12

Did Kant Hold that Rational Volition is *Sub Ratione Boni*?

Andrews Reath

1 Introduction

In an essay entitled "Personal Values and Setting Oneself Ends," Tom Hill casts doubt on what he rightly characterizes as a "strong thesis" about rational volition that several commentators have attributed to Kant. The thesis is "that anyone who has values at all is thereby implicitly committed to objective moral values" or that "adopting ends and maxims carries commitment to objective moral values that could ground moral claims."[1] The specific version of this thesis on which Hill focuses concerns the value judgments that "we implicitly make when we set ourselves personal ends and adopt maxims about how to achieve them," and claims that when one adopts a rationally optional end through practical reason, one takes one's end to have "objective value, in a robust sense."[2] This claim appears to initiate an argument to the effect that commitment to moral principle is built into rational choice per se, and that agents who violate moral principle act irrationally in the narrow sense of acting contrary to principles or commitments that they actually accept.[3]

Both Christine Korsgaard and Stephen Engstrom have defended versions of this thesis about the value commitments presupposed in rational choice of ends (including the end of happiness)—both as interpretation of Kant and on philosophical grounds. In her early well-known essay "Kant's Formula of Humanity," Christine Korsgaard attributes to Kant the view "when we act we take ourselves to be acting reasonably and so we suppose that our end is, in his sense, objectively good," that is, "provides reasons for action that apply to all rational beings."[4] I take it that the "objective goodness" that the agent supposes her end to have can be unpacked through the thoughts that

[1] Hill (2002: 244, 262). See also Hill (2002: 251, 258, 263, 266–7).
[2] Hill (2002: 246, 258).
[3] On the narrow sense of irrationality (practical irrationality is acting contrary to one's actual judgments about applicable principles or reasons), see Scanlon (1998: 25–30).
[4] Korsgaard (1996a: 116, 115).

pursuit of the end is supported by reasoning that justifies it to any rational agent, that it is a good thing from anyone's point of view that the agent achieve her end, and that others in certain relations to the agent have objective reasons to support her pursuit of it. As I understand it, the key idea (both to her reading of Kant and her own view) is that it is a necessary feature of rational choice of an end that the agent takes the end to be objectively good in this sense. This evaluative attitude is, as it were, a component of the self-consciousness of rational choice, part of the rational agent's own understanding of what goes into rational choice of an end. And she assumes that from this feature of rational self-consciousness one can uncover a commitment to morality that is likewise implicit in all rational choice of an end. Korsgaard does not deny that for Kant personal ends are "relative" or "subjective" ends that are of value to agents because of our interests or their fit with our nature. She argues that adoption of a relative or subjective end by an agent who is an end in itself confers objective value on the end in the above sense. The objective goodness that we suppose our relative ends to have presupposes that we understand rational choice to have a value-conferring capacity, and that in turn presupposes a valuing of rational nature as an end-in-itself that is implicit in all rational choice of ends.[5] The latter value commitment is, of course, fundamental to morality.

In the *Form of Practical Knowledge*, Engstrom develops what he terms a "practical-cognitivist conception of the will," one main element of which is that rational volition is based on practical judgments of goodness made under "the presupposition of universality." The cognitive aspect of volition is a judgment about good that is taken to be sharable by any rational subject—that is, it takes itself to satisfy a condition of universal validity.[6] This is a fact about the nature of rational volition—an essential feature of volition that is given to agents through practical self-consciousness, or the understanding of rational volition from the perspective of an agent engaged in it. Since human agents necessarily will own happiness as an end, we take our own

