Autonomy as the Supreme Principle of Morality

Jeppe von Platz

> "Morality, which discerns purely *a priori* the laws of freedom, is a metaphysics of freedom, or of morals, just as metaphysics is called a metaphysics of nature."
> (*V-Mo/Mron II*, 29:599)
>
> "Freedom must, if it is to be under laws, give the laws to itself."
> (*V-Naturrecht/Feyerabend*, 27:1322)

1. Introduction

In this essay I try to explain why Kant claims that autonomy of the will is the supreme principle of morality and what he means by it. I argue that the principle of autonomy captures the four definitive aspects of moral willing: its form, content, source, and ideal. The form is universality as specified in the formula of universal law. The content is humanity as specified in the formula of humanity. The source is the agent that is also the subject as specified in the formula of autonomy. And the ideal is the systematic realization of the three previous conditions as expressed by the idea(l)s[1] "of the will of every rational being as a will giving universal law" (*GMS*, 4:431) and of the "whole of all ends in systematic connection." (*GMS*, 4:433) In this manner, Kant's different statements of the categorical imperative identify and express the basic normative requirements of autonomous willing. Conversely, autonomy is the supreme principle of morality because it, and only it, captures all of these four aspects.

[1] As Kant defines it, an idea is "a concept of reason whose object cannot be met with in experience" (*Log*, 9:92; *V-Lo/Dohna*, 24:752); examples of ideas would include complete virtue, freedom, holiness, God, etc. An ideal, by contrast, is an imagined instance of this concept, a "representation of an individual adequate to the idea" (*KU*, 5:232; see also *KrV*, A567-9/B595-7; *V-Mo/Collins*, 27:423; *V-Mo/Mron II*, 29:604; *V-MS/Vigil*, 27:675, 27:680; *Anth*, 7:199-200). An ideal, on Kant's view, is thus a particular and not a concept, for example, the Christ-figure or the stoic sage as representations of virtue. All references to Kant are either to page number in the Akademy Edition or, in the case of the *Critique of Pure Reason* to page numbers in the first and second editions. For abbreviations I use the *Kant Studien* list of sigla.

I claim little originality for my interpretation. At most I add a variant perspective to the already detailed picture of Kant's *Groundwork* drawn by others.[2] I do, however, think that my interpretation provides a fresh perspective on some persistent questions about Kant's moral philosophy. Kant provides several different statements of practical principles in the *Groundwork*, but he also says that there is "only a single categorical imperative," (*GMS*, 4:421) and that the different principles "are at bottom only so many formulae of the very same law." (*GMS*, 4:436) This variety of principles and their supposed equivalence has been the source of much scholarly dispute. Three issues have been especially thorny: first, if and how the different principles are statements of one and the same law or principle. Second, if and how these statements are supposed to be coextensive as a matter of prohibited, permitted, and required maxims of action.

[2] Inevitably, my interpretation takes a stance in the ongoing dispute concerning which of the various statements of the categorical imperative offered by Kant is fundamental or most illuminating. My stance is that, though the formulation standardly referred to as the formula of universal law is a version of the principle of morality, they're all necessary but only jointly sufficient specifications of the conditions of full autonomy. Thus it may be said that I side with some recent attempts at emphasizing the principle of autonomy offered by Wood *Kant's Ethical Thought* (Cambridge, UK: Cambridge University Press, 1999); "The Supreme Principle of Morality," in *The Cambridge Companion to Kant and Modern Political Philosophy*, P. Guyer ed. (Cambridge, UK: Cambridge University Press, 2006), pp. 342-80); Reath *Agency & Autonomy in Kant's Moral Philosophy: Selected Essays* (Oxford: Oxford University Press, 2006); Shell *Kant and the Limits of Autonomy* (Cambridge, MA: Harvard University Press, 2009); Guyer "The Possibility of the Categorical Imperative," in *Kant's* Groundwork for the Metaphysics of Morals*: Critical Essays*; *Kant* (NY: Routledge, 2006), 203-7; *Kant's* Groundwork for the Metaphysics of Morals*: Reader's Guide* (NY: Continuum International Publishing Group, 2007) chapter 5; "Kant on the Theory and Practice of Autonomy," in *Kant's System of Nature and Freedom: Selected Essays* (Oxford: Oxford University Press, 2005), pp. 115-45; "Naturalistic and Transcendental Moments in Kant's Moral Philosophy," *Inquiry*, vol. 50, 5, October 2007, pp. 444-464; and Uleman *An Introduction to Kant's Moral Philosophy* (Cambridge, UK: Cambridge University Press, 2010), chapter 6. And surely I'm inspired by and owe much to these interpreters. Yet, as will become clear, my interpretation differs from theirs. Interpretations that, by contrast, emphasize the universality requirement of the formula of universal law include Paton, *The Categorical Imperative: A Study in Kant's Moral Philosophy* (NY: Harper and Row, 1947); Nell (later O'Neill) *Acting on Principle: An Essay on Kantian Ethics* (NY: Columbia University Press, 1975); "Consistency in Action," in Guyer ed. *Kant's* Groundwork of the Metaphysics of Morals*: Critical Essays* (NY: Rowman & Littlefield, 1998), pp. 103-31; Kerstein *Kant's Search for the Supreme Principle of Morality* (Cambridge, UK: Cambridge University Press, 2002); Pogge "The Categorical Imperative," in *Kant's* Groundwork of the Metaphysics of Morals: *Critical Essays*, pp. 189-213; Sedgwick *Kant's* Groundwork of the Metaphysics of Morals*: An Introduction* (Cambridge, UK: Cambridge University Press, 2008); Engstrom *The Form of Practical Knowledge: A Study in the Categorical Imperative* (Cambridge, MA: Harvard University Press, 2009). Interpretations that emphasize the formula of humanity include Dean *The Value of Humanity in Kant's Moral Theory* (Oxford: Oxford University Press, 2006); and Korsgaard *Creating the Kingdom of Ends* (Cambridge, UK: Cambridge University Press, 1996), chapter 4 and 7. Finally, Thomas E. Hill Jr. has done much to underscore the role of the formula of the kingdom of ends, cf. Hill Jr. *Dignity and Practical Reason in Kent's Moral Theory* (Ithaca, NY: Cornell University Press, 1992).

