one values its successful exercise. We would, for example, doubt whether someone truly valued the capacity of an acorn to grow if she denied that, other things being equal, it would be a good thing if it matured into an oak tree. However, this is a tricky point. It does not seem to be self-contradictory to value a capacity but not the "successful" exercise of it. Would there be anything irrational in valuing an acorn's capacity to grow but remaining indifferent as to whether it

just barely sprouted out of the ground or grew into a gigantic tree? Somewhat analogously, would there be anything irrational in valuing a person's capacity to pursue happiness but remaining indifferent as to whether he made little or much progress toward it?

Supposing that we grant that there would be something irrational in this, we see that an agent's actions are not consistent with his valuing humans' capacity to pursue their happiness unless they are consistent with his valuing their actually making progress toward happiness. Now an agent's actions are not consistent with this if he acts toward the goal of thwarting someone else in his pursuit of happiness, if we assume the other's pursuit is itself consistent with the other's appropriately valuing humanity. So an agent must refrain from acting toward this goal. If he thus refrains, then, in Kant's terms, his actions express "negative agreement" with humanity as an end in itself (GMS 430). According to Kant, however, the Formula of Humanity also requires "positive agreement" with humanity as an end in itself. An agent must also promote others' happiness. The idea here is that an agent's actions are not really consistent with his valuing others' progress toward happiness unless he aids them in making it.

Of course, this account of how a duty of beneficence would derive from the Formula of Humanity leaves important questions unanswered. For example, how robust a duty of beneficence would follow from the formula? Kant suggests that the formula requires that everyone try "as far as he can, to further the ends of others. For, the ends of a subject who is an end in itself must as far as possible be also *my* ends, if that representation is to have its *full* effect in me" (GMS 430). However, Kant considers beneficence to be an imperfect duty; and earlier in the *Groundwork* he characterizes a perfect duty as "one that admits no exception in favor of inclination," (GMS 421, note) apparently implying that an imperfect duty does admit of such an exception. If we must do all we can (morally permissibly do) to further the ends of others, how can we ever justifiably choose to satisfy our own inclinations (e.g., by watching a movie) instead of trying to promote the welfare of others (e.g., by working a few hours at a soup kitchen)?[33]

Our brief discussion of the duty of beneficence has illustrated that deriving duties from the Formula of Humanity is not a cut-and-dried business. Its difficulties will become more evident, I think, if we take a look at Kant's derivation of a duty of sincerity in promising.

According to Kant, the Formula of Humanity forbids an agent from making promises that he has no intention of trying to keep. "[H]e who has it in mind to make a false promise to others," says Kant, "sees at once that he wants to make use of another human being *merely as a means*, without the other at the same time containing in himself the end. For, he whom I want to use for my purposes by such a promise cannot possibly agree to my way of behaving toward him, and so himself contain the end of this action" (GMS 429–430). How are we to interpret this passage? According to Allen

Wood, Kant is arguing here that making a false promise would violate the Formula of Humanity since it would express disrespect for rational nature. "A false promise, *because its end cannot be shared by the person to whom the promise is made*, frustrates or circumvents that person's rational agency, and thereby shows disrespect for it."[34] Apparently, according to Wood, when Kant says that a promisee cannot "himself contain the end" of a false promisor's action, he is intimating that the latter cannot share the promisor's end. (Here "end" refers to an end to be produced.) That interpretation seems reasonable enough.

But what, precisely, does it mean to say that the promisee cannot share the promisor's end? Wood is not very helpful on this point. The end that the promisee *cannot* share is apparently the false promisor's end to deceive him into doing something (e.g., into giving him money), rather than the end for the sake of which the false promisor tries to deceive him into doing this thing. For the latter end *might* be that of diminishing world hunger, and there seems to be no reason why it would be impossible for the two to share *that* end. Perhaps in Kant's view the promisee cannot share the promisor's end of deceiving him into doing something in the sense that it would be irrational for him to share this end. Agents presumably share an end just in case each of them pursues the end. But, in ordinary circumstances, it would be irrational for the promisee to pursue the end of being deceived into doing something such as lending someone money. For this end's being brought about would prevent him from attaining other ends he is pursuing – for example, that of eventually buying himself a car.[35] The notion of irrationality at work here is familiar to us from our discussion of the Formula of Universal Law. In effect, if the promisee shared the false promisor's end, then the former would be willing that he be thwarted in attaining ends he is pursuing. In a practical sense, he would be irrational.

We might, then, take from Kant's discussion of false promising (GMS 429–430) that if an agent's action involves another, it expresses disrespect for the other's agency (and thus violates the Formula of Humanity) unless the other can share the agent's end. And the other can share the agent's end only if the other can pursue it without practical irrationality of the kind we have just described. To put the view briefly, a necessary condition for the moral permissibility of actions affecting others is that they be done to attain ends that others can share.

Unfortunately, there are difficulties with this view. First, suppose that Pete acts on the maxim: "In order to be the number-one ranked men's tennis player of the year, I will win every major tournament I enter." At first glance, it does not seem to be morally impermissible to act on this maxim. However, doing so might violate the Formula of Humanity as just interpreted. Acting on this maxim might frustrate some rival player's rational agency, thereby showing disrespect for it. For presumably some rival players cannot share Pete's end. Imagine that Pete and Andre are competing in the final of the

U.S. Open and that at stake is the number-one ranking for the year, which each player has as his goal. In pursuing the end of Pete's being number one – for example, by purposefully throwing the match – Andre would be willing to be thwarted in attaining his end of being number one. Andre cannot share Pete's end in the sense that it would be practically irrational for him to do so. In general terms, this reading of the Formula of Humanity has the following implication. Suppose an agent is pursuing an end in a competition. If his competitor cannot, rationally speaking, both pursue the agent's end and strive to secure his own end, the agent's action is morally impermissible.

One might respond that, although this implication initially seems to discredit the Formula of Humanity, reflection reveals otherwise. Granted, it would be worrisome if the Formula of Humanity entailed that pursuing an end in competitive sports (or some other competitive endeavor) is always wrong. But on the reading in question, the formula does not entail this. If Pete's end were not to be number one but to develop his capacities as a tennis player, then he would not be disrespecting Andre's agency. For this is an end that Andre can share. (Of course, if Andre perseveres in pursuing the end of being number one, then he is presumably violating the Formula of Humanity by disrespecting the rational agency of some other player who himself aims to be number one.) This reply has some force. According to reflective moral common sense, it seems, Pete and Andre would in some sense be more virtuous if each could share the other's end. Many of us do find the character of competitors who each have as an end to develop their own capacities morally more attractive than ones who each have as an end to defeat their rivals. There is something admirable in holding that, ultimately, one is "competing against" oneself. However, I think that ordinary moral reason would find unacceptably strong the judgment that it is *morally wrong* to act as Andre and Pete do in the example.

There is a second difficulty with the view that an agent's doing something to another expresses disrespect for the other (thus violating the Formula of Humanity), unless the other could, without practical irrationality, share the agent's end. Suppose a police officer has the end of preventing race-based attacks on law-abiding citizens. In pursuing this end, she arrests a white supremacist, someone she believes (correctly) to be planning an attack on a preschool frequented by Asian Americans. The difficulty is that in arresting the white supremacist she might be pursuing an end that he cannot share. Suppose, as the officer is aware, his end in planning the attack was to get revenge on a racial group that he thinks to be inferior to whites and thus undeserving of the rights and liberties its members possess. It would be practically irrational for him to pursue his end of revenge and at the same time to will the officer's end of preventing race-based attacks on law-abiding citizens. For, in willing the former, he would be thwarting his pursuit of the latter. Therefore, the view at issue forces us to embrace the counterintuitive

conclusion that, in making the arrest, the officer is expressing disrespect for the white supremacist's rational agency and thereby acting wrongly.

A natural reply to cases such as this is to fine-tune the view in question, perhaps by claiming the following. According to the Formula of Humanity, a person whom an agent is treating in a certain way must be able to share the agent's end, unless what would prevent his sharing the end is his acting in a morally impermissible way. On this modified view, it seems that the officer's arresting the white supremacist would conform to the Formula of Humanity. For what would prevent the white supremacist from sharing her end would be his (obviously immoral) attempt to get revenge.

But what is the standard by which we are supposed to determine whether the other is acting in a morally impermissible way? Perhaps Kant would appeal to the Formula of Universal Law, holding that a person whom an agent is treating in a certain way must be able to share the agent's end, unless what would prevent his sharing the end is his acting contrary to the Formula of Universal Law. But this appeal would be problematic. First, if in some cases such an appeal were necessary to make the Formula of Humanity work, then would it really be a viable candidate for the *supreme* principle of morality? The supreme principle is supposed to be such that all moral duties are derived ultimately from it, not from it in combination with some other moral principle. Of course, this difficulty would dissolve if, as Kant suggests, the two principles were equivalent. But as we have seen, it is very doubtful whether they are. Second, and more important, the Formula of Universal Law does not appear to be a reliable indicator of an action's moral permissibility. Indeed, the formula seems particularly ineffective in generating results that cohere with the ordinary conviction that actions such as that of the white supremacist are wrong. A maxim of attacking a racial minority to get revenge would seem to pass the Formula of Universal Law test. We proposed a modification in our understanding of when someone whom an agent treats in a certain way must (morally speaking) be able to share the agent's end. This modification is ineffective.

Perhaps the modification we need is not in our understanding of when someone must be able to share an end, but in our understanding of what it would mean to share an end. We have been employing an interpretation according to which someone shares an agent's end just in case he (actually) pursues the end. But this interpretation might be misguided. After all, in the false promising example, Kant says: "For, he whom I want to use for my purposes by such a promise *cannot possibly agree to my way of behaving toward him*, and so himself contain the end of this action" (GMS 429–430, emphasis added). Perhaps another's containing the end of an agent's action toward him (i.e., sharing this end) amounts to the other's being able to consent to the agent's pursuing his end in the way he does. (This strikes me as a rather tenuous sense of sharing an end, but so be it.) On this reading, the Formula of Humanity would escape the (in my view unwelcome) implication

that Pete's action toward Andre was morally impermissible. For there is no reason to suppose that Andre cannot consent to Pete's pursuing the end of being number one through beating him in the finals of the U.S. Open.

The obvious difficulty presented by this interpretation, however, is that of pinpointing what it means for a person to *be able to* consent to being treated in a certain way. It is clearly not the case that a necessary condition for an agent's being able to consent is that he would, upon reflection, consent if given the occasion to do so. Would the white supremacist, even if queried after calm deliberation, consent to the officer's action of trying to thwart his plot? Being able to consent in the requisite sense to being treated in a certain way must amount to being able, rationally speaking, to consent to it. But what does rational consent amount to? Echoing the preceding discussion, one might claim that a person can rationally consent to an agent's pursuit of his end just in case this pursuit would not in itself prevent the person from attaining his ends. Pete's *pursuit* of the number-one ranking would not itself block Andre from gaining this ranking. Yet the officer's pursuing his aim of preventing race-based attacks on law-abiding citizens through arresting the white supremacist may well preclude the latter from attaining his goal of revenge. So that strategy does not seem promising. One might instead claim that a person can rationally consent to an agent's pursuit of an end just in case this pursuit is morally permissible. Once again, however, we need to know what the standard of moral permissibility is supposed to be. If it is the Formula of Universal Law, then familiar difficulties arise. First, if we need to invoke this formula, then it is questionable whether the Formula of Humanity is really a candidate for the *supreme* principle of morality. Second, the Formula of Universal Law generates counterintuitive results. The white supremacist's maxim seems to pass its test, and thus we seem to arrive at the odious conclusion that his victims can rationally consent to being attacked. There is, I think, something philosophically attractive in the notion that an agent does not respect another's rational nature unless the other can rationally consent to the way he treats him. But specifying what is meant here by rational consent is a difficult task – one that I do not undertake here.

In the end, perhaps we need not focus at all on some inability of the recipient of a false promise to consent to or to share the promisor's end to explain why the promisor's action violates the Formula of Humanity. Here is a sketch of how one might proceed. To conform with this principle, the promisor's action must be compatible with his valuing the recipient's humanity as something absolutely and incomparably good. But in typical cases, making a promise to another that one has no intention of keeping is not compatible with so valuing the recipient's humanity. To value another's humanity is to value his capacity to set and to pursue ends. But if one values his

capacity to pursue ends, then, other things being equal, one must also value his capacity to pursue them successfully. (This move, which we discussed briefly in connection with Kant's derivation of a duty of beneficence, is admittedly controversial.) In a typical case, however, someone making a false promise realizes that he will thwart (or at least runs a significant risk of thwarting) the promisee in her pursuit of her ends. For example, the false promisor would realize that, because he obtains a loan from a person on the basis of a promise to her that he has no intention of keeping, the person will not have that money at hand to do what she wills with it. In short, making a false promise to someone expresses disrespect for the person's rational agency since it expresses indifference (or even contempt) for the agent's own projects.

8.9 Formula of Humanity: Further Challenges

The preceding section illustrated some of the challenges we face in deriving from the Formula of Humanity duties we take ourselves to have.[36] Although the details need to be worked out, it seems that the formula is capable of generating duties of beneficence and sincerity in promising. I will close my brief discussion of the Formula of Humanity by pointing out a few further hurdles defenders of it need to overcome if they are to establish that it fulfills Kant's eighth criterion for the supreme principle of morality.