[5] In this respect, in Korsgaard's conception the valuing of the self and its rational capacities that is implicit in all choice is the framing assumption of practical reasoning and judgment, and is needed to back up the objective goodness that an agent takes her ends to have. In her later work, she continues both to ascribe this view to Kant and to defend it herself. Her main points are that we do take our ends to be valuable, that doing so supposes that rational choice is value conferring and that the capacity for rational choice has unconditional value. See e.g. Korsgaard (1996b: ch. 3, esp. p. 122 (§3.4.8)), and Korsgaard (2009: 122–3 (§6.3)). Her view in (1996b) is that valuing one's human identity is a condition of having reasons for action (because practical reasons are based in practical identities and our practical identities get their normative grip on us through our valuing our human identity), and that valuing one's human identity is implicit in all rational choice or finding an end to be valuable or worth pursuing. In (2009: 123) she writes: "As a Kantian, I believe that it is our own choices that ultimately confer value on objects, even though our choices are responsive to features of those objects. In choosing objects, in conferring value on things that answer to our nature in welcome ways, an agent is affirming her own value. She takes what matters to her to matter absolutely and so to be worthy of choice."

[6] Engstrom (2009: 124–7). Barbara Herman has argued for a similar conception of the will as a "norm-constituted power," where the principle that defines the power (as I understand her view) is the moral law. See "The Will and its Objects" in Herman (2007). The conception of volition that I explore in this chapter is close to (and has drawn on) Herman's and Engstrom's views.

happiness to be objectively good (*ceteris paribus*) and will it under the presupposition of universality. On Engstrom's reading of Kant, the Formula of Universal Law is the internal norm of rational volition and we are committed to following it by the self-consciousness that is part of volition.

Hill expresses skepticism both about this thesis and the argumentative use to which it has been put. He has the good sense to be wary of the idea that ordinary choice of subjective ends carries any commitment to their objective value—this thesis appears to overreach in the value commitments that it attributes to ordinary agents—and he does not think that there is clear textual support for saddling Kant with such a view. He allows that for Kant choice of an end carries some limited value judgments—e.g. that an action advancing such an end is good as a means, or good insofar as it contributes to one's happiness and in that sense good for an agent. But he thinks that Kant's attitude towards subjective ends is closer to Hobbes than many recent commentators believe, and argues that "[n]one of these commitments in adopting ends and maxims attributes any impersonal, nonrelative, or moral goodness to our personal ends as individuals." Furthermore, it seems obvious that agents often willingly and knowingly (intentionally) make morally unjustifiable choices. All told Hill thinks that it is a "very thin reed" on which to base an argument that it is irrational to act contrary to moral principle.[7]

The strong thesis about the value commitments implicit in rational choice of an end readily extends to broader issues about Kant's understanding of the nature of rational volition—in particular did he think that all rational volition proceeds *sub ratione boni*, and if so in what sense? In this chapter I would like to explore what can be said on behalf of ascribing to Kant a conception of rational volition that I would state as follows: that rational volition constitutively understands itself to satisfy a condition of universal validity. That is, rational volition is based on practical reasoning aimed at judgments of goodness that make a tacit claim to universality. As one might say, it is part of the self-consciousness of rational volition that it proceeds on maxims that are taken to satisfy a condition of universal validity. This amounts to the admittedly controversial claim that all rational volition is tacitly guided by something like the Universal Law version of the categorical imperative as its formal or internal constitutive norm. I shall refer to this conception as the "rationalist thesis" about the will. It is a claim about the nature of rational volition (not the moral psychology of choice), and since volition is constituted by its own self-understanding, a claim about the nature of volition as given in a rational agent's self-understanding of what it is to exercise the will. It does bear on the rational authority of morality, since the authority of morality would follow if the fundamental principle of morality is indeed the internal principle of rational volition, the principle that, as it were, we impose on ourselves through our understanding of what it is to exercise the will. However I am more interested in exploring the contours of the rationalist thesis than in using it to argue that bad conduct is contrary to

[7] Hill (2002: 261, 267).

principles and commitments that an agent actually accepts and is therefore irrational in a narrow sense.