And, third, which, if any, of the stated principles that is the primary criterion of right and/or the basic principle in Kant's derivation of the system of duties. I argue that the different principles, that is, the different formulaic statements of the categorical imperative, each emphasize different aspects of a fundamental principle, namely, the principle of autonomy that is the supreme principle of morality. From this follows that, first, the different statements of the categorical imperative are *not* statements of one and the same principle, but are, rather, statements of different requirements of one and the same principle, namely the principle of autonomy. A permissible maxim satisfies all of the requirements, if a maxim violates one of the requirements it is impermissible. Second, though the different formulas must be consistent (that is, no formulation could require what another prohibits), they are *not* co-intensive, nor need they be co-extensive as a matter of prohibited, permitted, or required maxims. And, third, that none of the statements singularly presents Kant's criterion of right or is the primary principle in Kant's derivation of the system of duties, for each of them is necessary and none of them sufficient for the complete specification of what morality requires – which, I maintain, is simply the conditions of autonomous willing. So, my argument shows that the different formulas are not intensionally equivalent, need not be extensionally equivalent, and that neither of them is Kant's primary criterion of right, since each is a necessary member of the jointly sufficient set of conditions of autonomous, moral volition that the formulas taken together present.

2. Kant's *Groundwork*

In the *Groundwork* Kant identifies the basic principles that structure our use of the basic concepts of practical cognition. Kant labels these principles 'categorical' because they direct the use of the basic normative concepts (or categories) and as such are *a priori* valid for all rational agents independently of factual contingencies, and 'imperatives' because for embodied rational

agents like us these principles appear as constraints (*GMS*, 4:413-4; *KpV*, 5:19-20). The categorical imperatives structure (or ought to structure) our cognition of reality as a place where we pursue ends and decide to do things in a manner analogous to how the principles of the understanding structure the cognition of nature as a subject of knowledge. The difference is that the principles of the understanding apply concepts to objects that exist in accordance with natural laws, whereas the categorical imperatives apply concepts to ourselves as subjects that ought to exist in accordance with moral laws.[3]

In the preface to the *Groundwork* Kant identifies its aim as "nothing more than the search for and establishment of the *supreme principle of morality*." (*GMS*, 4:392) These two projects of search for and establishment of the supreme principle of morality are carried out in the two main parts of the *Groundwork*, which Kant labels, respectively, the metaphysics of morals and the critique of practical reason.[4] The metaphysics of morals is, Kant explains, the examination of the idea and principles of pure moral willing (*GMS*, 4:391) – in short, the *search* for the supreme principle of morality. This part of the argument is preceded by a clarification of common sense and philosophical notions of morality, notions that the metaphysics of morals then give a systematic analysis and restatement. Because the argument of the metaphysics of morals is analytical, it concludes in a conditional: that if there is a supreme principle of morality, then that principle is the principle of autonomy (*GMS*, 4:440). The metaphysics of morals is followed by the second main part of Kant's argument, the critique of practical reason, which presents the *establishment*

[3] E.g. *KpV*, 5:20. This is/ought gap allows Kant's reconstruction of special metaphysics that is the culmination of all each of the three *Critiques*. I am here glossing over all the important details about how 'reality as a subject of knowledge' must be understood given Kant's transcendental idealism.

[4] Here the term "metaphysics of morals" has a different meaning than it does in the title and text of the book *The Metaphysics of Morals*, where it concerns the *application* of the basic principles of practical reason to the human condition. The term "critique of practical reason" has roughly the same meaning as it does in the book of that title.

of the objective reality of freedom and therewith also validates the supreme principle of morality identified in the metaphysical part.

It is important not to mistake the progress of the sections of the *Groundwork* for the progress of the argument advanced in it. Section I moves from common sense to philosophical understanding of morality. Section II starts with a restatement and elaboration of the philosophical understanding and then moves on to the metaphysics of morals that concludes in the formulation of autonomy as the supreme principle of morality. Section III starts with a restatement and elaboration of the metaphysics of morals and then moves on to the critique of practical reason that concludes in the affirmation of the objective reality of freedom of the will and the correlative validity of autonomy as the supreme principle of morality. Thus, the *Groundwork* proceeds through four moments – common sense (4:393-7), philosophical analysis (4:397-426), metaphysics of morals (4:426-48), and critique of practical reason (4:448-63) – and the second and third of these moments span across two sections each. This is important, because one risks taking a result of one kind of argument for one of another. It is, for example, a mistake to take the conclusion of the philosophical analysis that is the statement of the categorical imperative at 4:421 for the conclusion to the metaphysics of morals. This mistake has, I think, been the source of some confused discussions of how the statement of the categorical imperative that concludes the philosophical analysis (the so-called 'formula of universal law') is supposed to include or entail Kant's other moral principles and cover all possible moral requirements through a simple procedure or test of the universalization of the maxim of action.[5] Such discussions mistakenly assume

[5] I generally do not think that any of the formulations of the categorical imperative or these taken together are meant to provide a decision procedure. In this I disagree with a number of interpretations that take the categorical imperative to provide a decision procedure, e.g. Onora Nell (later O'Neill), *Acting on Principle: An Essay on Kantian Ethics* (New York: Columbia University Press, 1975); John Rawls *Lectures on the History of Moral Philosophy* (Cambridge, MA: Harvard University Press, 2000), 167-70; Andrews Reath "Legislating the Moral Law," in *Agency and Autonomy in Kant's Moral Theory*, pp. 92-120, at 107; Thomas Pogge "The Categorical Imperative," in Paul Guyer ed. *Kant's Groundwork of the Metaphysics of Morals: Critical Essays* (Lanham, MD: Rowman and Littlefield Pub-

that Kant's analysis of morality concludes where it begins: the formulation of the categorical imperative in the form of the formula of universal law is the conclusion to the second, philosophical moment of the argument, and this conclusion then serves as the starting point of the third, metaphysical moment. Another example of how inattentiveness to the structure of Kant's argument can lead to mistakes is that one might mistake the propadeutic to the argument for the objective validity of the categorical imperative with which Kant begins section III, and which includes a restatement of the results of the metaphysics of morals (4:446-8), for the actual argument which is given only once the analysis is completed, and therefore confuse the analytical exposition of the concepts of freedom and morality for the synthetical argument that is the real achievement of the critique of practical reason.