First, a cluster of questions arise around Kant's claim that humanity has dignity and must be treated as such. As something with dignity, humanity has *incomparable* worth. According to Kant, it appears, this implies that it is never legitimate to treat humanity as a value to be exchanged for or compensated by either anything with mere price *or anything with dignity*. One fails to treat humanity as an end in itself in any situation in which one destroys the humanity in one person on the grounds that doing so is necessary to secure the "greater value" inherent in the humanity of two or more other persons. But what if the number of these other persons is 10, 1,000, or even 1 million, as it might be in some emergency? Even in extreme circumstances is it morally impermissible to treat humanity as a value to be sacrificed in order to secure more (even vastly more) of this very same value? It is not obvious that we would morally condemn the leader of a counterterrorist force for treating one innocent hostage held at the front of a plane as a value to be sacrificed (along with the value inherent in the terrorist) in order to preserve the greater value of the 350 remaining passengers and crew on board. Some emergency situations threaten to bring out disagreement with Kant's apparent view that the value of the humanity is not only incomparable to the value of things, but also to the value of other instances of humanity itself.[37]

In my view, though admittedly not in everyone's, it would be particularly damaging to the Formula of Humanity if it followed from it that it would never be permissible to kill one agent, where killing this agent would have the effect of saving (many) others. However, that it is wrong to treat the humanity in one individual as a value to be sacrificed for the sake of preserving the "greater value" inherent in the humanity of many others does not entail that it is morally impermissible to kill the one and thereby preserve the humanity in the many. It simply entails that one's *grounds* for killing the one cannot be that the humanity in him does not add up to a value as great as that of the humanity in the many others.[38] But the challenge then is to locate *other grounds* consistent with the Formula of Humanity's being the supreme principle of morality for killing in circumstances in which, according to common sense, this is appropriate.[39]

Another cluster of issues that must be addressed, if we are to see that the Formula of Humanity generates results acceptable to ordinary moral reason, concerns the formula's implications regarding how we must treat existing beings who do not have humanity (e.g., animals), as well as beings who do not yet exist but who will have humanity (future generations). This cluster of issues concerns the scope of humanity. In the *Groundwork* Kant contends that only persons, that is, beings who have humanity, are ends in themselves, and that all other beings "have only a relative worth, as means, and are therefore called *things*" (GMS 428). In the *Metaphysics of Morals*, he asserts that "a human being has duties only to human beings (himself and others), since his duty to any subject is moral constraint by that subject's will" (MS 442). Kant goes on to make clear that the term "human beings" here refers to "persons," beings who have humanity, and that all the persons of which we have experience *are* human beings. So, in effect, Kant is claiming here that a person has duties only to himself and to other persons; he has no duties to beings who do not possess the set of capacities that make up humanity. There seems to be no tension at all between this claim and the Formula of Humanity itself. After all, this formula commands us merely so to act that we treat *humanity* as an end in itself; it says nothing concerning how we must act toward beings without humanity. In light of Kant's suggestion that beings without humanity are valuable merely as means and that we have no duties to such beings, it might seem to follow that according to the Formula of Humanity, it is morally permissible to treat them as we will. Does the Formula of Humanity forbid, as ordinary moral reason (arguably) does, our causing tremendous pain to animals or to severely disabled human beings for the sake of making our lives a bit easier?

As readers of the *Metaphysics of Morals* are aware, Kant argues that, although we have no duty *to* beings without humanity, we do have duties *with regard to* such beings. More precisely, we have duties *to ourselves* that require us

to treat these beings in certain ways. For example, Kant argues that "violent and cruel treatment of animals" is opposed to "a human being's duty to himself . . . for it dulls his shared feeling of their suffering and so weakens and gradually uproots a natural predisposition that is very serviceable to morality in one's relations with other people" (MS 443). Kant's point here seems to be that by treating animals (and presumably severely disabled human beings) violently and cruelly, we desensitize ourselves to the suffering of persons, thus making it more difficult for us to fulfill our duties to them. As Kant himself suggests (MS 456–457), if we do not cultivate our capacity to recognize suffering in persons, then we might be less effective than we otherwise would be in fulfilling our duty of beneficence toward them. To help others effectively, we need to recognize when (and how) they need help. Training ourselves to share in their feelings of suffering can aid us in doing so. At any rate, Kant appears to base a prohibition on cruel treatment of animals on an agent's duties to other persons, so it is odd that he says it is based on an agent's duty to himself. Perhaps Kant has in mind that, by diminishing an agent's ability to fulfill his duty of beneficence, cruelty to animals would hinder his ability to fulfill a duty he discusses just a few pages later, namely his duty to himself to increase his moral perfection (MS 446). This duty requires one to strive to fulfill all of his duties, including, of course, that of beneficence.

Kant's claim that an agent must not be cruel to beings devoid of humanity, since doing so hinders her from fulfilling duties to the humanity in herself, may not satisfy reflective moral common sense. First, it seems to be a questionable thesis concerning human psychology that cruelty to animals (or, for that matter, to the severely disabled) always "weakens and gradually uproots" an agent's ability to feel the suffering of other persons. Is the equestrian who whips her horse in a competition necessarily diminishing her capacity to empathize with her fellow persons? Second, and more important, some might question whether the reason we should avoid treating nonpersons cruelly is really that (or just that) we thereby make it more difficult for us to fulfill our duties to ourselves, rather than that (or also that) we make them suffer unnecessarily. Unnecessary suffering, they might say, is a bad thing, whether it be the suffering of a horse or of a man in the late stages of Alzheimer's disease. Kant writes that "agonizing physical experiments [on animals] for the sake of mere speculation, when the end could also be achieved without these, are to be abhorred" (MS 443). But are they to be abhorred, as he suggests, simply because they diminish the experimenter's capacity to fulfill his duties to himself, or (also) because they cause needless pain to sentient beings?

The passage in the *Metaphysics of Morals* we have briefly discussed raises questions not only regarding the Formula of Humanity's implications concerning the treatment of existing beings devoid of humanity but also

concerning obligations we (might) take ourselves to have to future generations. Cited more fully, the passage reads:

> [A] human being has duties only to human beings (himself and others), since his duty to any subject is moral constraint by that subject's will. Hence the constraining (binding) subject must, *first*, be a person; and this person must, *secondly*, be given as an object of experience, since the human being is to strive for the end of this person's will and this can happen only in a relation to each other of two beings that exist (for a mere thought-entity cannot be the *cause* of any result in terms of ends). (MS 442)

We might get the impression here that according to Kant an agent has duties only to existing persons, not to any persons who will exist in the future. For the latter beings seem merely to be "thought-entities." This impression is not dissipated by the Formula of Humanity itself, which seems compatible with the view that we have no duties to future generations. For it merely commands that we so act that we treat the humanity *in* ourselves and *in* any other as an end in itself. But, arguably, since future generations do not yet exist, there is no humanity *in* them. If Kant's theory, specifically his advocacy of the Formula of Humanity as the supreme principle of morality, implies that we have no duties to future generations (of persons), then it might clash with reflective moral common sense. For many of us do hold that we have such duties – for example, a duty not to pollute the environment to such an extent that our descendants (none of whom now exist) will live in a quagmire of disease and malnutrition, and thus be unable to effectively pursue their happiness.

A first step toward meeting this challenge, which seems to pose fewer difficulties than Kant's views toward animals, the severely disabled, and so forth, might be to examine the dialectical context in which Kant's remarks occur. Kant's suggestion that persons do not have duties to mere "thought-entities" appears merely to be a premise in an argument he aims against the notion that we have duties to God. Shortly after making this suggestion, Kant claims that "we do not have before us, in [the idea of God], a given being to whom we would be under obligation; for in that case its reality would first have to be shown (disclosed) through experience" (MS 444). We can have obligations only to beings belonging to a kind which is such that we can experience its members' reality, Kant seems to be arguing. God is not such a being; therefore we cannot have obligations to God. Whatever the merits of this argument, it might be compatible with the notion that we have duties to future generations. For the future generations in question do belong to a kind, that of persons, which is such that we can experience its members' reality; we experience the reality of persons every day. The Formula of Humanity commands that we treat the humanity in ourselves and in others as an end in itself. I see no reason why a defender of the Formula of Humanity could not suggest that the scope of "others" include the future generations of persons whom we can reasonably be assumed to affect.

In sum, to fulfill Kant's eighth criterion, a candidate for the supreme principle of morality must generate moral prescriptions that square with those uncontroversially embraced by reflective moral common sense. The Formula of Humanity faces some difficulties on this score. But the prospects for it seem much brighter than those for the Formula of Universal Law.

8.10 Where We End Up

The argument of this book has taken shape against the background of the traditional reading of Kant's *Groundwork* derivation of the Formula of Universal Law. Kant's derivation fails miserably, according to this reading, since it involves a leap from a practically uninformative principle to the Formula of Universal Law. Kant is left with embarrassingly inadequate support for one of the foundational claims in his ethics. I have argued that we respond effectively to the traditional reading neither by appealing to the second *Critique* derivation of the Formula of Universal Law, as reconstructed by Allison, nor by focusing solely on Kant's derivation of the Formula of Humanity, as reconstructed by Korsgaard. Both of these reconstructed derivations suffer from fundamental flaws. We should instead challenge the traditional interpretation itself.

The central thesis of the book can be crystallized into a few sentences: There is a textually plausible reading of Kant's *Groundwork* derivation of the Formula of Universal Law, namely the criterial reading, that shows this argument to be far more philosophically engaging and forceful than does the traditional reading. With the help of this argument, Kant makes a convincing case against some key rivals to the Formula of Universal Law (e.g., some consequentialist principles). Even though in the end the Formula of Universal Law has serious and probably fatal shortcomings as a candidate for the supreme principle of morality, an argument of the sort Kant employs in deriving it (a criterial argument) holds substantial promise as a way of defending his Formula of Humanity – a principle that many philosophers, including me, find especially attractive.

Kant's criteria for the supreme principle of morality have been at the core of the discussion. By appealing to them Kant shows several rivals to his formulas not to be viable candidates for status as the supreme principle. So it might be helpful to summarize some main findings regarding Kant's criteria.

Four of them belong to Kant's basic concept of the supreme principle of morality. According to this concept, the supreme principle of morality must be (i) practical, (ii) absolutely necessary, (iii) binding on all rational agents, and (iv) the supreme norm for the moral evaluation of action. These criteria, which are all at least implicit in the *Groundwork* Preface, have not received nearly as much attention as the ones Kant develops in the course of *Groundwork* I. That is because the focus of the book has been the claim that if there is a supreme principle of morality, *in the basic sense of such a principle that Kant employs,* then it is the Categorical Imperative.

Although I do not pretend to have defended criteria i and ii here, I agree with Kant that according to reflective moral common sense, the supreme principle of morality would have to be both practical and absolutely necessary. It would have to be something on account of which agents might actually act – not, for example, merely a theoretical tool to be used by experts to determine the rightness of actions after they occur. It would also have to be a principle that each agent ought to obey no matter what she desired.

However, I have made a couple of critical points in connection with the criteria implicit in Kant's basic concept of the supreme principle of morality. First, and very briefly, Kant implies that if we embrace criterion ii, then we must embrace criterion iii. In other words, if we agree that the supreme principle of morality must be absolutely necessary – that is, unconditionally binding on the agents within its scope – then we are compelled to agree that the principle must have a scope that extends to all rational agents, including any nonhuman ones such as angels. Apparently, Kant believes this point to be obvious. In section 2.4 (in connection with Henry Allison's reconstruction of Kant's second *Critique* derivation of the Formula of Universal Law) I protested that it is not. Kant owes us an explanation of why a principle could not be unconditionally binding on all human rational agents – that is, one that each of us ought to obey no matter what her inclinations might be, yet not binding on some other type of rational agent, for example, a type that is necessarily incapable of conforming to it. I do not have any particular reason for rejecting iii; it is just that, contrary to Kant, I do not believe that it follows quickly and easily from ii.

More important, I have argued that unless Kant modifies criterion iv, his own candidates for the supreme principle of morality face elimination on the grounds that they manifestly fail to meet it (section 8.2). At least in the *Groundwork*, Kant suggests that a principle is the supreme norm for the moral evaluation of action only if just those actions done in accordance with the principle can have moral worth. But as I have contended, Kant is rationally compelled to hold that actions done in accordance with an indefinite number of different principles can have moral worth, as long as the actions are done from duty. (An agent acts from duty just in case her [in itself sufficient] incentive for acting stems from the notion that a principle, represented by her as a law, requires the action; she acts against the background of conscientious reflection; and she does her best to realize her action's end.) Kant needs to modify criterion iv so that it demands that the supreme principle of morality be such that actions' moral permissibility and requiredness, but not their moral worth, be defined in terms of this principle.

Let me now turn to the criteria for the supreme principle of morality that go beyond those contained in Kant's basic concept of this principle. Since the end of Chapter 6, we have been operating with four further criteria.

The supreme principle of morality must be such that: (v) every case of willing to conform to it because the principle requires it has moral worth; (vi) the moral worth of willing to conform to the principle because the principle requires it stems from its motive, not from its effects; (vii) an agent's representing the principle as a law – that is, a universally and unconditionally binding principle – provides him with sufficient incentive to conform to it; and, finally, (viii) a plausible set of duties (relative to ordinary rational moral cognition) can be derived from the principle. I will begin the discussion with criterion viii and work my way to v.

Much of this chapter has been devoted to the question of whether Kant's candidates for the supreme principle of morality should be eliminated on the basis of a manifest failure to fulfill viii. This is, I think, an important question, since I believe that viii is a criterion that Kant is correct to embrace. Given that Kant bases his *Groundwork* I derivation on reflective moral common sense, it would, I think, be unwarranted for him to endorse a candidate for the supreme principle of morality that generated prescriptions totally unacceptable to common sense (and that thus violated viii). After all, what grounds would Kant have for trusting ordinary moral consciousness in its endorsement of several criteria for the supreme principle of morality, yet ignoring its view of which duties would stem from a plausible candidate for this principle?

