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 396). 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). 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). 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. 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). 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—those 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. 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

© Cambridge University Press  www.cambridge.org
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 chimirical 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. 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). 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. 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.
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.” 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.” 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.

### 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:

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. 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.” L, suggests Aune, “is a higher-order principle telling us to conform to certain lower-order laws.” 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. 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.” 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.”

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.\textsuperscript{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.\textsuperscript{22} In their view, in neither \textit{Groundwork I} nor \textit{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 \textit{Groundwork I} and \textit{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.”\textsuperscript{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”:\textsuperscript{24}

\begin{quote}
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.
\end{quote}

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.\textsuperscript{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.\textsuperscript{46} 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