Not having left all of Tom Hill's good sense behind, I have some misgivings about ascribing the rationalist thesis to Kant, and I don't claim to have resolved the interpretive issue in this chapter. While I believe that there is strong support for this reading, there are also clearly passages where Kant seems to allow that we can freely and intentionally choose actions and ends that we see are not universally justifiable. The texts are not decisive. My aim will be to see what support for the rationalist thesis as a reading of Kant can be drawn from the texts and to explore what it would do for his moral conception. In section 2, I canvas various discussions of the will for possible textual support, and in section 3 I argue that Kant's well-known theses about autonomy and free agency provide stronger philosophical support for this thesis. In section 4, I'll consider how this thesis can accommodate certain forms of bad willing and will address some counter-intuitive implications that it might be thought to have.

2 Does Volition Understand Itself to Satisfy a Condition of Universal Validity?—Initial Textual Support

As I've indicated, the rationalist thesis about the will that I wish to explore is the idea that rational volition constitutively understands itself as part of its self-consciousness to satisfy a condition of universal validity. If so, all rational volition, including morally bad choice, is tacitly guided by the Universal Law version of the categorical imperative in some form. (As I will sometimes put it: the Formula of Universal Law is the formal principle of rational volition.[8]) What might this conception of volition involve and is there textual evidence that Kant accepted it? In this section, I'll lay out some initial ideas.

In various well-known passages, Kant appears to identify the will with practical reason. In the *Groundwork*, Kant writes: "Only a rational being has the capacity to act according to the representation of laws, i.e., according to principles, or [has] a will. Since reason is required for deriving actions from laws [*die Ableitung der Handlungen von Gesetzten*], the will is nothing other than practical reason" (*G* 4: 412[9]). In a perfectly rational agent whose will is unfailingly determined by objective reason, "the will is a

[8] For this use of "formal principle" see Reath (2010: esp. section III).

[9] Citations to Kant are to the volume and page in the Berlin Academy edition of Kant's *Gesammelte Schriften* (Berlin: de Gruyter, 1900–), using the translations (with some modifications) in the *Cambridge Edition of the Works of Immanuel Kant*. I use the following abbreviations:

G	*Groundwork of the Metaphysics of Morals* (in Kant 1996a).
KpV	*Critique of Practical Reason* (in Kant 1996a).
KU	*Critique of the Power of Judgment* (in Kant 2000).
MS	*The Metaphysics of Morals* (in Kant 1996a).
Rel	*Religion within the Boundaries of Mere Reason* (in Kant 1996b).

capacity to choose *only that* which reason, independently of inclination, recognizes as practically necessary, i.e., as good." In imperfectly rational agents such as ourselves, the will is a capacity to choose what reason recognizes as good, but whose exercise, experience tells us, is not unfailingly directed at objective good, and the determination of the will by objective principles is "necessitation" [*Nöthigung*]. In our case practical reason constrains volition through imperatives. Even so imperatives represent actions as good and, moreover, determine volition by representing actions as "practically good . . . on grounds that are valid for every rational being as such" (*G* 4: 413).[10] In this discussion, the identification of the will with practical reason is based on the idea that volition is a capacity for "deriving actions from laws" (or principles), thereby representing an action as good in some respect. Presumably the will is a complex capacity to move from rational principle to action through an agent's representational capacities, by representing or judging an action to be good.

Likewise, in the *Metaphysics of Morals*, Kant defines the will as "the faculty of desire in accordance with concepts" whose "inner determining ground, hence even what pleases it, lies within the subject's reason" (*MS* 6: 213). Call this a "rational faculty of desire." This passage contains the important distinction between *Wille* (will) and *Willkür* (choice), which are the legislative (or rational) and executive (or causal) aspects of the rational faculty of desire, or will [*Wille*] in a broad sense.[11] The power of choice is the rational faculty of desire "insofar as it is joined with one's consciousness of the ability to bring about its object by action." It is the causal aspect of volition, involving the ability to realize the object or action that one judges to be supported by rational grounds. *Wille* in a narrower sense is the rational faculty of desire considered "in relation to the ground determining choice to action, and has itself properly no determining ground, but is, insofar as it can determine the power of choice, practical reason itself" (*MS* 6: 213). Kant's identification of the will in a narrow sense with practical *reason* indicates that volition has a rational or cognitive aspect, while its identification with *practical* reason references the fact that it is a kind of *faculty of desire*—the capacity in a living being "to be, by means of its representations, the cause of the objects of these representations" (*MS* 6: 213; cf. *KpV* 5: 9n. and *KU* 5: 220). Will is a faculty of desire in which the representations through which the living being guides its causal activity are based in reason. For example, the representations that guide the being to realize their objects are rational principles, or judgments that represent an action or end as good by "deriving" it from a principle. The practical dimension