Kant suggests two arguments in support of vii (see section 5.7). One of the arguments turns on a dictum that Kant should abandon, namely (a particular understanding of) "ought implies can" (see section 8.2). But the other does not turn on this dictum. According to this argument, denying vii would amount to holding that the supreme principle of morality must be such that one's expectation of the effects of conforming to it necessarily constitutes (at least part of) his incentive for conforming to it. Now consider an agent who denied vii and embraced some principle as a viable candidate for the supreme principle of morality. He would be committed to the view that the *worth* of his conforming to the principle necessarily derived (at least in part) from its results. For if he thought that conforming to the principle was valuable in itself, then he would not hold that he *necessarily* needs to look to the action's results to find a sufficient incentive to do so. But if the agent ties inextricably the worth of his conforming to a principle to its results, then he is rationally compelled to deny that his conforming to it can have intrinsic worth. He is rationally compelled to deny a basic tenet of reflective moral common sense. Therefore, he must agree that we can hold a principle to be the supreme principle of morality only if we can maintain that our representing it as a law governing our actions gives us a sufficient incentive to conform to it.

It would be remiss not to acknowledge this argument to be controversial. (Let me reiterate that the argument is not one Kant explicitly gives but rather one that I believe he suggests.) One possible difficulty with it is the following.

The argument turns on the notion that, rationally speaking, an agent cannot hold conforming to a certain principle to be intrinsically valuable, yet at the same time hold that she needs necessarily to rely on the prospect of the actions' effects in order to have sufficient incentive to conform to it. But this notion is disputable. Recall the principle EU: "Always perform a right action, one that you expect will yield as great a sum total of well-being as would any alternative action available to you." Suppose that according to a particular agent, her conforming to EU because she expects that doing so will maximize aggregate well-being is intrinsically valuable. The agent holds further that in order to have sufficient incentive to conform to EU, she not only does but must rely on the prospect of maximizing aggregate well being.[40] It is not obvious that there is anything irrational in this agent's views. At the very least, the argument in question would need to be bolstered in order to deal with cases such as this. In any event, if I am correct, Kant does not have to rely on vii in order to eliminate rivals to his candidates for the supreme principle of morality.

But he does need to rely on criteria v and vi. It almost goes without saying that, in my view, v and vi are very plausible criteria for the supreme principle of morality. I have discussed them in detail. So here I would like merely to emphasize that v represents a *modification* of a criterion Kant advocates in the *Groundwork*. According to Kant, the supreme principle of morality must be such that all and only actions conforming to this principle because the principle requires it (i.e., all and only actions done from duty) have moral worth. I have advocated two significant changes to this criterion. First, Kant must acknowledge that some actions contrary to duty can be from duty and can thus have moral worth (section 6.6). He needs to hold, as criterion v contends, that the supreme principle of morality must be such that every case of *willing* from duty to conform to it has moral worth. Second, Kant offers no good reason for holding that only actions done from duty have moral worth. He does not undermine the view that actions done with an overriding commitment to morality but from other motives (e.g., sympathy) have such worth (6.10). (Recall that on the account I have sketched, an agent has an overriding commitment to morality just in case he acts against the background of conscientious reflection, and if after such reflection, he determines that an action is contrary to what he takes to be morally required, he will for this reason refrain from performing it.) In my view, Kant fails to establish that the supreme principle of morality must be such that only instances of willing to conform to it because the principle requires it have moral worth. Yet this failure has no significant impact on his derivations. With the help of other criteria, especially v and vi, he can construct powerful arguments against many rivals to his candidates for the supreme principle of morality.

It is one thing for Kant to show that rivals founder as candidates for the supreme principle of morality; it is quite another for him to establish

that his formulas remain afloat. Fulfilling criterion viii in particular poses a formidable challenge. I have offered reasons for thinking that the Formula of Universal Law fails to generate prescriptions acceptable to reflective moral common sense. And I believe it remains an open question whether the Formula of Humanity succeeds. It would be irresponsibly optimistic to claim that Kant has demonstrated that if there is a supreme principle of morality, then it is the Formula of Humanity.

A down-to-earth approach to the Kantian project in ethics emerges from this book. To advance toward the ideal of showing that if there is a supreme principle of morality, then it is some particular principle, it does not suffice to rely merely on abstract premises such as that we are transcendentally free rational agents or that a categorical imperative requires an unconditionally good "ground." We need to enter concrete controversies regarding which duties the principle would generate and whether these duties would be acceptable to reflective moral common sense. In searching for the supreme principle of morality, we need to follow the twists and turns of everyday moral experience. There is no royal road to a successful derivation.

Notes

Introduction: Derivation, Deduction, and the Supreme Principle of Morality

1. By Kant's "critical writings" in ethics, I mean simply the works in the field he published after the appearance of the *Critique of Pure Reason*.
2. In *Groundwork* II, Kant says that the "the categorical imperative," the principle he takes to be the supreme principle of morality, is "the canon of moral appraisal of action in general" (GMS 424). On the next page (GMS 425), Kant says: "we have . . . set forth distinctively and as determined for every use the content of the categorical imperative, which must contain the principle of all duty (if there is such a thing at all)."
3. As Schopenhauer notes with irritation, Kant never seems to tire of reminding us of this feature. See Arthur Schopenhauer, *On the Basis of Morality*, tr. E. F. J. Payne (1841; Providence: Berghahn Books, 1995), 63.
4. Evidence that Kant holds this can be found in his *Metaphysics of Morals* discussion of ends that are also duties (e.g., MS 386). Kant argues that an agent cannot have an obligation to promote the end of his own happiness, since each agent unavoidably has this end.
5. I discuss Kant's notions of the will and its determining grounds in Chapter 1.
6. "Common rational moral cognition" or, translated slightly differently, "ordinary rational knowledge of morals" is Kant's starting point in the *Groundwork*. He entitles Section I "Transition from common rational to philosophic moral cognition" (GMS 393).
7. In the cited passage, Kant moves seemingly without argument from the notion that a moral law must be absolutely necessary to the notion that it must have wide universal validity. I think this move is problematic, as I explain in section 2.4.
8. Allison appears to employ this usage of "derivation." See Henry E. Allison, *Idealism and Freedom* (Cambridge: Cambridge University Press, 1996), 143–144.
9. See Henry E. Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990), 214.
10. I have adapted this notion of moral particularism from O'Neill's helpful discussion. See Onora O'Neill, *Towards Justice and Virtue* (Cambridge: Cambridge University Press, 1996), 11–13.

11. For some interesting arguments against moral particularism, see ibid., especially chap. 3.
12. Rüdiger Bittner, *What Reason Demands* (Cambridge: Cambridge University Press, 1989), 89.
13. See also KpV 93 on this point in particular.
14. In my view, *Groundwork* III is one of the most enigmatic texts Kant published. The argument is hard to follow, and philosophers differ significantly in how they interpret it. For reconstruction and criticism of Kant's argument (or crucial aspects of it), see Allison, *Kant's Theory of Freedom*, 214–229; Karl Ameriks, "Kant's Deduction of Freedom and Morality," *Journal of the History of Philosophy* 19 (1981): 45–65; Rüdiger Bittner, "Wer frei ist, ist gebunden," in *Philosophiegeschichte und Logische Analyse*, ed. U. Meixner and A. Newen (Paderborn: Mentis, 2000), 209–221; Dieter Henrich, "Die Deduktion des Sittengesetzes," in *Denken im Schatten des Nihilismus*, ed. Alexander Schwan (Darmstadt: Wissentschaftliche Buchgesellschaft, 1975), 55–112; Allen W. Wood, *Kant's Ethical Thought* (Cambridge: Cambridge University Press, 1999), 171–182.
15. For Aune's discussion of a gap in the *Groundwork* II derivation, see Bruce Aune, *Kant's Theory of Morals* (Princeton: Princeton University Press, 1979), 42–43.
16. Ibid., 28–29.
17. Ibid., 30.
18. Ibid., 31.
19. As Aune acknowledges (29), this statement of the Categorical Imperative, which stems from *Groundwork* II (GMS 421), differs a bit from the one Kant actually offers in *Groundwork* I. The statement Kant gives there is this: "*I ought never to act except in such a way that I could also will that my maxim should become a universal law.*" The differences between the two statements are not important here.
20. Aune, *Kant's Theory of Morals*, 30; see also 32 and 86–87.
21. Ibid., 34.
22. See, for example, Allison, *Idealism and Freedom*, 144, 150; David Cummiskey, *Kantian Consequentialism* (Oxford: Oxford University Press, 1996), 57; Thomas E. Hill Jr., "The Rationality of Moral Conduct," in *Dignity and Practical Reason in Kant's Moral Theory* (Ithaca: Cornell University Press, 1992), 121–122; Allen W. Wood, *Hegel's Ethical Thought* (Cambridge: Cambridge University Press, 1990), 164–166; and Wood, *Kant's Ethical Thought*, 47–49, 81–82.
23. Wood, *Kant's Ethical Thought*, 81.
24. See Allison, *Idealism and Freedom*, 146, for a slightly more complex version of this principle.
25. Ibid., 145.
26. I discuss the moral impermissibility of acting on this maxim of false promising in Chapter 8.
27. I take it that this is the sense (or at least one of the senses) in which Kant employs the notion of a categorical imperative in *Groundwork* II before he officially introduces "the categorical imperative." See GMS 416. A principle that was a categorical imperative in this sense, that is, an unconditionally and universally binding principle, would presumably not manifest itself *as an imperative* to beings incapable of violating it (e.g., God and angels).
28. In taking this to be a statement of the Formula of the Kingdom of Ends, I am following O'Neill. See Onora O'Neill, *Constructions of Reason* (Cambridge: Cambridge University Press, 1989), 127.

29. I will also leave aside another formula Kant introduces, namely what has been referred to as the Formula of Autonomy: "[C]hoose only in such a way that the maxims of your choice are also included as universal law in the same volition" (GMS 440). See O'Neill, *Constructions of Reason*, 53. According to Allen Wood, Kant presents the Formula of the Kingdom of Ends as a "more intuitive version" of the Formula of Autonomy. See Wood, *Kant's Ethical Thought*, 163–166.
30. Of course, Kant holds at the very least that both the Formula of Universal Law and the Formula of Humanity each generate the duties (roughly) not to commit suicide, not to make false promises, to develop one's natural talents, and to help others in need. See GMS 421–424, 429–430.

Chapter 1: Fundamental Concepts in Kant's Theory of Agency

1. MS 211–213 gives one a taste of just how challenging Kant's discussions of agency can be.
2. "*Maxime* ist das subjective Prinzip zu handeln" (GMS 421, note). See also GMS 400, note.
3. The discussion of maxims in this section has been heavily influenced by Bittner. Bittner disagrees, however, with the view I set out that, when fully described, maxims incorporate descriptions of an agent's end and incentive in acting. See Rüdiger Bittner, *Doing Things for Reasons* (Oxford: Oxford University Press, 2001), chap. 3.
4. This maxim is, of course, a slight variant of one that Kant employs in the *Groundwork*. See, for example, GMS 422.
5. If nobody held M, then it would be a possible or potential maxim.
6. I am assuming here that Kant is justified in his view that acting on M is forbidden by the Categorical Imperative. See GMS 421–422.
7. See, for example, Onora O'Neill, *Constructions of Reason* (Cambridge: Cambridge University Press, 1989), 151.
8. Of course, on Kant's view all morally *obligatory* actions also count as morally *permissible* ones. That Kant held all of an agent's actions to be either morally permissible or impermissible becomes clear in the *Religion*. In the main text, Kant says: "It is . . . of great consequence to ethics in general to avoid admitting, so long as it is possible, of anything morally intermediate, whether in actions (*adiaphora*) or in human characters" (Rel 22, English ed. 18). This suggests that Kant does not *want* to admit there to be actions that are neither morally permissible nor impermissible. And, as a matter of fact, central claims he makes in the *Religion* (and elsewhere in his practical philosophy) prevent him from admitting it. Kant holds that all human action is free and that free action is action that does not result merely from natural laws (Rel. 41, English ed. 36). But if there were human actions that were neither morally permissible nor impermissible, claims Kant, then they would result merely from natural laws (Rel 23, note; English ed. 18, note): they would not be free. Therefore, according to him, there are no human actions that are neither morally permissible nor impermissible.
9. On this point, see Nelson Potter, "Maxims in Kant's Moral Philosophy," *Philosophia* 23 (1994): 62, 71.
10. The term "incentive" translates the German term *Triebfeder*. Translated literally *Triebfeder* means something like "push spring." (A *Triebfeder* in a clock would be

the spring that makes it run.) Of course, in contemporary English "incentive" can mean something like object to be gained, as when we say that the child's incentive to clean her room is a trip to the amusement park.