My concern is with the analytical argument that Kant performs in the metaphysics of morals part of the *Groundwork*. In particular, I want to understand why Kant concludes section II with the claim that autonomy is the supreme principle of morality, what this principle is, and the relation between this and the various other practical principles that Kant formulates.

It is useful to start with an overview of the interpretation I offer. To provide this overview I begin with an analogy between Rousseau's conception of the legitimacy of political laws and Kant's account of morality.[6]

3. Outline of my Interpretation: an Analogy between Rousseau and Kant

---

lishers, Inc, 1998), chapter 8. For discussion see Barbara Herman's "Moral Deliberation and the Derivation of Duties," in *The Practice of Moral Judgment* (Cambridge, MA: Harvard University Press, 1993), chapter 7.

[6] I am not the first to draw attention to analogies between Rousseau and Kant, but I believe that the analogy I present in the following has been overlooked. See Ernest Cassirer's *The Question of Jean-Jacques Rousseau*, P. Gay transl. (Bloomington: Indiana University Press, 1963); *Rousseau, Kant, and Goethe*, J. Gutmann transl. (NY: Harper and Row, 1963); Frederick Neuhouser *Rousseau's Theodicy of Self-Love: Evil, Rationality, and the Drive for Recognition* (Oxford: Oxford University Press, 2008); part IV; Andrew Levine *The Politics of Autonomy: A Kantian Reading of Rousseau's Social Contract* (Amherst: The University of Massachusetts Press, 1976); Richard L. Velkley *Freedom and the End of Reason* (Chicago: University of Chicago Press, 1989); Wood *Kant's Ethical Thought*, part II; Reath "Legislating the Moral Law," in *Agency & Autonomy in Kant's Moral Theory*, 94-6.

Rousseau presents a set of three necessary and jointly sufficient conditions for the legitimacy of political laws: conditions of form, content, and source.[7] First, laws must have the *form* of laws, that is they must command unconditionally and be directed universally. Second, laws must have the right *content*, that is, they must serve the common good. Third, laws must be issued by rightful authority, which can be none other than those who are also subjected to the laws. All three conditions are derived from the basic idea of popular sovereignty and its basic principle that laws are legitimate only if they genuinely express the general will.

The idea of popular sovereignty entails the three conditions of form, content, and source, because all and only laws that satisfy the three conditions are genuine expressions of the general will. Even if a putative law issues from a process of democratic will-formation and serves the common good, it isn't a genuine law without the proper form of universality; for without the proper form, it does not address or concern all citizens simply as citizens, but, say, gives some special treatment by function of their particular features (- and as such expresses a particular and not the general will). Even if a putative law has the proper form and issues from a proper process of democratic will-formation, it is not an expression of the general will unless it serves the common good; the proper genesis and form of the law may suffice to create an obligation to obey it, but the law may nevertheless be bad or unjust and therefore not a genuine expression of the general will. Finally, even with the right form and the proper content, a putative law that is not democratically generated cannot claim to express the general will and therefore cannot obligate those subjected to it; no matter how benevolent a dictator is, he cannot claim that his commands

[7] An overstatement, for at least one condition is left out; that laws must be backed by proper sanctions. The interpretation of Rousseau I present in this section is based on the *Discourse on Political Economy*, *The Social Contract*, the *Geneva Manuscript*, the *Letters Written from the Mountain*, and his occasional writings on political issues (such as the projects for Corsica and Poland). I have argued for this interpretation elsewhere, but since I here only use Rousseau to draw out the structure of a certain approach to the connection between freedom, morality, and law nothing hangs on the accuracy of my interpretation and I thus abstain from supporting my interpretation with quotes and references to Rousseau's writings.

express the general will and thus cannot issue legitimate laws. So, any law that fails on one of these three conditions is not a genuine expression of the general will and is, therefore, illegitimate. On the other hand, laws with the proper universal form, that properly serve the common good, and have the proper democratic genesis are genuine expressions of the general will and are, therefore, legitimate. Moreover, if all three conditions are jointly and generally satisfied in a political system, this system will be a legitimate and just republic, wherein any member by obeying the laws obeys only herself. A political system that systematically satisfies all three conditions is, therefore, Rousseau's political ideal – the true republic where all political authority properly express the will of the people.

In sum, Rousseau's analysis of the legitimacy of law has the following elements: the *basic idea* of popular sovereignty, the *basic principle* that all laws must be genuine expressions of the general will, the *validity-conditions* of form, content, and source that follow from this principle, and, finally, the *ideal* of the republican rule of law that results if the validity conditions are satisfied. I claim that we can understand Kant's analysis of moral willing in terms of an analogue set of elements. First, the *basic idea* is the idea of autonomy: the idea of the agent governing herself in accord with moral reasons. Second, the *basic principle* is the principle of autonomy: the principle that maxims of action are permissible only if they conform with the moral laws given by the will to itself. Third, Kant identifies three *validity-conditions* as individually necessary and jointly sufficient for maxims to be genuine expressions of proper moral, autonomous willing. The three validity-conditions are: first, the condition of form: that maxims must have the form of universal law; second, the condition of content: that maxims must have the content of humanity; and, third, the condition of source: that the moral commands must be issued by rightful authority, namely by the will of the agent that is also their subject. When these three conditions are jointly

satisfied the agent wills autonomously and, therefore, acts on permissible maxims of action. Conversely, when one or more of these conditions is not satisfied, there is a failure of moral deliberation and the maxim of action that results is impermissible. Thus each of the conditions alone can serve as a source of prohibitions (- as illustrated by the examples in the *Groundwork*). But permission requires that the maxim satisfies all three conditions. Finally, the consistent compliance with the validity conditions results in three *ideals*: the individual ideal of the good will and the collective ideals of the autonomy of every rational will and the system of ends that is thereby realized (as presented by the formula of the realm of ends).