[10] Cf. *KpV* 5: 61: "What we are to call good must be an object of desire in the judgment of every reasonable human being, and evil an object of aversion in the eyes of everyone; hence for the appraisal of action, reason is needed in addition to sense."

[11] Beck (1960: 199–202) and Allison (1990: 129) have suggested that *Wille* and *Willkür* distinguish the legislative and executive functions of volition. Another important discussion on which I have drawn is Engstrom (2010). Engstrom has argued persuasively that volition has a cognitive dimension and argues that the distinction between *Wille* and *Willkür* tracks the distinction between the cognitive and the causal moments of volition. For further discussion see Reath (2013: 41–8).

of volition is not in dispute, and I am concerned to explore the implications of its rational or cognitive dimension.

That volition involves "deriving action from laws" (or principles) suggests the following picture. Volition is initiated by ends, principles, or values taken to be good (and which presumably are genuinely good in some circumstances or subject to certain limitations).[12] Practical reasoning and judgment takes into account relevant features of one's circumstances and available alternatives, and it moves from its starting point through these features to a representation of an action that will achieve the end or is called for by the principle or value under the circumstances (perhaps through one or more subsidiary ends or principles for which means are required, perhaps directly). Since it is a species of reasoning, it understands itself to *move correctly* or in a *warranted* way from initiating end through circumstances to choice, thus concluding in a representation of the action (or action kind) *as rationally supported*. Obviously the practical reasoning may be quite complex when multiple or competing ends and values that are assigned differing degrees of priority are on the table, some of which may set limiting conditions on the satisfaction of others. Kantian maxims represent actions (or action kinds) as intelligible occurrences with rational support and are intended to capture the practical reasoning and judgment that guides volition. A maxim is a representation of an action or action kind as rationally supported by facts about an agent's ends, principles, circumstances, and so on, and thus represents the action *as good*. (Since the reasoning is general and applies to all relevantly similar cases, the maxim itself is a principle to the effect that given a certain end or value and a certain set of circumstances, one is to φ, or φ-ing is good.) And it is through such a representation of the action as rationally supported or good that the agent guides the deployment of his *causal powers*. That is, the maxim is a representation of an action or end, through which an agent directs his causal powers, and it is efficacious in realizing the represented action through the agent's self-understanding as moving correctly from an initiating end or value to choice, that is, by representing the action as (judging it to be) rationally supported or good—viz., rational agents *act* according to the representation of laws or principles.[13]

Though it addresses a different issue about volition, the so-called "Incorporation Thesis"—that "the power of choice . . . cannot be determined to action through any incentive except so far as the human being has incorporated it into his maxim (has made it a general rule for himself, according to which he wills to conduct himself)" (*Rel* 6: 23–4)—readily fits into the conception of will as the capacity to derive actions

[12] The end or principle that initiates practical reasoning is not its starting point in time, but rather the first premise in its rational reconstruction (which admittedly is often artificial).