11. At Rel 35 (English ed. 30), Kant says that "in the absence of all incentives the will [*Willkür*] cannot be determined."
12. For discussion of this point, see Potter, "Maxims in Kant's Moral Philosophy," 63.
13. Suppose that, after much reflection, an agent comes to believe that her incentive in performing a particular action was the notion that it was morally required. In the *Groundwork*, Kant claims that it is impossible for an agent *to know* that her incentive in performing the action was the notion that it was morally required (was "duty"), rather than some "covert impulse of self-love" (GMS 407). When coupled with the reading of maxims I have suggested, this claim implies that, in cases where it appears to an agent that she has acted from duty, the agent cannot know which maxim she acted on. For if the maxim were fully specified, it would include a description of her incentive. This implication would be problematic if it turned out that in some cases, since an agent could not fully specify the maxim of her action, she would not be able by using the Categorical Imperative to determine whether her action was morally permissible. For if such cases might arise, then it is hard to see how Kant could maintain as he does that the Categorical Imperative is the canon of the moral estimation of all of our action (GMS 423). Perhaps Kant holds that an agent is always capable of specifying each of the elements in her maxim besides her incentive and that, through specifying these elements, she would always get a sufficient grasp on her maxim to gain a reliable indication of its moral permissibility from the Categorical Imperative test.
14. As I indicated, the conclusion that acting on the maxim in question would be impermissible is based on a traditional reading of the Categorical Imperative. For a different reading of it, according to which acting on this maxim would not turn out to be morally impermissible, see Thomas W. Pogge "The Categorical Imperative," in *Kant's "Groundwork of the Metaphysics of Morals": Critical Essays*, ed. Paul Guyer (Lanham, Md.: Rowman & Littlefield, 1998), 189–196. As I argue in Chapter 8, this reading has problems of its own.
15. This account of what differentiates maxims from other rules of the same form has been heavily influenced by O'Neill who says: "The maxim of an act is the principle that *governs* the selection of *ancillary* principles of action that express or implement the maxim in a way that is adjusted to the agent's (perceived) circumstances." O'Neill, *Constructions of Reason*, 129; see also 151–152. Another proposal regarding what distinguishes maxims from other rules of the same form has been made by Bittner. He suggests that maxims are "rules of life" (*Lebensregeln*), whereas the others are not. (See Rüdiger Bittner, "Maximen," in *Akten des 4. Internationalen Kant-Kongresses*, ed. G. Funke and J. Kopper [Berlin: De Gruyter, 1974], 489.) An agent's maxims express the kind of life he wants to lead, the course that he intends his life as a whole to take. If Kant did indeed hold maxims to be *Lebensregeln*, it is clear that he would deny status as a maxim to the rule "From self-love, every Monday at 3 P.M. I take live karate lessons in order to improve my endurance and flexibility." For surely this rule does not express an agent's conception of the direction he wants his life as a whole to

take. Yet the more general rule of keeping oneself in shape by exercising during one's free time does presumably express such a conception. However, there are, I believe, fairly good grounds for rejecting the view that for Kant all maxims are *Lebensregeln*. Consider, once again, the practical rule of false promising that Kant discusses in the *Groundwork*: "When I believe myself to be in need of money, I shall borrow money and promise to repay it, even though I know that this will never happen." This practical rule is not a *Lebensregel*. Granted, if we knew that an agent acted on it, we would have a clue as to what his *Lebensregeln* might be like. We might, for example, suspect that he has adopted one akin to: "Whenever my happiness is threatened, if need be I will lie in order to secure it." But the false-promising rule does not, *in itself*, express the direction that an agent who acts on it wants his life as a whole to take. (Allison makes this point. See Henry E. Allison, *Kant's Theory of Freedom* [Cambridge: Cambridge University Press, 1990], 92–93.) Kant refers explicitly to this rule as a *maxim* (GMS 422). Since it is not a *Lebensregel*, it appears that Kant does not think of all maxims as *Lebensregeln*.

16. At KpV 32 Kant identifies the will (*Wille*) of rational beings with "the ability to determine their causality by the representation of rules." See also KpV 45, 55, 58–59, 125. For examples of such usage of *Wille* in the *Groundwork*, see GMS 412, 427. Laberge offers an excellent discussion of Kant's definition of the will at GMS 412. See Pierre Laberge, "La définition de la volonté comme faculté d'agir selon la représentation des lois," in *Grundlegung zur Metaphysik der Sitten; Ein kooperativer Kommentar*, ed. Otfried Höffe (Frankfurt am Main: Klostermann, 1989), 83–96.
17. In using the terms "executive *Wille*" and "legislative *Wille*," I am following (roughly) the suggestion of Beck. See Lewis White Beck, *A Commentary on Kant's "Critique of Practical Reason"* (Chicago: University of Chicago Press, 1960), 202.
18. For example, Greene and Hudson use this translation in their edition of the *Religion within the Limits of Reason Alone*.
19. Some translators render *Willkür* as the "[power of] choice," thus implying that when one chooses to act, one exercises *Willkür*. (See, e.g., Gregor's translation of the *Metaphysics of Morals*, 650.) This rendering is potentially misleading. For Kant, exercising *Willkür* involves acting on choice: it involves choosing to realize an object and acting in the sense of *trying* to realize it. Some readers (myself included) conceive of choosing to realize an object as distinct from trying to realize it: one might do the former (e.g., choose to see a certain play) without doing the latter (e.g., making any effort to see it). That Kant takes exercising *Willkür* to involve trying to realize a chosen object is, I think, suggested in his (admittedly dense and difficult) definition of *Willkür* at MS 213. For a detailed discussion of this definition, see Samuel J. Kerstein, "Action, Hedonism, and Practical Law: An Essay on Kant" (Ph.D. diss., Columbia University, 1995), 30–34. Now I have suggested that, for our purposes, it is safe to consider exercising *Willkür* to be the same thing as exercising executive *Wille* (i.e., acting on self-given rules). But one might think that it is possible for an agent to exercise *Willkür* (i.e., to act on choice), without acting on any rule. However, as we have already seen, for Kant all of our (rational agents') acting is acting on some rule, that is, some maxim. Therefore, any time an agent exercises her *Willkür*, she also exercises her executive *Wille*.

20. I am *not* here invoking the distinction Kant draws at GMS 427 between an incentive and a motive (*Bewegungsgrund*). Kant there writes: "The subjective ground of desire is an *incentive*; the objective ground of volition is a *motive*; hence the distinction between subjective ends, which rest on incentives, and objective ends, which depend on motives, which hold for every rational being." Kant himself does not appear to maintain this distinction. See, for example, KpV 72, where he writes at length concerning the moral law as incentive.
21. Here I am following Allison. See Henry E. Allison, *Idealism and Freedom* (Cambridge: Cambridge University Press, 1996), 131. I discuss this notion further in section 2.2.
22. See Allison, *Kant's Theory of Freedom*, 204. Also see Barbara Herman, *The Practice of Moral Judgment* (Cambridge, Mass.: Harvard University Press, 1993), 217–218.
23. Material in sections 1.6–8 stems from Samuel J. Kerstein, "Kant's (Not So Radical) Hedonism," in *Kant und die Berliner Aufklärung. Akten des IX. Internationalen Kant-Kongresses,* vol. 3, ed. V. Gerhardt, R.-P. Horstmann, and R. Schumacher (Berlin: Walter de Gruyter, 2001), 245–253.
24. See, for example, T. H. Green, *Lectures on the Philosophy of Kant*, in *Works of Thomas Hill Green*, vol. 2, ed. R. L. Nettleship (London: Longmans, Green, 1900), 141; Terence Irwin, "Kant's Criticisms of Eudaemonism," in *Aristotle, Kant, and the Stoics*, ed. S. Engstrom and J. Whiting (Cambridge: Cambridge University Press, 1997), 74; A. Phillips Griffiths, "Kant's Psychological Hedonism," *Philosophy* 66 (1991): 211; Bernard Williams, *Ethics and the Limits of Philosophy* (Cambridge: Cambridge University Press, 1985), 15, 64.
25. Griffiths, "Kant's Psychological Hedonism," 212.
26. See Andrews Reath, "Hedonism, Heteronomy and Kant's Principle of Happiness," *Pacific Philosophical Quarterly* 70 (1989): 42–72. Christine Korsgaard seems to embrace Reath's reading. See Christine M. Korsgaard *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 56.
27. See Reath, "Hedonism, Heteronomy and Kant's Principle of Happiness," 46–49.
28. Perhaps Allison suggests the alternative interpretation in a discussion of Kant's "philanthropist" (or "friend of humanity"). See Allison, *Kant's Theory of Freedom*, 103. Allison says: "Thus, whereas helping others in need provides [the philanthropist] with satisfaction and he would not act in that way unless it did so, the end he has in mind is nevertheless the improvement of the lot of others and not the satisfaction of his own needs." I am unsure whether Allison is indeed urging the alternative interpretation, since the context of his remark makes it doubtful whether by "satisfaction" he means pleasure, as the alternative would require. But if he is suggesting the alternative interpretation, then my aim here could be construed as developing his suggestion by (a) distinguishing it from Reath's proposal and (b) shedding light on its textual basis.
29. Beck treats the concept of the capacity of desire in a way that is typical for contemporary scholars. In his classic commentary on the second *Critique*, he employs the concept, but does not focus on explicating it. See, for example, Beck, *A Commentary on Kant's "Critique of Practical Reason,"* 94–97.
30. When an agent exercises her capacity of desire, she tries to realize an object. Kant, it is worth noting, does not exclude mental objects from the set of those

an agent might try to realize in exercising this capacity. For example, guided by my idea of remembering the name of my first mathematics teacher, I might try to remember it.

31. Of course, an agent's *pathological interest* in something is not necessarily an interest that stems from illness or an interest in something unhealthy. In this context, "pathological" does not connote disease. It means rather *sensory*, that is, having to do with sensation.
32. In Kant's vocabulary, the term "agreeableness" (*Annehmlichkeit*) designates a kind of sensation. See, for example, KpV 22, where Kant speaks of the "sensation of agreeableness." To say that something is agreeable to a person is to say that it enables him to experience the sensation of agreeableness. Furthermore, on Kant's view, to experience the sensation of agreeableness is to experience pleasure: agreeableness is a kind of pleasure. For the most part, Kant seems to use the terms "pleasure" (*Lust*) and "agreeableness" interchangeably. See, for example, KpV 23. It appears, however, that "pleasure" has a wider extension than "agreeableness," since Kant employs the term "moral pleasure" (e.g., at MS 378) but, to my knowledge, not "moral agreeableness."
33. According to one common usage, the German *sofern* is equivalent to the German *im Fall, daß* (in case that) or *vorausgesetzt, daß* (supposing that) (Gerhard Wahrig, ed., *Deutsches Wörterbuch* [Gütersloh: Bertelsmann, 1986], 1188). See also Jacob Grimm and Wilhelm Grimm, eds., *Deutsches Wörterbuch*, vol. 10, pt. 1 (Leipzig: Hirzel, 1905), 1402. Not surprisingly, we find Kant using *sofern* in a way that makes it seem equivalent to "only if." For example, at MS 223 Kant says: "An action is called a *deed* insofar [*sofern*] as it comes under obligatory laws." But *sofern* can also mean "to the extent that." See Grimm and Grimm, 1402.
34. The conclusion that Kant's claim regarding acting from inclination amounts to this seems to be supported by the following assertion Kant makes at GMS 442, note: "every empirical interest promises to contribute to our well-being by the agreeableness that something affords, whether this happens immediately and without a view to advantage or with regard for it."
35. Reath, "Hedonism, Heteronomy and Kant's Principle of Happiness," 50.
36. In his *Anthropology*, Kant also defines inclinations as habitual desires. See Anth 251 and 265. In the *Religion*, Kant employs a different, narrower sense of inclination. He states that to have an inclination, we must be acquainted with the object of our desire. He contrasts inclination with instinct, "which is a felt want to do or to enjoy something of which one has as yet no conception (such as the constructive impulse in animals, or the sexual impulse)" (Rel 28–29, note; English ed. 24, note).
37. The question may arise of why Kant here writes of desire in the narrow sense: what would be desire in the broad sense? In my view, Kant holds that both actions done from duty and ones not done from duty involve an agent's determining his capacity of desire. To put the point roughly, both kinds of action involve the agent's choosing to realize some object. An agent has a desire in the broad sense for an object when he has determined his capacity of desire with respect to this object. Yet only if his ground for determining his capacity of desire included the prospect of his own pleasure does the agent have a desire in the narrow sense for the object. If the agent's ground for choosing to realize some object

did not include the prospect of his own pleasure – for example, if the agent set himself to realize it solely because he thought doing so was commanded by the Categorical Imperative – then the agent has a desire in the broad sense for the object but not a desire in the narrow sense for it. Of course, this reading of Kant's notion of desire in the narrow sense rests on my particular analysis of MS 212.

38. See Reath, "Hedonism, Heteronomy and Kant's Principle of Happiness," 47.
39. At this point, a defender of Reath's interpretation might make the following argument: "Granted, the capacity of desire is not the capacity to have a desire. Nevertheless, to say that an agent's capacity of desire has been "determined" is only to say that she has come to have a desire. It is *not* to say that she has, through her representation of an object, chosen to realize it. Therefore, as Reath argues, Kant's *Metaphysics of Morals* account merely suggests that pleasure plays a role in our coming to have inclinations." In response to this argument, let me say that I find no evidence that Kant conceived of the determination (*Bestimmung*) of the capacity of desire in this way. Actually, there is evidence that he conceived of it in the way I suggest. In a note to the *First Introduction to the Critique of Judgment*, Kant analyzes empty wishes and longings in terms of the "determination" of the capacity of desire. An agent has an empty wish, he suggests, just in case, through her representation of an object, she determines her forces to realize this object, even though it is impossible for her to realize it. He goes on to say: "It is indeed a not unimportant problem for anthropology to investigate why it is that nature has given us the predisposition to such fruitless expenditure of our forces as [we see in] empty wishes and longings (which certainly play a large role in human life). It seems to me that here, as in all else, nature has made wise provisions. For if we had to assure ourselves that we can in fact produce the object, before the representation of it could determine us to apply our forces, our forces would presumably remain largely unused" (KUE 231, note). Kant seems to suggest here that the determination of an agent's capacity of desire is equivalent to her actually setting herself to try to realize an object.
40. At GMS 400, Kant says that the will "must be determined by the formal principle of volition as such when an action is done from duty, where every material principle has been withdrawn from it."
41. Typically, Kant defines happiness (roughly) as the complete satisfaction of all inclinations over a lifetime. See, for example, KrV A 806/B 834; GMS 399, 405; KpV 73, 124; Rel 58 (English ed. 51). Precisely how his hedonistic and inclination-based definitions of happiness are supposed to cohere with one another is a complex issue. For one attempt to resolve it, see Virginia Wike, *Kant on Happiness in Ethics* (Albany: State University of New York Press, 1994), chap. 1.
42. Thanks to Michael Slote for this example.