So, it is the case for both Rousseau and Kant that a basic idea of self-governance turned into a basic normative principle is the source of the validity-conditions on, and ideals of, rightful willing. For Rousseau this analysis applies to political willing. For Kant it applies to individual willing. But their fundamental idea is the same: that rightful willing is a kind of self-governance consists in subjection to valid laws issued by the subject itself. And though their different subject matters yield different analyses of the condition of content (the common good versus humanity) and differing ideals (the just republic versus the good will and realm of ends), their analyses of the validity conditions on moral legislation are analogous in terms of the set of conditions (form, content, source) as well as in terms of what the conditions of form and source require. These analogies can, of course, be explained by the analogy of their basic idea and principle: the idea and principle of just and moral self-determination – popular sovereignty and personal autonomy.

So far my interpretation. The next two subsections present the textual basis for it.

4. Outline of Kant's Argument

The structure of Kant's argument for the conclusion that autonomy is the supreme principle of morality is clear enough. As a sort of transition from philosophical analysis to the metaphysics of

morals is a brief paragraph that ties the concepts of the good will and worth to the concept of freedom (at 4:426). This is the first mention of freedom since the preface and it signals that the philosophical analysis just completed was a preliminary, now the real work begins: to ascertain the grounds and content of the categorical imperative with which the philosophical analysis concluded through an analysis of the internal connection between freedom, morality, and rational being. And so, "we must step forth [...] into the [...] metaphysics of morals." (*GMS*, 4:426-7)

The topic of the metaphysics of morals is restated as a question of proper self-determination, the "question of objective practical laws and hence of the relation of the will to itself insofar as it determines itself only by reason." (*GMS*, 4:427) The first step of the metaphysics of morals is, accordingly, a discussion of the faculty of the will as a capacity for self-determination in accord with the representation of laws (*GMS*, 4:427). Kant moves directly from the will as a capacity for self-determination in accord with laws to the introduction of a distinction between subjective (contingent) and objective (necessary) ends (*GMS*, 4:428), and then redefines the question of the possibility of a categorical imperative in terms of the possibility of objective ends.[8] There is, Kant claims, only one thing that could have the status as objective end, namely, humanity.[9] Thus Kant concludes the introduction to the metaphysics of morals: "the practical imperative will therefore be the following: So act that you use humanity whether in

[8] "[S]uppose there were something the *existence of which in itself* has an absolute worth, something which as *an end in itself* could be a ground of determinate laws; then in it, and in it alone, would lie the ground of a possible categorical imperative, that is, of a practical law." (*GMS*, 4:428) The possibility of a necessary connection between rational being and morality thus depends on the possibility of necessary ends. This claim is stated quite clearly in notes from Kant's lectures: "[m]orality leads us to the principle of *necessary ends*, without which it would itself be only a chimera." (*V-Phil-Th/Pölitz* 28:1075) See also *GMS*, 4:428-9;. In emphasizing necessary ends I'm indebted to Paul Guyer, *Kant's Groundwork for the Metaphysics of Morals*, 89-92; "The Possibility of the Categorical Imperative," 228-34; *Kant*, 205-6; "Form and Matter of the Categorical Imperative," in *Kant's System of Nature and Freedom: Selected Essays* (Cambridge, MA: Cambridge University Press, 2005), pp. 146-68; "Ends of Reason and Ends of Nature: The Place of Teleology in Kant's Ethics," also in *Kant's System of Nature and Freedom*, 169-97.

[9] *GMS*, 4:429, 4:436. Happiness is a subjective ultimate end.

your own person or in the person of any other, always at the same time as an end, never merely as a means." (*GSM*, 4:429, italics removed)

Next, Kant introduces "a third practical principle of the will [...] the principle of the will of every rational being as a will giving universal law through all its maxims." (*GMS*, 4:431, 4:432, italics removed) Kant explains that according to this principle any maxim that is inconsistent with the will's own giving of universal law is repudiated, so that "the will is not merely subject to the law but subject to it in such a way that it must be viewed as also giving the law to itself and just because of this as first subject to the law (of which it can regard itself as the author)." (*GMS*, 4:431) All previous attempts at identifying the supreme principle of morality, Kant claims, failed to understand that even in his subjection to the moral commands the agent must remain sovereign: he must "be subject only to laws given by himself but still universal and that he is bound only to act in conformity with his own will, which, however [...] is a will giving universal law." (*GMS*, 4:432) Accordingly, Kant contrasts this "principle of autonomy of the will," with all other principles of morality as principles of heteronomy.[10] Kant then suggests that the principle of autonomy underwrites a "very fruitful concept [...] namely that *of a realm of ends*," (*GMS*, 4:433) which he presents as the idea of "a whole of all ends in systematic connection." (*GMS*, 4:433)

[10] *GMS*, 4:433. Kant's argument here is an abbreviated version of the argument we find throughout the Kantian corpus, that all other moral philosophies identify the wrong supreme principle of morality that has the will determined by material determining grounds and as such heteronymous. Kant's argument has the structure of an argument by exclusion of two times two sets of alternatives, for example: "Autonomy is legislation of another sort, where there is neither feeling [empirical-external], nor inclination [empirical-internal], nor speculative reason [objective-internal], nor another will [objective-external]; my actions, in this case, are good insofar as I can consider my will to be self-legislating therein." (*V-Mo/Mron II*, 29:629) Kant repeats variations of this argument against previous moral philosophies throughout his ethical writings, see *GMS*, 4:441-4; *KdV*, 5:39-41; 5:64-5; *V-Mo/Collins*, 27:252-60; *V-MS/Vigil.* 27:497-500; 27:274-8; *V-Mo/Mron II*, 29:619-29. See also *Refl.*, 6631, 6637. I am not here concerned with the soundness of the argument, but largely agree with Wood's and Schneewind's critical assessments; cf. Wood's "The Supreme Principle of Morality," 369-73; Schneewind's "Kant against the 'spurious principles of morality'," in *Kant's Groundwork of the Metaphysics of Morals: A Critical Guide*, J. Timmerman ed. (Cambridge, UK: Cambridge University Press, 2009).