[13] For another discussion of Kant's conception of practical reasoning, see Herman (2006: 44–61). As the causal aspect of volition, *Willkür* is the ability to guide one's causal powers to realize the action represented as good, through that very representation. It is volition moving all the way to action. And as a rational faculty, it likewise understands itself to move in a warranted way from that representation, viz. from a maxim to action.

from principles. The point of the Incorporation Thesis is to clarify the freedom of the power of choice, denying that incentives influence choice causally by holding that they do so only through an act of spontaneity on the part of an agent. The act of spontaneity is here characterized as "incorporating" or taking up (*aufnehmen*) an incentive into a maxim, or "making it a general rule" for one's conduct. What the thesis claims is that an incentive of any sort can influence choice only by an agent taking the incentive up into practical reasoning and judgment that can be expressed as a maxim, or general rule for one's conduct. The "incorporation of an incentive into a maxim" could consist of one's taking the incentive up as an end (making its object an end) that is to initiate practical reasoning to action, or of taking the incentive up into practical reasoning from some pre-existing end (happiness, for example, or some more specific end) as a way of specifying or furthering that end. The "general rule" that the agent adopts is a function of the generality of the relevant practical reasoning. As I see it, the central claim here is that an incentive influences choice, not causally, but by being taken up into practical reasoning that leads to a judgment of an action as good or rationally supported, on which the agent acts. This is a thesis about freedom of choice because it addresses the role of incentives in practical reasoning carried all the way to action (the maxim on which an agent *acts*).[14]

The rationalist thesis would hold that the practical reasoning that guides volition constitutively understands itself as part of its self-consciousness to satisfy a condition of universal validity—e.g. to derive or move toward action with sufficient rational support. That is, the thesis is that practical reasoning and judgment aimed at affirming a detachable conclusion about action begins from ends, principles, or values that are taken to be "good . . . in the judgment of every reasonable human being," and that it takes itself to move correctly from there to a representation of an action as good in just that sense. Moreover, the thesis is that it is part of the self-understanding of rational volition that it is initiated by an end or principle that satisfies a condition of universal validity and that it moves correctly or in a warranted way from that starting point to a practical conclusion (a judgment) about the objective goodness of an action. Assuming that the reasoning is captured in the maxim of action, that is to say that it is part of the self-consciousness of rational volition that agents take their maxims to satisfy a condition of universal validity. In this sense volition would tacitly be guided by and would understand itself to be guided by something like the Universal Law version of the categorical imperative. (This is not to say that all rational volition does satisfy a condition of universal validity, but only that the agent tacitly understands it to.)

[14] On one common understanding of the Incorporation Thesis, the act of spontaneity that is essential to the freedom of choice resides in the adoption of a maxim (as a principle representing an action as good), that is, in the decision to act on the maxim. My suggestion here is that the spontaneity referred to is present throughout the derivation of an action from a principle. It is a feature of all moments of volition, including the taking up of the incentive into practical reasoning as well as the movement from judging an action to be good to action. The Incorporation Thesis is a thesis about *Willkür* because it concerns volition that moves all the way to action.

The example of the deposit in the second *Critique* is a case of volition that illustrates features of this structure. The agent has made it his maxim "to increase my wealth by every safe means." The end of increasing one's wealth is good through its contribution to happiness, which is an objectively good end when its pursuit is properly constrained. The fact that the agent has in his hands a deposit whose owner has died without leaving a record presents "a case" for his maxim. Applying the maxim of increasing one's wealth by every safe means to this case leads the agent to ask whether "he can give through his maxim such a law as this: that everyone may deny a deposit which no one can prove has been made" (*KpV* 5: 27). I take it that the agent here applies the initial end to his circumstances (prompted by the owner's death and realization that he can keep the deposit without detection) to formulate what appears to be a second maxim of denying the existence of the deposit as a means to safely increasing his wealth—a maxim that expresses a specific intention to act. That maxim represents the action of denying the deposit as good by representing it as rationally supported by the initial maxim (the end of safely increasing one's wealth by all safe means) in the agent's circumstances. [15] Now of course it is not good, and if the agent acts under, in Engstrom's phrase,[16] the presupposition of universality, the not terribly taxing realization that the secondary maxim fails to satisfy this condition should lead the agent to change his initial judgment and to abandon this intention.[17]