Chapter 2: Transcendental Freedom and the Derivation of the Formula of Universal Law

1. For example, in section 2 he aims to show that no "material" practical principle could be the supreme principle of morality (KpV 21–22).

2. Henry E. Allison, *Idealism and Freedom* (Cambridge: Cambridge University Press, 1996), xx.
3. Henry E. Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990), 204.
4. Allison, *Idealism*, 204.
5. Ibid., xviii.
6. Ibid.
7. Ibid., 131.
8. Allison acknowledges this point. See ibid., 130.
9. See, for example, ibid., 130–131. At *Idealism*, 130, Allison says that "the Incorporation Thesis is best seen as a general thesis about how motives function in the case of finite rational agents." He goes on (130–131) to say that "although a finite rational agent is sensuously or "pathologically" affected, that is to say, it finds itself with a set of given inclinations and desires, which provide possible motives or reasons to act, it is not causally necessitated to act on the basis of any of them."
10. See, for example, ibid., 131.
11. Ibid., 132.
12. See Allison, *Kant's Theory of Freedom*, 207, and *Idealism*, 151–152.
13. Allison, *Idealism*, 152. I am simply assuming here that Allison's account of transcendental freedom is on target. One might argue that, at least in the *Critique of Pure Reason*, transcendental freedom amounts to bare independence from natural causes.
14. Allison credits Karl Ameriks and Paul Guyer with raising this sort of question. See *Idealism*, 124.
15. Ibid., 123–128.
16. Ibid., 138.
17. Ibid., 152; see also *Kant's Theory of Freedom*, 207–208.
18. See Thomas Nagel, "Universality and the Reflective Self," in Christine M. Korsgaard, *The Sources of Normativity*, ed. Onora O'Neill (Cambridge: Cambridge University Press, 1996), 201–203.
19. Allison, *Idealism*, 152, and *Kant's Theory of Freedom*, 208.
20. Allison, *Kant's Theory of Freedom*, 208.
21. Ibid., 209. See also *Idealism*, 152.
22. Allison, *Kant's Theory of Freedom*, 210.
23. That Allison agrees that each maxim contains, if only implicitly, a description of an end becomes evident at *Idealism*, 119, and *Kant's Theory of Freedom*, 90–91.
24. Allison, *Kant's Theory of Freedom*, 210.
25. Reath challenges Allison's argument on the grounds that it does not eliminate the possibility that the "Principle of Happiness," rather than the moral law, could be legitimately adopted by a transcendentally free rational agent as his fundamental principle. See Andrews Reath, "Intelligible Character and the Reciprocity Thesis," *Inquiry* 36 (1993): 427. Reath's challenge is aimed at step 4 rather than step 3 of Allison's argument.
26. This sort of response is suggested by Allison's reply to the challenge of Andrews Reath (mentioned in note 25). See Allison, *Idealism*, 117.
27. For an example of Kant's making this move, see GMS 416: "[O]nly law brings with it the concept of an unconditional and objective *and hence universally valid*

necessity" (emphasis added; Kant's emphasis omitted). Kant, of course, makes it abundantly clear that universally valid means binding on all rational beings (see, e.g., KpV 32).

28. Allison, *Kant's Theory of Freedom*, 212, and *Idealism*, 152.
29. Kant acknowledges that sometimes "reason employs the unity of the maxims in general, a unity which is inherent in the moral law, merely to bestow upon the incentives of inclination, under the name of *happiness*, a unity of maxims which otherwise they cannot have. (For example, truthfulness, if adopted as a basic principle, delivers us from the anxiety of making our lies agree with one another and of not being entangled by their serpent coils)" (Rel 36–37, English ed. 32).
30. Allison, *Kant's Theory of Freedom*, 213.
31. Allison, *Idealism*, 152–153. See also his *Kant's Theory of Freedom*, 213.

Chapter 3: The Derivation of the Formula of Humanity

1. Here I am following Hill. See Thomas E. Hill Jr., *Dignity and Practical Reason in Kant's Moral Theory* (Ithaca: Cornell University Press, 1992), 38–41. For a slightly different account of what Kant means by "humanity," see Allen W. Wood, "Humanity as End in Itself," in *Proceedings of the Eighth International Kant Congress*, vol. 1, ed. Hoke Robinson and Gordon Brittan (Milwaukee: Marquette University Press, 1995), 306.
2. On my reading, it simply belongs to Kant's concept of an agent's having a particular practical principle that he have a sufficient ground (motive) to act on it. See section 1.8.
3. He makes the claim at KpV 60–61. By the time he makes it, Kant is obviously operating not only with the view that he has offered a successful derivation of the Formula of Universal Law (see KpV 41) but also that the Formula of Universal Law is valid. At KpV 31, he states that consciousness of the moral law, that is, the Formula of Universal Law, may be called a "fact of reason."
4. Amartya Sen, "Evaluator Relativity and Consequential Evaluation," *Philosophy and Public Affairs* 12 (1983): 114.
5. One might be concerned that this distinction Kant makes between natural laws and practical laws threatens his thesis of the unity of reason. According to this thesis, practical and speculative reason are unified in a common principle. "[T]here can, in the end," claims Kant, "be only one and the same reason, which must be distinguished merely in its application" (GMS 391). Now Kant *seems* to assert that speculative reason is constitutive of natural laws, laws of what is, and practical reason is constitutive of practical laws, that is, laws of what ought to be. Yet how could this be if practical and speculative reason were unified in a common principle? Recently, Neiman has claimed that for Kant reason, whether speculative or practical, has the role of providing laws "that tell us what ought to happen, even if it never does, not laws of nature, which tell us what does happen" (Susan Neiman, *The Unity of Reason* [Oxford: Oxford University Press, 1994], 108). Neiman suggests that on Kant's considered view speculative reason is not constitutive of natural laws but, when employed properly, is regulative. Speculative reason urges us toward the end of complete and systematic knowledge of the realm of nature. See, for example, 125–129. If, as Neiman asserts, for Kant both speculative and practical reason are regulative

in the sense of providing ends and standards for activity, then Kant's unity of reason thesis seems less problematic than it would if he held speculative reason to be constitutive of natural laws.

6. Of course, Kant might, in the end, simply claim that it belongs to his concept of a practical law that it contain "the *very same determining ground* of the will in all cases and for all rational beings." If we embrace this concept, it does follow that, if there is a practical law (categorical imperative), then each rational agent must hold that some object is (or objects are) unconditionally good. But then the question arises: why should we embrace this concept?
7. See the second full paragraph at GMS 428.
8. See Christine M. Korsgaard, *The Sources of Normativity*, ed. Onora O'Neill (Cambridge: Cambridge University Press, 1996), 122.
9. Parts of sections 3.4–7 stem from Samuel J. Kerstein, "Korsgaard's Kantian Arguments for the Value of Humanity," *Canadian Journal of Philosophy* 31 (March 2001): 23–52.
10. For a brief criticism of the textual accuracy of the argument Korsgaard attributes to Kant, see Berys Gaut, "The Structure of Practical Reason," in *Ethics and Practical Reason*, ed. Garrett Cullity and Berys Gaut (Oxford: Oxford University Press, 1997), 173–174.
11. Korsgaard says: "In the argument for the Formula of Humanity, as I understand it, Kant uses the premise that when we act we take ourselves to be acting reasonably and so we suppose that our end is, in his sense, objectively good" (Christine M. Korsgaard, "Kant's Formula of Humanity," in *Creating the Kingdom of Ends* [Cambridge: Cambridge University Press, 1996], 116). Later Korsgaard asks: "Suppose that you make a choice, and you believe what you have opted for is a good thing. How can you justify it or account for its goodness?" (ibid., 121). Korsgaard appears to use the terms "good end" and "objectively good end" interchangeably. I employ the simpler term "good end."
12. Ibid., 115.
13. Ibid., 114.
14. Ibid., 115; see also 120, 122.
15. Korsgaard says: "If one's end cannot be shared, and so cannot be an object of the faculty of desire for everyone, it cannot be good" (ibid., 116).
16. Ibid., 122.
17. Christine M. Korsgaard, "Aristotle and Kant on the Source of Value," in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 227. See also Korsgaard, "Two Distinctions in Goodness," in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 259.
18. Korsgaard, "Humanity," 114–119.
19. See Korsgaard, *The Sources of Normativity*, 122. Korsgaard there suggests that the argument she attributes to Kant in "Kant's Formula of Humanity," namely what I call the "regressive argument," supports the following conclusion. Unless an agent takes humanity to be unconditionally valuable, he must embrace complete practical skepticism. If the regressive argument is effective, it demonstrates that, if an agent maintains that some of his ends are good, then he must hold rational nature to be unconditionally good. In other words, an agent must either hold humanity to be unconditionally good or give up the notion that he has good ends in Korsgaard's very robust sense. Korsgaard appears to believe that an agent's abandoning the notion that he has good ends in her sense would force him into

complete practical skepticism. If an agent gives up this notion, then, rationally speaking, he finds himself without reasons for his actions. Unless Korsgaard holds this, it is totally unclear how the regressive argument is supposed to support her conclusion. She needs to block the possibility that it would be rational to deny that one has good ends (and thereby deny the necessity of holding humanity to be unconditionally good), yet to affirm that one has reasons for his actions.

20. This example stems from Gaut, "The Structure of Practical Reason," 174.
21. In a context similar to that of the present discussion, Allen Wood recognizes that one might raise an objection to the sort of argument Korsgaard (and presumably Kant) want to make here. The objection is that "if *y* is something valuable and *x* is its source, it does not in general follow that *x* is something valuable, still less that it is objectively valuable or an end in itself" (Allen W. Wood, *Kant's Ethical Thought* [Cambridge: Cambridge University Press, 1999], 130). Wood responds to this objection by claiming that it misunderstands the argument. In the argument, says Wood, "rational nature is not being viewed as the source of *good things* (i.e., of their *existence*), but instead as the source of the fact of their *goodness*" (ibid., 130). I do not believe that Wood's response here threatens my counterexample. To use Wood's terms, in my example rational disapproval is being viewed as the source of the fact of the badness of bad things. And that rational disapproval is being viewed as the source of this fact does not entail that rational disapproval must itself be viewed to be bad at all, let alone unconditionally bad. So the question remains: why must the source of the fact of good things' goodness be viewed to be good at all, let alone unconditionally good?
22. In *The Sources of Normativity*, 122, Korsgaard offers a "fancy new model" of the regressive argument. For criticism of this new version of the argument – a version that appeals to the notion of a "practical identity" – see Kerstein, "Korsgaard's Kantian Arguments for the Value of Humanity," 42–51.
23. Korsgaard, "Humanity," 117.
24. Ibid., 110.
25. See ibid., 111, and Christine M. Korsgaard, "Motivation, Metaphysics, and the Value of the Self," *Ethics* 109 (1998): 55.
26. Korsgaard, "Humanity," 110–111.
27. See ibid., 111, 113.
28. For evidence that Korsgaard is using "humanity" as a synonym for rational nature, see ibid., 110–114. For evidence that she equates rational nature with the power of rational choice, see ibid., 123. At 123 Korsgaard says that "humanity is the power of rational choice."
29. See ibid., 119–124. For a summary of the regressive argument, see Korsgaard, "Two Distinctions in Goodness," 256–262. See also Korsgaard, "Aristotle and Kant on the Source of Value," 239–243.
30. For evidence that this is the notion of a sufficient condition that Korsgaard has in mind, see Korsgaard, "Humanity," 122.
31. Ibid., 121.
32. See GMS 393–395. For a recent criticism of Kant's argument, especially as a response to the kind of value realism I have discussed, see Gaut, "The Structure of Practical Reason," 165–170.

33. Korsgaard, "Humanity," 121.
34. Ibid.
35. Ibid., 122.
36. Ibid.
37. Ibid.
38. Ibid.
39. Ibid., 123.
40. Ibid.
41. Ibid.
42. For this conception of happiness in Kant, see, for example, GMS 399, 405.
43. It is not logically impossible for everyone to be happy, even on Kant's desire-satisfaction account of happiness. For we can coherently imagine a world in which every person always gets what he wants. Of course, in this imagined world, no one person's satisfying any of his desires would preclude any other person from satisfying any of her desires.
44. Kant says that "an impartial rational spectator can take no delight in seeing the uninterrupted prosperity of a being graced with no feature of a pure and good will" (GMS 393).
45. This reading of "good will" would have to be broadened to accommodate Kant's view that perfectly rational beings such as God cannot act from duty. To them the "ought" of duty does not apply, since their willing is necessarily in accord with the law. See GMS 414. We might attribute to Kant the view that these beings have a good will (engage in unconditionally good willing) just in case they act "for the sake of the law." Presumably such beings are capable of doing this. And Kant does not seem averse to the idea that acting from duty is a species of acting for the sake of the law.
46. See Karl Ameriks, "Kant on the Good Will," in *Grundlegung zur Metaphysik der Sitten; Ein kooperativer Kommentar*, ed. Otfried Höffe (Frankfurt am Main: Klostermann, 1989), 54–59.
47. See Christine M. Korsgaard, "Kant's Analysis of Obligation: The Argument of *Groundwork* I," in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 60–61.
48. I am here following Ameriks, "Kant on the Good Will," 53.
49. Ibid., 51–54.
50. In focusing on the possibility of a value realist posing this sort of question, I am following Gaut. See Berys Gaut, "The Structure of Practical Reason," 176. I do not wish to suggest that Gaut defends the environmentalist position I have mentioned. He does support a version of value realism but not environmentalism. In *The Sources of Normativity*, Christine Korsgaard offers an argument against value realism. I critically discuss this argument in "Korsgaard's Kantian Arguments for the Value of Humanity." 44–51.
51. See Gaut, "The Structure of Practical Reason," 167. My criticism of Kant's claim that nothing but the good will is good without qualification is indebted to Gaut's treatment. See also H. J. Paton, *The Categorical Imperative* (New York: Harper & Row, 1967), 38.
52. For a forceful challenge to Kant's view that courage, cleverness, and knowledge are not unconditionally good, see Gaut, "The Structure of Practical Reason," 167–168.