After the introduction of the principle of autonomy and the idea of a realm of ends, Kant ties the principles of humanity as an end in itself, autonomy, and the realm of ends to the concepts of freedom, virtue, and dignity, and finds that autonomy "is the ground of the dignity of human nature and of every rational nature." (*GMS*, 4:436) Kant then concludes this initial analysis and exposition of the principles with a statement of their interrelation. Kant writes that the three identified principles are "three ways of presenting the principle of morality [...] at bottom only so many formulae of the same law," and explains that they are related in terms of the form, matter, and complete determination of morality (*GMS*, 4:436). The formula of the law of nature expresses the requirement of universality of form of maxims, the formula of humanity as an end in itself presents the necessary matter, and the realm of ends presents the idea (we might say *ideal*[11]) of a "complete determination of all maxims by means of that formula, namely that all maxims from one's own lawgiving are to harmonize with a possible kingdom of ends as with a kingdom of nature." (*GMS*, 4:436) After this statement of the relation between the principles, Kant provides a summary of the argument of *Groundwork* section II and explains how the analysis has provided a metaphysics of the main concepts of *Groundwork* section I – the good will, moral worth, and dignity (*GMS*, 4:437-40). Kant finally concludes the metaphysics of morals part of the *Groundwork* by stating that "morality is [...] the relation of actions to the autonomy of the will [...] an action that can coexist with the autonomy of the will is permitted; one that does not is *prohibited*."[12] And so, we reach our destination: that autonomy of the will is the supreme principle of morality; "the sole principle of morals."[13]

[11] As Kant indeed says, *GMS*, 4:433.
[12] *GMS*, 4:439.
[13] *GMS*, 4:440-1. Similarly Kant is recorded as saying in lectures given around the same time he was working on the *Groundwork* that "The principle of morality is the Idea of a will, insofar as it is a law unto itself, for what it will is always a universal law, and that is the good will. [...] The agreement of an action with the principle of my will, as universal legislator, is thus the principle of morality." (*V-Mo/Mron II*, 29:628)

I won't here pursue the further argument for the objective reality of the concept of autonomy and the validity of the moral principles that is unfolded in *Groundwork* III.[14] Instead, I want to pursue two questions: What is the supreme principle of morality (autonomy)? And, how is this principle related to the other principles stated in *Groundwork* II?

5. Autonomy as the Supreme Principle of Morality

Whereas the structure of Kant's argument for the conclusion that autonomy is the supreme principle of morality is clear enough, it is less clear what the argument is or, indeed, what the conclusion means. Two issues in particular need clarifying. First, what the principle of autonomy is. And, second, if and how it is related to the various principles of morality that Kant states.

1. The first issue may seem contrived insofar as Kant actually provides a fairly clear answer. When he first introduces the 'principle of autonomy' it is stated as "the *principle* of every human will as *a will giving universal law through all its maxims*." (*GMS*, 4:432) And, in the paragraph titled Autonomy as the Supreme Principle of Morality, he writes: "the principle of autonomy is, therefore, to choose [zu Waehlen] only in such a way that the maxims of your choice [Wahl] are included as universal law in the same volition." (*GMS*, 4:440) Moreover, the idea of autonomy as the supreme principle of morality is presented as follow in the *Critique of Practical Reason*:

> *Autonomy* of the will [des Willens] is the sole principle of all moral laws and of duties in keeping with them [...] the sole principle of morality consists in independence from all matter of the law (namely, from a desired object) and at the same time in the determination of choice [Willkuer] through the mere form of giving universal law that a maxim must be capable of. That *independence*, however, is freedom in the *negative* sense, whereas this *lawgiving of its own* on the part of pure and, as such, practical reason is freedom in the *positive* sense. Thus the moral law expresses nothing other than the *autonomy* of

[14] I have tried to make sense of Kant's argument for the objective reality of freedom in "Freedom as Both Fact and Postulate," *Proceedings of the XI International Kant Congress* (Berlin: Walter de Gruyter, *forthcoming*).

> pure practical reason, that is, freedom, and this is itself the formal condition of all maxims, under which alone they can accord with the supreme practical law.[15]

The principle of morality mentioned in this passage, it seems, is the "fundamental law of pure practical reason" (*KpV*, 5:30) that was presented in two ways in the preceding section. First, as "[s]o act that the maxim of your will could always hold at the same time as a principle in a giving of universal law." (*KpV*, 5:30) And, second, the corollary of the imperative: "[p]ure reason is practical of itself alone and gives [...] a universal law which we call the *moral law*." (*KpV*, 5:31)

With these statements in mind, it is fair to assume that the following chain of equivalents presents Kant's analysis of the supreme principle of morality: the supreme principle of morality is the principle of autonomy and the supreme principle of autonomy is the categorical imperative and the categorical imperative is the principle stated in both the *Groundwork* and the second *Critique*, namely, the imperative to act only on maxims that can at the same time be willed as principles in a giving of universal law.