The rationalist thesis about the nature of volition is certainly consistent with Kant's remarks about the will and with the conception of practical reasoning suggested by some of his examples. But how will it deal with a kind of case that is clearly possible (since it commonly occurs) in which an agent decides to keep the deposit even after realizing that the relevant maxim fails of universal validity. Bad choice is one obvious difficulty for this conception of volition.[18] The thesis must hold that bad willing is due

[15] This discussion illustrates an ambiguity in Kant's understanding of maxims: are they general principles or ends that are sometimes called "*Lebensregeln*," or are they more specific principles with an action-end-circumstances schema that can express an intention to act? What Kant initially calls the agent's "maxim" in this passage is a general end or principle ("to increase my wealth by every safe means"), and its application to the case leads to a maxim of the second kind (to deny the existence of the deposit when one can do so without detection as a means to increasing one's wealth), which is a representation of the practical reasoning that leads to a specific action under the circumstances. We need not resolve this question about the nature of maxims, but there is no deep problem here. Presumably principles of the first kind are adopted as having rational support, and to articulate that support, they can be stated in the second form as means to happiness (or some other general end) given standing facts about human life. Happiness is my end and something I take to be good; and given various general facts about life in certain social settings along with facts about what I need to lead a satisfactory life, adopting the principle of increasing my wealth by all safe means is a means to happiness—a valid "counsel of prudence."

[16] See Engstrom (2009: 124–7).

[17] If the initial maxim has the form of or qualifies as a practical law, it is a principle that all can agree that all are to act from and will apply consistently to all cases that fall under it, including this one. But since applying initial the maxim to this case leads to a contradiction in conception, it does not have the form of a law, nor does the subsidiary principle that results from its application to this case.

[18] It is important to distinguish genuine vice, which involves endorsing and acting from bad principles (bad willing), and weakness, which involves knowing failure to live up to principles that one endorses. For an important discussion (aimed at showing how genuine weakness is consistent with the Incorporation Thesis), see Johnson (1988: esp. 358–62). A defense of the strong thesis (both philosophical and as Kant

to either of two kinds of defects in rational commitment or reasoning (or to a combination of both). One is erroneous or untenable assignment of value to the initiating end—for example, the agent takes the initiating end to be objectively good or good without restriction when it is not, or assigns it a priority in the circumstances that cannot be sustained. Or perhaps the agent mistakenly takes his happiness to have greater objective importance than that of others (thereby licensing privileges not extended to other agents). The second kind of error is bad reasoning from the initiating end to action—e.g. the agent might begin from a just assessment of the standing of the end and take the reasoning that it initiates to satisfy a condition of universal validity, when it does not. In the first case, the agent who keeps the deposit might take own happiness to be good without restriction or to have a value that overrides applicable common sense moral norms. Or—the second kind of case—the agent might rightly acknowledge the conditional value of own happiness but "make an exception for herself" in this case and mistakenly judge that the maxim of denying the deposit as a means of furthering her own happiness is endorsable by anyone. (That sounds pretty bloodless, but what precludes defective value commitments or errors in practical reasoning from being "willful" or "motivated"?)

Is this a plausible reading of what goes on in bad choice? Perhaps the following "alternative picture," is more compelling. Practical reasoning (*Wille* in the narrow sense) reveals that denying the deposit fails to meet a condition of universal validity and thus would be wrong. But even an agent who sees this clearly still needs to decide whether or not to comply with moral principle (i.e. to abide by the conclusions of *Wille*), and in this case elects not to, fully aware that the action is *Gesetzwidrig*. Rational volition does require practical reasoning and judgment about what is good or what one has reason to do (the province of *Wille* in the narrow sense). But action also requires choice (*Willkür*), which is the further decision about whether or not to follow the conclusions of practical reason. The conclusions of (objective) practical reason ideally guide choice, but because "the will is not *in itself* completely in conformity with reason," they necessitate or constrain.[19]

This "alternative picture" of bad choice—one that does not ascribe the rationalist thesis to Kant—is committed to a certain understanding of the distinction between *Wille* and *Willkür*. To give it a name, I'll call it the "elective conception" of the will. I think that there are problems with the picture of bad choice that results from the elective conception of the will, and with this conception of the will more generally, and these problems push one towards that rationalist thesis.

interpretation) needs to accommodate both. What I say in these paragraphs (about "bad willing") applies to what Kant calls "true vice" (*MS* 6: 408), and I make some remarks about weakness at the end of section 4.