53. This point derives from David Cummiskey, *Kantian Consequentialism* (New York: Oxford University Press, 1996), chap. 5.

Chapter 4: The Derivation of the Formula of Universal Law: A Criterial Reading

1. Parts of this chapter have been adapted from Berys Gaut and Samuel Kerstein, "The Derivation without the Gap: Rethinking *Groundwork* I," *Kantian Review* 3 (1999): 18–40.
2. Korsgaard's reconstruction can be found in Christine M. Korsgaard, "Kant's Analysis of Obligation: The Argument of *Groundwork* I," in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 43–76. For evidence that in Korsgaard's view the derivation succeeds, see her last paragraph on page 67.
3. The numbering of the steps here is not Korsgaard's.
4. Korsgaard, "Kant's Analysis," 60 (emphasis omitted).
5. Ibid., 61.
6. Ibid.
7. Ibid., 63 (emphasis omitted).
8. Ibid. (emphasis omitted).
9. Ibid., 61–62.
10. Korsgaard in her interpretation appears to be trying to exploit some of the considerations on which an account of autonomy might draw to make the derivation work. But autonomy is not mentioned once in the derivation in *Groundwork* I. Its deployment belongs to the second section of the *Groundwork*.
11. Korsgaard, "Kant's Analysis," 62.
12. There is an additional ground I have for rejecting the notion that the derivation as interpreted by Korsgaard succeeds. Korsgaard endorses the view expressed in step i, namely that the reason why a good-willed person does an action and the reason why the action is right are the same. Here the assumption is that if an action expresses good will, for example, if it is done from duty, then it is right. But I will argue in Chapter 6 that, actually, Kant is rationally compelled to acknowledge that an action can express good will even if it is not right.
13. Kant's main task in *Groundwork* II also seems to be to derive the supreme principle of morality – in all the complexity of its various formulas. Of course, in neither of the first two sections of the *Groundwork* does Kant claim to show that there *is* a supreme principle of morality. For as he explicitly acknowledges he has not there eliminated the possibility that our view that we are bound by moral requirements is a "phantom of the brain" (see section I.3).
14. For a similar point, see Marcia W. Baron, *Kantian Ethics Almost without Apology* (Ithaca: Cornell University Press, 1995), 181–182.
15. In the roughly two pages that remain in *Groundwork* I, Kant tries to show the need to engage in a more rigorously philosophical discussion of the supreme principle of morality. This discussion takes place in *Groundwork* II.
16. Strictly speaking, the principle is a preliminary version of the Formula of Universal Law. Kant sets out the canonical version of this formula later, in *Groundwork* II (GMS 421). The preliminary version is this: "*I ought never to act except in such a way that I could also will that my maxim should become a universal law*" (GMS 402).

17. This reconstruction is based on Bruce Aune, *Kant's Theory of Morals* (Princeton: Princeton University Press, 1979), 32–33.
18. At GMS 397, Kant says that he is going to explicate the concept of a good will through exploring the concept of duty, "which contains that of a good will though under certain subjective limitations and hindrances." It is clear that these subjective limitations and hindrances are inclinations the fulfillment of which would involve acting contrary to what morality requires.
19. For Kant's suggestion that moral worth is unconditional worth, see GMS 400. Kant refers to the "unqualified worth" of actions at GMS 411.
20. See, for example, Barbara Herman, *The Practice of Moral Judgment* (Cambridge, Mass.: Harvard University Press, 1993), 1; Henry E. Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990), 107, and Baron, *Kantian Ethics Almost without Apology*, 28, n. 19. That Kant held that only actions from duty have moral worth is strongly suggested in his discussion of cases from GMS 397–399. In the second *Critique*, Kant confirms that this is his view: "[M]oral worth must be placed solely in this: that the action takes place from duty, that is, for the sake of the law alone" (KpV 81). As Thomas E. Hill Jr. has pointed out to me, Kant does not in the *Groundwork* explicitly state that all actions from duty have moral worth. However, I think Kant strongly implies that he holds this near the end of GMS 399: "The second proposition is this: *an action from duty has its moral worth* not in the purpose to be attained by it but in the maxim in accordance with which it is decided upon" (emphasis added, Kant's emphasis omitted).
21. At GMS 400, Kant suggests that in an action done from duty, the will is determined by the law.
22. For Kant's usage of the term "impulse," see GMS 434, 444.
23. See also the end of GMS 427 and the beginning of GMS 428.
24. Yet, as we noted in section 1.8, this is not the end of the story. Under Theorem II, Kant states that all material principles place the determining ground of the will in the pleasure or displeasure to be received from an object. On Kant's conception, an agent has sufficient motive to conform to a material practical principle only if he expects that doing so will result in the realization of some object he desires (e.g., his visiting Grant's tomb) *and* he expects that realizing this object will have a hedonic payoff. Whenever an agent acts on a material practical principle, that is, follows the principle's prescription for trying to realize an object, she has some hedonic motivation.
25. For what might be a different way Kant has of distinguishing between material and formal principles, see the end of GMS 427 and the beginning of GMS 428. Allison offers a helpful discussion of some difficulties in Kant's use of the term "formal." See Henry E. Allison, *Idealism and Freedom* (Cambridge: Cambridge University Press, 1996), 150.
26. See Henry E. Allison, *Kant's Transcendental Idealism* (New Haven: Yale University Press, 1983), 78.
27. With regard to theoretical rules, Kant says the following: "[E]xperience never confers on its judgments true or strict, but only assumed and comparative *universality*, through induction. We can properly only say, therefore, that, so far as we have hitherto observed, there is no exception to this or that rule. If, then, a judgment is thought with strict universality, that is, in such a manner that

no exception is allowed as possible, it is not derived from experience, but is valid absolutely *a priori*. Empirical universality is only an arbitrary extension of a validity holding in most cases to one which holds in all" (KrV B 3–4). Just as empirically based universality is insufficient to ground theoretical rules that allow of no possible exception, so it is insufficient to ground practical rules that allow of none. See KpV 26.

28. Kant claims to prove the validity of the principle, "[A]ct on no other maxim than that which can also have as object itself as a universal law" (GMS 447).

Chapter 5: Criteria for the Supreme Principle of Morality

1. Chapter 6 explores objections to this first criterion.
2. This section has been adapted from my paper "The Kantian Moral Worth of Actions Contrary to Duty," *Zeitschrift für Philosophische Forschung* 53 (1999): 45–66.
3. This interpretation is not wholly unproblematic. In the Preface to the *Groundwork*, Kant remarks that moral laws require "a power of judgment sharpened by experience, partly in order to distinguish in which cases they are applicable and partly to provide them with access to the will of the human being and efficacy for his fulfillment of them" (GMS 389). The question is: according to Kant, does everyone, that is, every rational agent, acquire this power of judgment in the course of maturing? Of course, if Kant would answer affirmatively, then the interpretation at hand is not really threatened by this remark.
4. Allison, for example, says: "Starting with the assumption, itself questionable, that actions performed from duty cannot, objectively speaking, be contrary to duty, he proposes to limit his consideration to actions that are at least in agreement with duty (*pflichtmäßig*)." See Henry E. Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990), 109.
5. Curzer adopts this interpretation. See Howard Curzer, "From Duty, Moral Worth, Good Will," *Dialogue* 36 (1997): 290–291.
6. That Kant would hold this is not as obvious as it might at first appear. Kant asserts that duties cannot conflict: if at the same time two moral rules prescribe different actions, then it cannot be a duty to act in accordance with both. See MS 224. But suppose that disagreeing with Kant on this score, an agent holds that duties can conflict. Further suppose that here and now she believes that she has conflicting duties, a duty to keep a promise to meet a student and a duty to take her mother-in-law to the airport. It seems that in performing the latter action, she could both believe that she was acting contrary to duty (i.e., her duty not to break her promises), yet be acting from duty. Moreover, Kant might acknowledge this possibility, although he would insist that the agent was mistaken in her belief that duties can conflict. I think it is clear that *if* an agent believes along with Kant that duties cannot conflict, then she could not, in doing something she believed to be contrary to duty, be acting from duty.
7. It is striking that in a work where Kant is at pains to explore the concept of duty, he mentions not one example of an action done from duty but which conflicts with duty.
8. At KpV 81, Kant says that moral worth "must be placed solely in this: that the action takes place from duty, that is, for the sake of the law alone."

9. At KpV 79, Kant says: "A maxim is . . . morally genuine only if it rests solely on the interest one takes in compliance with the law."
10. Here one might suggest the following. Supererogatory actions are not morally required; they are optional in the sense of beyond what duty requires. Nevertheless, in performing one (e.g., by jumping on a live grenade to save one's comrades), one might put aside entirely the influence of inclination. So there are some actions that are neither morally required nor at all influenced by inclination. In response, note first that Kant does not recognize the category of supererogatory actions. For discussion of why, see Marcia W. Baron, *Kantian Ethics Almost without Apology* (Ithaca: Cornell University Press, 1995), chap. 1. Moreover, even if Kant did recognize this category, a supererogatory action would not be done from duty in his sense. In an action from duty, an agent's will is determined by her representation of the law in itself (GMS 401). But in a supererogatory action, her will is presumably not determined simply by her representation of the law, since she is aiming "above and beyond" what the law requires.
11. See GMS 402 for evidence that Kant recognizes this point.
12. For discussion of Kant's notion of respect, see Allison, *Kant's Theory of Freedom*, 120–128, and Ralph C. S. Walker, "Achtung in the *Grundlegung*," in *Grundlegung zur Metaphysic der Sitten: Ein kooperativer Kommentar*, ed. Otfried Höffe (Frankfurt am Main: Klostermann, 1989), 97–116.
13. See Barbara Herman, *The Practice of Moral Judgment* (Cambridge, Mass.: Harvard University Press, 1993), 13–17. See also Baron, *Kantian Ethics Almost without Apology*, 129–130.
14. See Herman, *Practice*, 16.
15. It is striking that Herman does not cite a single case of Kant's mentioning that an agent performs a particular action from duty as a "limiting condition" or "secondary motive."
16. This point is made by H. J. Paton, *The Categorical Imperative* (New York: Harper & Row, 1967), 47–50; Herman, *Practice*, 12; Allison, *Kant's Theory of Freedom*, 110–111; and Baron, *Kantian Ethics*, 150. Both Allison and Baron note that Kant implies early in his *Groundwork* I discussion of duty that an agent's having an inclination to do something does not in itself preclude him from doing it from duty. Kant suggests that to determine whether an action is from duty is difficult in cases "when an action conforms with duty and the subject has, besides, an *immediate* inclination to it" (GMS 397). But why, Allison and Baron rightly ask, would he suggest this to be difficult if he adopted the view that having an immediate inclination to do something itself precluded one from doing it from duty? If he adopted this view, then, it seems he would simply say that if one has an immediate inclination to an action, then it is thereby impossible for him to do it from duty. (I discuss Kant's distinction between mediate and immediate inclination in section 5.5.) For an opposing interpretation of Kant, see Noa Latham, "Causally Irrelevant Reasons and Action Solely from the Motive of Duty," *Journal of Philosophy* 91 (1994): 599–618.
17. Of course, Kant does hold that it is impossible for me to be certain that I have kept the promise from this motive. See GMS 407.
18. Allison puts the point in this way. See Allison, *Kant's Theory of Freedom*, 111.
19. In addition to Baron, *Kantian Ethics*, 146–187, see, for example, Paul Benson, "Moral Worth," *Philosophical Studies* 51 (1987): 365–382; Paul Guyer, *Kant on*

Freedom, Law, and Happiness (Cambridge: Cambridge University Press, 2000), 287–303; Richard G. Henson, "What Kant Might Have Said: Moral Worth and the Overdetermination of Dutiful Action," *Philosophical Review* 88 (1979): 39–54; Tom Sorell, "Kant's Good Will and Our Good Nature," *Kant-Studien* 78 (1987): 87–101.