So, the immediate answer to my first question, concerning what the principle of autonomy is, is that the principle of autonomy is the statement of the categorical imperative standardly referred to as the formula of universal law.[16]

However, while I think it is indisputable that Kant affirms something like the chain of equivalents presented above and that a version of the answer just given is correct, I nevertheless think that it is a mistake to think that the supreme principle of morality, the principle of autonomy, simply is the universality requirement that is implied by the so-called formula of universal law.

---

[15] *KpV*, 5:33. Compare *GMS*, 4:440-1; 446-7; *V-MS/Vigil.* 27:499.

[16] This appears confirmed in the lecture notes: "The agreement of an action with the principle of my will, as a universal legislator, is thus the principle of morality. If we cannot consider our will to be universally legislative, we reject the action." (*V-Mo/Mron II*, 29:628)

The categorical imperative is first stated as the result of the philosophical analysis part of the *Groundwork*. At that point it requires that maxims could be willed as *universal* law. And Kant illustrates how the requirement of willed universality alone yields various prohibitions. But this statement of the imperative is not its final statement and what it requires is not merely universalization. Its final statement and the full analysis of the principled requirements of autonomy are the topics of the metaphysics of morals part of the groundwork. Here Kant analyses what could be willed as universal law. Universality is implied, of course, but maxims must also be in accord with principles that could be *willed* in a *giving* of *law*. And how these requirements are fulfilled is barely touched upon prior to the analysis of the metaphysical part of the *Groundwork*. It is only at the end of *Groundwork* II that Kant has articulated the meaning and requirements of the categorical imperative. In a sense one might say that the categorical imperative that is stated prior to the metaphysics of morals part of the groundwork both is and is not the same as the categorical imperative that is the principle of autonomy and supreme principle of morality. It is the same in the sense that philosophical analysis and the metaphysics of morals conclude in the same principle, so that the requirement that maxims must be of a kind that could be willed as universal law is the same principle. But it is not the same in the sense that the grounding and content of this principle are undetermined prior to the analysis given with the metaphysics of morals. In short, and trivially, the principle lacks grounding and systematic content prior to the metaphysics of morals – this is trivial, because of how Kant understands the relation between philosophical understanding and metaphysics. So, the question remains: what is the principle of autonomy that Kant at the end of the metaphysics of morals concludes is the supreme principle of morality?

To answer this question, I will move in the opposite direction of the progression of the metaphysical part of the *Groundwork*. That is, I start from the concept and requirements of au-

tonomy and work back to the categorical imperative. Along the way I try to distinguish more clearly than Kant did between the concept, idea, principle, and requirements of autonomy.

Kant's *concept of autonomy* is a conception of freedom, namely, freedom as the will's self-determination in accord with self-legislated moral laws: "the property of the will by which it is a law to itself."[17] Since every rational being is (or ought to be) free in this sense, the concept of autonomy yields the *idea* of autonomy, again, the "idea *of the will of every rational being as a will giving universal law*." (*GMS*, 4:431)

Before we proceed to what the requirements of autonomy are or what the property of autonomy consists in, we should ask what autonomy is a property of. Since pure practical reason (*Wille*) could not choose other laws to legislate than the moral laws it is, properly speaking, not free. Properly speaking, it is only choice (*Willkuer*) that is free:

> Laws proceed from the will [Wille], *maxims* from choice [Willkuer]. In man the latter is a free choice; the will [Wille], which is directed to nothing beyond the law itself, cannot be called either free or unfree, since it is not directed to actions but immediately to giving laws for the maxims of actions (and is, therefore, practical reason itself). Hence the will [Wille] directs with absolute necessity and is itself *subject to* no necessitation. Only *choice* [Willkuer] can therefore be called *free*. (*MdS*, 6:226)

So, pure practical reason is not free. But neither is it unfree, for it is not determined, since its legislation is external object independent: "[t]he will [Wille] itself, strictly speaking, has no determining ground; insofar as it can determine choice, it is instead practical reason itself."[18] It is tempting, but would be misleading, to say that pure practical reason is autonomous. Tempting, because it is reason legislating the moral law to the will as a whole. Misleading, since pure prac-

[17] *GMS*, 4:440, 4:447, see also *KpV*, 5:33. In a note from the late 1780s Kant equates the positive concept of freedom with "autonomy through reason." (*R*, 6076, 18:443)

[18] *MdS*, 6:213. Likewise: "Everything in nature works in accordance with laws. Only a rational being has the capacity to act *in accordance with the representation* of laws, that is, in accordance with principles, or has a *will* [einen Willen]. Since *reason* is required for the derivation of actions from laws, the will is nothing other than practical reason." (*GMS*, 4:412)

tical reason is not legislating the law to itself, but to the will of which it is a part. The faculty of choice, on the other hand, is free but not autonomous. Autonomy, then, is a property of the will as a whole and not of any of its subfaculties – it is the property of the will giving the moral law to itself.[19]

In its most abstract form, the basic *principle of autonomy* is the concept of autonomy turned into a normative principle, that is, the principle that persons ought to be autonomous and, therefore, ought to be self-determining in accord with the requirements of morality. We can state the principle of autonomy in terms of faculties: the will as executive faculty ought to choose maxims of action that are in accord with the moral laws issued by the will as legislative faculty so that the will as a general faculty is self-determining. Or, the principle can be stated in terms of the proper configuration of the reasoning of the autonomous agent: she acts only on maxims that she could also will as principles in the giving of universal laws.

Since autonomy is the proper configuration or activity of moral willing, the principle of autonomy can be further specified in terms of the negative and positive aspects and requirements of autonomous willing. The basic *requirement of autonomy* is that the will must be self-determining (or self-governing) and therefore both source and subject of the moral laws or commands.[20] The negative aspect of this requirement is that the will must be sovereign authority, and thus cannot be unconditionally subject to the commands of any external authority (the '*auto*' part

[19] Though Kant in the second *Critique* says calls both pure practical reason and Willkuer autonomous, I think this is just an expression of careless terminology. Examples: "the moral law expresses nothing other than the *autonomy* of pure practical reason, that is, freedom, and this is itself the formal condition of all maxims, under which alone they can accord with the supreme practical law;" (*KpV*, 5:33) "...autonomy of choice [*Willkuer*]," (*KpV*, 5:36).