20. This paragraph has been influenced by Henry Allison. See Allison, *Kant's Theory of Freedom*, 116–118.
21. Typically, Kant defines happiness (roughly) as the complete satisfaction of all inclinations over a lifetime – thus seemingly not in *purely* hedonistic terms. See, for example, KrV A 806/B 834; GMS 399, 405; KpV 73, 124; Rel 58 (English ed. 51).
22. Here I am following Marcia Baron's initial account of overdetermined actions. See *Kantian Ethics*, 150. (She offers a more detailed account on 156–157). In concluding that for Kant there can be no overdetermined actions done from duty and from inclination, I am also following Baron. She writes: "Overdetermined actions involving duty and inclination do not merely lack moral worth; they are not intelligible" (161). However, the grounds Baron offers for this conclusion differ significantly from the ones I have given (see *Kantian Ethics*, 156–161).
23. Kant implies this, for example, in his "second proposition": "an action from duty *has its moral worth* not in the purpose to be attained by it, but in the maxim in accordance with which it is decided upon" (GMS 399; emphasis added, Kant's emphasis omitted).
24. I have been influenced here by Guyer's discussion of this Kantian notion. See Paul Guyer, *Kant and the Experience of Freedom* (Cambridge: Cambridge University Press, 1993), 344–351. See also Guyer, *Kant on Freedom, Law, and Happiness*, 107–117.
25. See also Rel 30–31 (English ed. 26): "For when incentives other than the law itself (such as ambition, self-love in general, yes, even a kindly instinct such as sympathy) are necessary to determine the will [*Willkür*] to conduct *conformable to the law*, it is merely accidental that these causes coincide with the law, for they could equally well incite its violation."
26. This example is based on one offered by Barbara Herman. See Herman, *Practice*, 4–5.
27. Here I am following ibid., 5–6.
28. See ibid., 5.
29. I offer some criticism of this argument in section 8.10.
30. Of course, Kant expresses this view not only in the *Groundwork*, but in the second *Critique* as well. See, for example, KpV 20.
31. Kant seems to think that we must interpret this principle as a material one. But I do not see why we must. See section 7.2.
32. Ultimately, I do not believe that Kant should rely on this argument. For reasons I crystallize in section 8.2, I think Kant should abandon one of the argument's key premises, namely that ought implies can.

Chapter 6: Duty and Moral Worth

1. I do not wish to imply that Williams's and Stocker's criticisms are meant to apply to Kantian morality exclusively. They both aim their criticisms at utilitarian theories as well, just to name one additional target.

2. Both my understanding of this objection and my response to it have been heavily influenced by Baron's work. See Marcia W. Baron, *Kantian Ethics Almost without Apology* (Ithaca: Cornell University Press, 1995), 118–135.
3. Michael Stocker, "The Schizophrenia of Modern Ethical Theories," *Journal of Philosophy* 73 (1976): 462.
4. At MS 456, Kant states that we, human beings, have a duty to use our capacity to have sympathetic feelings as a means to promote "active and rational benevolence." Perhaps the person Kant describes in the *Groundwork* is lacking in this capacity. Kant suggests that "nature had put little sympathy" in his heart (GMS 398).
5. Oakley appears to defend this position. See Justin Oakley, *Morality and the Emotions* (New York: Routledge, 1992), 57–63, 83.
6. See Bernard Williams, *Moral Luck: Philosophical Papers, 1973–1980*, (Cambridge: Cambridge University Press, 1981), 18. Williams discusses a case put forth by Charles Fried. The case I discuss is inspired by, but differs from, the one Williams discusses.
7. I have been influenced in my thinking regarding cases of this kind by Baron, *Kantian Ethics*, 136–140; Paul Guyer, *Kant and the Experience of Freedom* (Cambridge: Cambridge University Press, 1993), 392–393; and Barbara Herman, *The Practice of Moral Judgment* (Cambridge, Mass.: Harvard University Press, 1993), 41–42.
8. My treatment of this thesis (sections 6.5–9) mirrors my discussion in Samuel Kerstein, "The Kantian Moral Worth of Actions Contrary to Duty," *Zeitschrift für Philosophische Forschung* 53 (1999): 45–66.
9. Some commentators ignore (what I take to be) Kant's *Groundwork* denial of moral worth to all morally impermissible actions. See, for example, H. J. Paton, *The Categorical Imperative*, 5th ed. (New York: Harper & Row, 1965), 46–50. Others take note of this denial but do not explore in any detail Kant's grounds for it or its plausibility. See, for example, Henry E. Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990), 109. Roger Sullivan offers some helpful references to remarks Kant makes relating to the sources of error in moral judgment. See Roger J. Sullivan, *Immanuel Kant's Moral Theory* (Cambridge: Cambridge University Press, 1989), 57–60. However, Sullivan does not directly address the issue I discuss here, namely that of whether the logic of Kant's own moral theory rationally compels him to embrace the possibility that morally impermissible actions can have moral worth.
10. Kant sums up what he takes to have accomplished in *Groundwork* I when he says: "Thus, then, we have arrived, within the *moral cognition of common human reason*, at its principle" (GMS 403, emphasis added).
11. By writing of an action's (potentially) *expressing* good will rather than being identical to it, I am implicitly invoking the "whole character" understanding of a good will. The point at issue would not be altered if I instead invoked the "particular action" understanding of a good will. For the distinction between these two notions of a good will, see section 3.7.
12. Thomas Nagel focuses on this aspect of Kant's doctrine in the *Groundwork*. See Thomas Nagel, "Moral Luck," in *Mortal Questions* (Cambridge: Cambridge University Press, 1979), 24–25. See also Thomas E. Hill Jr., *Respect, Pluralism, and Justice* (Oxford: Oxford University Press, 2000), 159–160.

13. Recall Kant's dualistic view of the ultimate grounds of human action. Either we act from inclination, or we act from duty (section 1.6). Since he denies that actions from duty can conflict with duty, Kant must hold that actions that conflict with duty are done from inclination.
14. One might wonder whether, in Kant's view, an agent's failure to perform a morally permissible action is *always* a failure of will. For it seems that when a person is drunk, he can perform a morally impermissible action that does not stem from a failure of will – if only because he is so intoxicated that, at that moment, he *has no* will to fail. In response, I think Kant would argue that the drunk person's morally impermissible actions really do stem ultimately from a failure of will, namely a failure to suppress the inclination to drink in the first place. In the *Lectures on Ethics*, Kant says: "everything is imputable that pertains to freedom, even though it may not have arisen directly through freedom, but indirectly nevertheless. E.g., what a person has done in a state of drunkenness may well not be imputed; but he can be held accountable for having got drunk" (LE 291).
15. Strictly speaking, Kant suggests that we use for moral appraisal "the universal formula of the categorical imperative: *act in accordance with a maxim that can at the same time make itself a universal law*" (GMS 436–437). In the example, Colonel Mikavitch takes "the universal formula of the categorical imperative" to be a version of the Formula of Universal Law. Wood has argued recently that it is actually a version of the Formula of Autonomy; see Allen W. Wood, *Kant's Ethical Thought* (Cambridge: Cambridge University Press, 1999), 188–189.
16. Admittedly, Kant sometimes comes out against the permissibility of suicide in any circumstances. For example, in the *Lectures on Ethics* he says: "There are many conditions under which life has to be sacrificed; if I cannot preserve it other than by violating the duties to myself, then I am bound to sacrifice it, rather than violate these duties; yet on the other hand, suicide is not permitted under any condition" (LE 372).
17. An orthodox Act Utilitarian would define the good as (something like) the maximum happiness of all sentient beings. He would then conceive an action's moral value purely in terms of the degree to which it promotes the good thus defined. For the Act Utilitarian, if Stram's action diminishes the general happiness, then it has no moral value – unless, perhaps, all of the actions open to him would diminish the general happiness, but this action is the one that would diminish it least. As an orthodox Act Utilitarian, Stram would himself hold the moral value of his lying to the politician to be contingent on its effects. Of course, we need not agree with Stram's own take on when his actions have moral worth. For a far more detailed discussion of utilitarianism and moral worth, see sections 7.3–5.
18. Actually, there seems to be a third kind of case of an action's being morally impermissible, yet having moral worth. In this kind of case, an agent does his best to adopt the correct moral principle but fails. He also does his best to apply properly the principle he has adopted but fails at that as well. In this kind of case, there is a failure both of principle choice and of principle application.
19. For an interesting discussion of the Kantian conception of conscience, as well as two other conceptions of it, see Hill, *Respect, Pluralism, and Justice*, 260–274.

20. In the *Anthropology from a Pragmatic Point of View*, one might argue, Kant accepts that a person who adopted a standard of moral judgment other than the Categorical Imperative might perform morally impermissible actions, which nonetheless had moral worth. Consider his discussion of "character" (*Charakter*). In much the same language he uses in *Groundwork* I to describe a good will, he tells us that character has intrinsic worth (*inneren Wert*) and is beyond all price. A few sentences earlier, he says: "to have character relates to that property of the will by which the subject has bound himself to certain practical principles which he has unalterably prescribed for himself by his own reason. Although these principles may sometimes indeed be false or defective, nevertheless the formal element of the will as such, which is determined to act according to firm principles (not shifting hither and yon like a swarm of gnats), has something precious and admirable to it, which is also something rare" (Anth 292). Here, one might argue, Kant is advocating the idea that, even if a person has adopted false or defective practical principles, she can have character. Since Kant believes character to have intrinsic worth, he also holds that *actions expressing character* have a special worth. Therefore, concludes the argument, Kant thinks that a person could perform a morally impermissible action that had moral worth: an action of obeying a self-given, yet false, principle because she believed obeying it was the right thing to do. In response, note that Kant does not really embrace the idea that a person who has adopted false practical principles can have character. Shortly after the cited passage, Kant says: "Character requires maxims, which proceed from reason and from ethicopractical principles" (Anth 293). Kant then lists principles that, he suggests, a person of character would have to live by, including those of not speaking an untruth intentionally, not dissembling, and not breaking one's (legitimate) promise (Anth 294). These are, of course, just the sort of principles that Kant believes to stem from the Categorical Imperative. At the very least, *if* Kant believed someone living by false principles could have character, among her false principles could be none that prevented her from also embracing those on Kant's list. In addition, Kant suggests that a person of character would act on principles that are valid (*gelten*) for everybody (Anth 293). Kant would hardly claim that false principles would be valid for everybody! Despite initial appearances, Kant is not suggesting in the *Anthropology* that a person who lives by false principles can have character, nor, by extension, that the person's action on such principles can have moral worth. He is praising the quality of sticking to one's principles much in the same way that, in *Groundwork* I, he acknowledges the value of self-control and calm deliberation. These qualities, he says, are rightly held in high esteem, but they have no absolute (i.e., moral) worth (see GMS 394).

21. Some of this disagreement is manifest in our discussion of the Formula of Universal Law in Chapter 8. At any rate, Korsgaard discusses three ways of interpreting the Formula of Universal Law, each one of which is found in the literature: Christine M. Korsgaard, "Kant's Formula of Universal Law," in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 77–105. For another way of reading this formulation, see Thomas W. Pogge, "The Categorical Imperative," in *Kant's "Groundwork of the Metaphysics of Morals": Critical Essays*, ed. Paul Guyer (Lanham, Md.: Rowman & Littlefield, 1998), 189–196.

22. At GMS 411, for example, Kant says: "From what has been said it is clear that all moral concepts have their seat and origin completely a priori in reason, and indeed in the most common human reason just as in reason that is speculative in the highest degree."
23. Kant, of course, holds that it is only by virtue of having its source in reason alone that the Categorical Imperative *could* be valid.
24. Michael Slote has suggested this sort of objection, using the example of a conscientious Nazi prison guard. See Michael Slote, *Goods and Virtues* (Oxford: Oxford University Press, 1983), 63. See also Jonathan Bennett, "The Conscience of Huckleberry Finn," *Philosophy* 49 (1974): 123–143.
25. Hannah Arendt, *Eichmann in Jerusalem* (New York: Penguin, 1994), 135–137.
26. As Thomas Pogge has suggested to me, one wonders how, in Kant's view, an inquisitor can be sure that it is right to *spare* a defendant who, in the inquisitor's view, has violated divine doctrine.
27. Kant himself holds that "it is a basic moral principle, which requires no proof, that one ought to hazard nothing that may be wrong" (Rel 185, English ed. 173). Hazarding nothing that may be wrong involves being sure that what one proposes to do is right. For Kant goes on to say that "concerning the act which *I* propose to perform I must be *sure* that it is not wrong; and this requirement is a postulate of conscience" (Rel 186, English ed. 174). I think it is questionable whether Kant has highlighted here a basic moral principle that requires no proof. Why doesn't it require any proof? Also, one might wonder what relation Kant takes to hold between this principle and the Categorical Imperative. Does the Formula of Universal Law (or perhaps the Formula of Humanity) entail this principle? How, precisely, does it do so?
28. See Lawrence Blum, *Friendship, Altruism and Morality* (London: Routledge and Kegan Paul, 1980), 12–15. For an account of emotions in general that is in the spirit of Blum's account of sympathy, see Oakley, *Morality and the Emotions*, 7–16.
29. This discussion has been influenced by Barbara Herman's distinction between the motive of an action and its object. See Herman, *Practice*, 25.
30. Oakley, *Morality and the Emotions*, 83–84.
31. See Michael Slote, *Morals from Motives* (New York: Oxford University Press, 2001), 51–58.
32. See Oakley, *Morality and the Emotions*, 101–102.
33. A defender of the notion that all actions from sympathy have moral value might argue that a soldier who displayed such a lack of sympathy toward ethnic minorities would, contrary to what has been suggested here, be incapable of acting from sympathy at all, even toward his fellow soldier. The idea would be that an utter lack of sympathy toward one group is incompatible with genuine sympathy toward another group. I suppose that this is possible, but I see no good reason to believe it.

Chapter 7: Eliminating Rivals to the Categorical Imperative

1. See also Berys Gaut and Samuel Kerstein, "The Derivation without the Gap: Rethinking *Groundwork* I," *Kantian Review* 3 (1999): 18–40.
2. In the *Groundwork* version of this argument (GMS 444, especially the lower half of the page), Kant does not use the term "material principle," although

he does employ it earlier in the text (GMS 400). As we will see, he instead writes of "heteronomy of the will." Nor does Kant in the *Groundwork* version explicitly invoke the notion that material principles are conditional for their motivational force on the agent's expectation of a hedonic payoff. At GMS 444, however, he does suggest that the motivational force of such principles depends on an "impulse" that the representation of an object exerts on the will.