[20] What I call the requirement of autonomy is roughly what Reath calls the Sovereignty Thesis (in "Autonomy of the Will as the Foundation of Morality," in *Agency & Autonomy in Kant's Moral Theory*, 122, see also 126, 137, 205). While my reading agrees with Reath's in some respects, I depart from his reading of the formula of autonomy and, accordingly, from his claim that the formula of universal law is equivalent to the formula of autonomy. More generally, I cannot follow his reading of Kant as a constructivist, but that's a beef I have with most of the children of Rawls.

of autonomy).[21] The positive aspect (the '*nomos*' part of autonomy) is that the will must determine itself through self-legislated moral laws. Since the will must be bound by the unconditional commands of morality, these two requirements together entail that the will must be both source and subject of the moral laws.

The two parts of the requirement of autonomy result in two different requirements. First, the principle of autonomy presents the negative requirement that maxims of action must not subject the will to any external authority. This presents a necessary condition on maxims of action, namely, that persons ought always to act on "maxims of one's will as a will that could at the same time have as its object itself as giving universal law." (*GMS*, 4:432) When this condition is satisfied, the agent "is subject *only to laws given by himself but still universal* and [...] bound only to act in conformity with his own will." (*GMS*, 4:432) In this manner, the constraint that is the necessitation of duty is self-constraint: "since the human being is still a *free* (moral) being [...] the constraint that the concept of duty contains can be only self-constraint [...] for only so can the *necessitation* [...] be united with the freedom of his choice." (*MdS*, 6:380) The negative requirement thus identifies a necessary condition of moral volition, namely, that the *source* of the moral commands to which the will is a subject must be the will itself; thus "all maxims are repudiated that are inconsistent with the will's own giving of universal law." (*GMS*, 4:431) This, then, is the requirement that the will must be the source of the commands to which it is a subject.

Second, the positive requirement concerns all of the conditions necessary for autonomous willing, including the source, and thus simply specifies the validity-conditions implied by the principle of autonomy.

[21] Conditionally binding commands, such as orders from a legitimate superior (say from a officer to a private) are not rejected by the requirement of autonomy. Such commands are conditionally binding because they are derived from unconditionally binding, autonomous acts of the subject (say, volunteering for the army).

It follows that the requirement that the will must be the source of the reasons it sees as commands is entailed by, but does not entail, the principle of autonomy. The analogy with Rousseau illustrates that there is nothing strange about the idea that the principle of autonomy entails or includes in it the source requirement, whereas the source requirement does not entail or include the principle of autonomy. This relation is analogous the relation between popular sovereignty, the principle of the general will, and democracy: the principle of popular sovereignty (that laws must express the general will) entails that laws must be generated by democratic procedures, by contrast, it is not the case that the claim that laws must have a democratic genesis entails the idea of popular sovereignty or the principle that laws must express the general will.

I must admit that the textual basis for a sharp distinction between the source-requirement and the principle of autonomy is limited. However, while Kant may not draw the distinction as clearly as we would like, there is a clear distinction to be drawn between the requirement that the will must not be subject to any external authority but must be self-governing (the source-requirement), and the requirement that the will ought to be self-governing through moral self-legislation (the principle of autonomy). Between, on one hand, the requirement that the will must remain sovereign and therefore the source of any commands to which it is subject, and, on the other hand, the positive account of what it is to be sovereign and what commands the will must be a subject of. The source-requirement captures the former, the principle of autonomy the latter.

2. With the preceding, we have already begun the answer to the second question: What is the relation between the principle of autonomy and the other principles presented by Kant?

The answer is that the different statements of the categorical imperative emphasize the different requirements of the principle of autonomy: the validity conditions of autonomous and moral willing. One of the requirements of the principle of autonomy was identified above: max-

ims must be compatible with the sovereignty of the will. This requirement entails that the will must be the *source* of the moral commands to which it is also a subject. Another requirement of the principle of autonomy is universality: that the maxims of action must be from a principle that is universal in *form*. Likewise, the principle of autonomy leads to the requirement that humanity must be treated as an end in itself. No maxim is without a matter, no rational act without an end.[22] So, the maxims by which the autonomous will determines itself have the *content* of humanity.

Finally, since the general realization of the principle of autonomy entails general, systematic, and complete moral willing, it entails the ideals of autonomy and the realm of ends. The idea of autonomy in this manner supports both an individual and a collective ideal. Full individual autonomy is complete self-determination in accord with the moral laws as issued by pure practical reason. This ideal can, of course, be described also in terms of the form or content of the fully autonomous will. And so, full individual autonomy is coincident with the good will. At the collective level, the idea of autonomy envisions the complete autonomy of all persons. The formal description of this ideal is simply the idea of autonomy, again, "idea of the will of every rational being as a will giving universal law." (*GMS*, 4:431) Stated in terms of the content of morality we get a different description of the same ideal, namely, the ideal of the realm of ends. The realm of ends, moreover, ties the moral ideals to the natural end of happiness. Humanity is the capacity to set and pursue ends subject to the requirements of morality, and part of what it is to treat humanity as an objective and final end (an 'end in itself') is to respect and promote the morally respectable ends set by others. This includes ends related to their pursuit of the necessary

[22] I here aim to remain agnostic as to the dispute over whether the formula of universal law yields positive duties (in part, because I think the positive / negative duties distinction is unstable). For an argument that it does not, see Wood *Kant's Ethical Thought*, 100-2; "The Supreme Principle of Morality," 344, 355. For an argument that it does, see Engstrom's *The Form of Practical Knowledge*, 209-23.

subjective good, happiness, which is why one of the ends that is also a duty is the promotion of the happiness.[23] If the formula of humanity is permanently and universally adhered to, the result is, therefore, the realization of the ideal of systematic unity of the ends of all rational beings, "a whole of all ends in systematic connection [...] that is, a realm of ends." (*GMS*, 4:433) In this manner, the formula of the realm of ends is an expression of the ideal envisioned by the principle of autonomy.[24] The connection between the individual and collective ideals and the concept of autonomy is indicated in notes from Kant's lectures given in 1785:

> If I picture to myself a kingdom of natural things, that are purposively ordered, even though the things themselves neither entertain the purposes, nor are causes of their existence, then that is the kingdom of nature under heteronomy. But I can also picture a kingdom of purposes with autonomy, which is the kingdom of rational beings, who have a general system of ends in view. In this realm, we consider ourselves as those who obey the law, but also as those who give laws. (*V-Mo/Mron. II*, 29:629)

While the collective ideal thus supports the hope for a necessary connection between what we ought to do and what we can hope for – in a term, the highest good – the realization of the moral ideal is not yet the realization of the highest good. For, while morality requires the promotion of morally permissible happiness and an internal connection between virtue and happiness is thereby implied by the formula of the realm of ends,[25] this internal connection is not yet a necessary connection between virtue and happiness and, therefore, not a sufficient basis for belief in the possibility of the highest good. The highest good requires complete virtue and happiness distributed in proportion to virtue and for this human effort is not enough; the cooperation of nature is

[23] Cf. *GMS*, 4:430; *MdS*, 6:385, 6:393; 6:450-1.
[24] Here it should be noted that my interpretation requires a departure from Gregor's translation. The passage where Kant mentions the connection between the three formulas is: "Die angefuehrten drei Arten [...] sind aber im Grunde nur so viele Formeln eben desselben Gesetzes, *deren die eine die anderen zwei von selbst in sich vereinigt*." (*GMS*, 4:436, my italics) Gregor translates the italicized part "and any one of them of itself unites the other two in it," I prefer the translation "and one of these unites the other two in it," the unifying one being the formula of the realm of ends. In *Kant's Ethical Thought* (p. 187) Wood follows Gregor's translation and therefore tries to construct an argument that each of the formulas contains the others, but Wood corrects this and provides a helpful discussion of how to translate the key sentence in "The Supreme Principle of Morality," see pp. 356, 376n12.
[25] See Guyer, "The Form and Matter of the Categorical Imperative," pp. 163-5.

needed as well. And so the gap between ought and hope remains unless and until we can have rational faith in the providential design of nature.[26]

Returning now to where the metaphysics of morals began, with the categorical imperative that concluded the philosophical analysis part of the *Groundwork*, we can say that this principle really is the supreme principle of morality; it is the same principle as the principle of autonomy. But what it requires, and why it requires it, only becomes clear once we have gone through the metaphysics of morals. By then it is also clear that the requirement of universality emphasized by this principle is merely one of the necessary conditions of autonomous willing that the principle of autonomy makes the supreme principle of morality.

The preceding paragraphs have indicated how the different formulaic statements of the categorical imperative can be viewed as spelling out the principled requirements and ideals of the principle of autonomy. Taken together they provide a complete determination of morality as autonomy. Neither of these requirements alone provides a complete account of moral willing. Willing in the universal form alone does not make a moral maxim. Having the right end, humanity, alone does not make a moral maxim. For the maxim must also have the right form and be issued by the rightful authority – the agent's own will. Nor is it sufficient that the maxim is issued by the agent's own will, for it must also have the right form and content. Only where all three conditions are satisfied is there an instance of a moral volition. So, Kant's basic deontic modalities are as follows: it is permissible to act on maxims that satisfy all three conditions; it is impermissible to act on maxims that violate at least one of the conditions. Moreover, only where all three conditions are consistently adhered to, individually and collectively, are the ideals of morality realized. Thus, the principle of autonomy, and only the principle of autonomy, identifies and entails

[26] So goes the argument in the Critique of Teleological Judgment.

all the necessary and jointly sufficient conditions of moral volition and specifies the idea and ideal of autonomy. The principle of autonomy is, therefore, the supreme principle of morality.

6. Conclusion

On my reading, the metaphysics of morals part of the *Groundwork* is an analysis of how freedom and morality come together in autonomy as the supreme principle of morality. I have tried to distinguish more clearly than Kant did between the concept, the principle, the validity conditions, and the idea(l)s of autonomy. Autonomy is moral self-determination. The principle of autonomy is that persons ought to determine themselves to act only for reasons that they could will as principles in the giving of universal law. When analyzed, this principle presents validity conditions of form, content, and source of maxims, and identifies the individual and collective ideals that are realized through free, moral, autonomous being.

If my interpretation is correct, it follows: first, that there is only one supreme principle of morality, namely the principle of autonomy; second, that the different formulas are formulas of the same law, because each express a requirement of the principle of autonomy; third, that since each is a necessary part of the jointly sufficient set of conditions of autonomous willing, the formulas are *not* intensionally equivalent; and, finally, that Kant provides no *a priori* reason that the formulas must be co-extensional.

Admittedly, my interpretation leaves many questions unanswered. I'll mention just one of these: Why does Kant think that the four formulas together provide a *complete* determination of the principle of autonomy and, therefore, of moral volition? I do not have an adequate answer. A tempting way to approach it would be in terms of the forms of judgment so that each of the formulas corresponds to one group of categories of practical judgment. But I don't think that this approach can work. It is more promising, I think, to draw an analogy between Aristotle's four

causes and the different formulas, so that each principle corresponds to one sort of explanation of free, teleological causality. Thus, without too much violence to Kant's texts, we might say that the form-condition of universality corresponds to the formal cause of moral actions, the content-condition of humanity corresponds to the material cause, the source-condition that the will must be the source of reasons corresponds to the efficient cause, and the ideals of autonomy and the realm of ends to the final cause. Yet, suggestive as the idea might be, it still seems to me to lack textual backing. So, I leave this question open.

***