3. Here I am following David Cummiskey who argues that not all consequentialist principles must be considered to be "material" ones. See David Cummiskey, *Kantian Consequentialism* (New York: Oxford University Press, 1996), 46–48.
4. At KpV 34, Kant writes: "Thus, the happiness of other beings can be the object of the will of a rational being. But if it were the determining ground of the maxim, one would have to presuppose that we find not only a natural satisfaction in the well-being of others but also a need, such as a sympathetic sensibility brings with it in human beings."
5. Amartya Sen, "Utilitarianism and Welfarism," *Journal of Philosophy* 76 (1979): 464.
6. Ibid., 464.
7. This is not an unusual conception of states of affairs. See Sen, "Utilitarianism," 464–465; "Evaluator Relativity and Consequential Evaluation," *Philosophy and Public Affairs* 12 (1983): 128–129; and "Well-Being, Agency, and Freedom: The Dewey Lectures 1984," *Journal of Philosophy* 82 (1985): 181–182. See also Bernard Williams, "A Critique of Utilitarianism," in J. J. C. Smart and B. Williams, *Utilitarianism For and Against* (Cambridge: Cambridge University Press, 1973), 83.
8. Thanks to Thomas Pogge and Michael Slote for pushing me on this point.
9. An agent's conforming to EU amounts to her performing an action that she *expects* will yield as great a sum total of well-being as would any alternative action available to her. Suppose that from duty an agent wills to conform to EU. From duty she wills to do what she expects will yield as great a sum total of well-being as would any alternative action available to her. From the perspective of a Kantian conception of acting from duty, it is hard to see how, in this case, she could fail to do what she expects will yield this result. It seems that she could only fail if she indulged her inclinations. But since she has acted from duty, she has *not* indulged them.
10. Actually, as we will see in section 8.2, arguments such as the one summarized in this paragraph will require a modification of one of Kant's criteria, namely criterion iv.
11. For a concise criticism of Act Utilitarianism on the grounds that it generates a set of duties that clashes with common sense, see Richard B. Brandt, "Toward a Credible Form of Utilitarianism," in *Contemporary Utilitarianism*, ed. M. Bayles (New York: Doubleday, 1968), 146–147.
12. At least for advocates of the Formula of Universal Law, this seems to be a somewhat dangerous argument to make. For, as I contend in Chapter 8, it is very doubtful whether this formula generates a set of duties that squares with ordinary moral thinking.
13. For a defense of this sort of principle, see Thomas Hurka, *Perfectionism* (New York: Oxford University Press, 1993), especially chaps. 2 and 4.
14. This point can also be illustrated with reference to rational perfection. An increase in an agent's rational perfection might not result from an agent's

willing to develop his rational capacities. For example, in an attempt to develop these capacities, the agent might take an experimental "brain-enhancing" drug that ends up diminishing them.

15. For Cummiskey's claim that the "first proposition" is consistent with consequentialism, see *Kantian Consequentialism*, 27; for his claim that the "second proposition" is also consistent with it, see 39.
16. This is nearly a direct quotation from ibid., 99. Cummiskey's detailed statement of his principle spans four paragraphs, 98 to 99.
17. Cummiskey denies that this requirement entails that we ought to maximize the number of rational beings. See ibid., 91.
18. Ibid., 150.
19. Cummiskey suggested this sort of reading of his principle in a paper presented at the annual Pacific Division meeting of the American Philosophical Association, Berkeley, California, March 1997.
20. See Cummiksey, *Kantian Consequentialism*, 4, 79, 156.
21. See ibid., 6, 16.
22. See ibid., 11–12.
23. Cummiskey discusses this argument at length in chapter 4 of his book. See ibid., 62–83.
24. As did we, Cummiskey examines Kant's argument as reconstructed by Korsgaard. See ibid., 81, n. 11.
25. See ibid., 156.
26. TC would have to be rephrased to accommodate Kant's view that a viable candidate for the supreme principle of morality must be capable of being binding on all rational agents, including perfectly rational ones such as angels. Instead of "You ought to honor your father and mother; you ought not to kill; you ought not to commit adultery . . . ," the principle would have to read something like "Honor your father and mother; do not kill; do not commit adultery. . . . " Presumably, angels would, by virtue of their perfect rationality, necessarily act in accordance with whatever principle was the supreme principle of morality. Therefore, with respect to angels, the "ought" in TC would be out of place.

Chapter 8: Conclusion: Kant's Candidates for the Supreme Principle of Morality

1. This is an argument Kant suggests in his discussion of his "second proposition" at GMS 399–400.
2. Recalling an argument we discussed (and criticized) in Chapter 3, one might suspect that Kant himself implies that the principles could not fulfill the criterion. To summarize this argument, if an agent holds there to be a supreme principle of morality, then she must also hold there to be something unconditionally good, claims Kant. For if she did not take there to be something unconditionally good, then she might find herself without sufficient motive to conform to the principle and thus, for reasons that require no repeating here, the principle could not be the supreme principle of morality. In light of this argument, it might appear that Kant (perhaps without being aware of it) commits

himself to the view that the representation of a principle as a law does not provide an agent with sufficient incentive to abide by it. After all, according to the argument, to have sufficient incentive, an agent must (at least in some cases?) take conforming to the principle to promote or secure something unconditionally valuable. In response, note that for Kant it is an agent's representing a principle as universally and unconditionally binding that gives rise to her conception of the good. So ultimately her incentive for acting lies nevertheless in this representation.

3. Kant, of course, actually tests the maxim of false promising using the Formula of the Law of Nature, stated in the third full paragraph at GMS 421.
4. Christine M. Korsgaard, "Kant's Formula of Universal Law," in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 80, 92–93. Baron joins Korsgaard in holding the Practical Contradiction Interpretation to offer the most plausible account of how, precisely, a maxim such as that of false promising fails the Formula of Universal Law test. See Marcia W. Baron, "Kantian Ethics," in Marcia W. Baron, Philip Petit, and Michael Slote, *Three Methods of Ethics* (Oxford: Blackwell, 1997), 69–70.
5. According to Korsgaard, on the Practical Contradiction Interpretation, "the contradiction is that your maxim would be self-defeating if universalized: your action would become ineffectual for the achievement of your purpose if everyone (*tried to*) use it for that purpose"; see Korsgaard, "Kant's Formula of Universal Law," 78 (emphasis added).
6. Here one might object that at an early stage in the imagined world, that is, before everyone has realized that money borrowed simply on a promise will not be repaid, an agent acting on FPM may well attain her end. Perhaps Korsgaard would respond to this objection by saying that, nevertheless, in willing the imagined world, the agent would be willing a world in which her chances of getting money on a false promise were severely diminished. And that would be enough to render practically irrational willing the imagined world at the same time as acting on FPM.
7. As Korsgaard acknowledges, one might also read the Formula of Universal Law to land a person acting on such a maxim of false promising in a logical contradiction. According to (what Korsgaard calls) the Logical Contradiction Interpretation, the universalization of the maxim would be as follows: from self-love, when anyone believes himself to be in need of money, he borrows (*rather than tries to borrow*) money on a promise to repay it, even though he knows that this will never happen. In order to be able to will this world, an agent needs to be able to conceive of it. However, she cannot really conceive of the world, suggests Kant. For not everyone in financial need could get a loan based simply on a promise if no such person ever repaid a loan she received in this way. Creditors would not part with their money. In other words, the practice of lending money to those in need based simply on their promise to repay would cease to exist if none of them ever repaid their loans. There simply *is* no world in which when each and every agent finds herself in financial need, she gets money through false promising. The agent considering the false promising maxim has been forced into a logical contradiction. The Formula of Universal Law requires her to hold that she can conceive of the world of her universalized maxim, since

it requires her to will this world and (as she must acknowledge) she could not will this world without being able to conceive of it. However, she concludes that she cannot conceive of this world. There is obviously a logical contradiction in holding, as the agent would (presumably) have to, that something both is and is not conceivable. Korsgaard grants that this interpretation is well supported by Kant's text. (See Korsgaard, "Kant's Formula of Universal Law," 81–82.) It is worth mentioning that a philosophical difficulty seems to arise in connection with the Logical Contradiction Interpretation. It is not obvious that the world of the universalized maxim is inconceivable. Granted, it is very unlikely that people would continue to lend money simply on a promise that they would be repaid, even though whenever they did lend it, they were not repaid. But this unlikely scenario is (arguably) not inconceivable. Thanks to Thomas Pogge for this point.

8. See Korsgaard, "Kant's Formula of Universal Law," 93, and Barbara Herman, *The Practice of Moral Judgment* (Cambridge, Mass.: Harvard University Press, 1993), 138.
9. This maxim and my analysis of it stem from Herman, *Practice*, 138–139.
10. See Baron, "Kantian Ethics," 73.
11. This maxim stems from Pogge. See Thomas W. Pogge, "The Categorical Imperative," in *Kant's "Groundwork of the Metaphysics of Morals": Critical Essays*, ed. Paul Guyer (Lanham, Md.: Rowman & Littlefield, 1998), 190.
12. Moreover, in this world it is questionable whether *anyone* would be earning a comfortable living (as Jack conceives of one). With everyone trying to become a professor to earn a living, who would do the work necessary to sustain an economy in which it is possible for (many) people to have their own houses, cars, and computers?
13. See Pogge, "The Categorical Imperative," 189–196.
14. Ibid., 190.
15. See ibid., 191.
16. Ibid.
17. Ibid., 192.
18. Ibid.
19. Ibid., 191.
20. As Korsgaard acknowledges, the Practical Contradiction Interpretation also faces difficulties regarding maxims of violence. See Korsgaard, "Kant's Formula of Universal Law," 100. On this interpretation, it is not obvious that acting on the one we have been discussing – that of killing for revenge – turns out to be morally impermissible. And a maxim such as "In order to release my anger, I will punch anyone who offends me" seems to sail through.
21. See Pogge, "The Categorical Imperative," 196.
22. We have already noted Korsgaard's reservations as to whether, on the interpretation she champions, the Formula of Universal Law generates adequate results regarding certain maxims of violence (see Korsgaard, "Kant's Formula of Universal Law," 100). Pogge, it appears, does not think that the Formula of Universal Law itself produces an adequate set of duties; for, in his view, it fails to generate a duty of beneficence (see Pogge "The Categorical Imperative," 196). According to Herman, if we read the Formula of Universal Law as "a method

of judgment to be used by agents in determining the permissibility of their own maxims" (and we have read the formula in this way), then it is not effective. See Herman, *Practice*, 143. Herman (147–157) offers an innovative account of what the Formula of Law procedure might actually accomplish, but I do not discuss this here.

23. Herman, *Practice*, 143.
24. Here I am following Allen W. Wood, "Humanity as End in Itself," in *Proceedings of the Eighth International Kant Congress*, vol. 1, ed. Hoke Robinson and Gordon Brittan (Milwaukee: Marquette University Press, 1995), 317, n. 2.
25. I am assuming here (and I take Kant to hold) that there is no way of treating humanity such that one is treating it neither as a means nor as an end.
26. Here I am following Thomas E. Hill Jr., *Dignity and Practical Reason in Kant's Moral Theory* (Ithaca: Cornell University Press, 1992), 41–42.
27. That for Kant in the Formula of Humanity "end" is equivalent to "end in itself" is clearly implied at GMS 428.
28. See Hill, *Dignity and Practical Reason*, 47–49.
29. See Thomas W. Pogge, "Kant on Ends and the Meaning of Life," in *Reclaiming the History of Ethics*, ed. Andrews Reath, Barbara Herman, and Christine M. Korsgaard (Cambridge: Cambridge University Press, 1997), 361–362.
30. Here "reason" refers to a Kantian motivating reason.
31. In the *Metaphysics of Morals*, Kant seems to derive the duty of beneficence from the Formula of Universal Law. See MS 393, 453.
32. According to Kant, we can safely presuppose that each human agent has this end by a necessity of nature (GMS 415).
33. For discussion of how demanding a duty of beneficence Kant endorses (or is compelled by his own views to endorse), see Marcia W. Baron, *Kantian Ethics Almost without Apology* (Ithaca: Cornell University Press, 1995), chaps. 1–3; Cummiskey, *Kantian Consequentialism*, chap. 6; Hill, *Dignity and Practical Reason*, chap. 8.
34. Allen W. Wood, *Kant's Ethical Thought* (Cambridge: Cambridge University Press, 1999), 153 (emphasis added).
35. There is something very odd about setting oneself to pursue the end of being deceived. It would seem that pursuing the end would likely involve knowing (at least roughly) in what circumstances one is to be deceived (e.g., by a card shark in Las Vegas). But if one knows (even roughly) in which circumstances one is to be deceived, then there is a sense in which one is not entirely deceived.
36. For far more detailed discussion of some of the practical implications of the Formula of Humanity, see Thomas E. Hill Jr. "Respect for Humanity," in *The Tanner Lectures on Human Values*, vol. 18, ed. Grethe B. Peterson (Salt Lake City: University of Utah Press, 1997), 3–76.
37. For an interesting discussion of this issue, see Cummiskey, *Kantian Consequentialism*, chap. 8.
38. Thomas Hill makes this point. See Hill, *Dignity and Practical Reason*, 49, 206.
39. Hill discusses grounds Kant might offer for destroying the humanity in one person in circumstances in which this would preserve the humanity in others. However, it is questionable whether these grounds are limited to ones implicitly endorsed in the Formula of Humanity. To locate the grounds, Hill invokes

Kant's views on justice and the "kingdom of ends." See Hill, *Dignity and Practical Reason*, 207–225.

40. Granted, the agent's view that in order to have sufficient incentive to conform to EU she must rely on the prospect of maximizing aggregate well-being does seem a bit odd. In *some* circumstances, might not she derive sufficient incentive for abiding by EU from another source – for example, from the notion that doing so would promote her own happiness?

Index