[19] *G* 4: 413. One is tempted to take "the will" [*der Wille*] in this remark—what is not in itself in complete conformity with reason and is thus subject to rational necessitation—to refer to the causal moment of volition, to what Kant in later works specifies as *Willkür*. But it should refer to all moments of volition. In finite rational agents, any moment of volition—our practical reasoning as well as our implementation of our practical judgments—can deviate from objective norms and thus is subject to rational necessitation.

What are these problems? I've suggested that Kant understands volition in the broad sense to be a complex capacity to move from rational principle to action through one's own representational activity that has both a cognitive dimension involving reasoning and judgment (*Wille* in the narrow sense) and a causal or executive dimension (*Willkür*) that are linked. Practical reasoning concludes in a representation of an action as good, and the causal dimension of willing is the ability to actualize the action thus represented. But does the cognitive dimension of volition do any real work in the alternative picture of bad choice—and if not, is it genuine volition? Assume here that practical reasoning leads the agent to see that there are compelling reasons not to deny the deposit, then (perversely) ignores this conclusion. The judgment of *Wille* then drops out of the picture and does not figure in what he does. Choice is then divorced from practical reasoning and judgment and appears arbitrary. If so, in what sense is bad choice a case of volition? Perhaps the proper analysis is that the agent reasons instrumentally from the end of happiness, which is mistakenly taken to be good in an unrestricted sense, or decides on further reflection that an exception to ordinary moral norms is (after all) warranted in this case. But that story belongs to the rationalist thesis, since it takes the choice to result from defective principle, reasoning, or judgment of *Wille*. It is volition that is initiated by an unsustainable assessment of the standing of the initial end, or leads to choice through unsound practical reasoning.

A general problem with the alternative elective conception of the will is that it detaches choice from practical reasoning—as though once one sees what practical reason judges to be worth doing, one still needs to decide whether or not to do it. (And how would one go about deciding that?) I am inclined to think that the distinction between *Wille* and *Wilkür* at *MS* 6: 213 suggests a different conception of volition that is more coherent. *Wille* in the narrow sense is the cognitive aspect of volition that it is constitutively aimed at determining what there is good and sufficient reason to do, and it takes its exercises to satisfy a condition of universal validity. Choice is the causal aspect of volition whose function is to execute the conclusion of practical reason, e.g. a representation of a specific action as good through which the agent deploys his causal powers. (Until *Wille* has put an action or end on the table by representing it as good, choice has nothing to do.[20])

One might grant that *Wille* in the narrow sense as the cognitive aspect of volition is constitutively guided by a standard of universal validity and that its fundamental principle is indeed the categorical imperative.[21] But surely, one might object, *Willkür*

[20] I develop this reading further (which is indebted to Stephen Engstrom) in Reath (2013: 41–9). Among other things, I suggest there that the internal norm of *Willkür*, understood as the causal or executive dimension of volition, is the hypothetical imperative.

[21] Hence Kant can say that "Laws proceed from the will . . . the will, which is directed to nothing beyond the law itself, cannot be called free or unfree . . . since it is directed immediately to giving laws for the maxims of actions (and is therefore practical reason itself)" (*MS* 6: 226). Presumably this passage implies that objective practical reason (practical reason correctly applied) follows the categorical imperative. In that case one should hold that individual uses of practical reason, while admittedly fallible, understand themselves to satisfy a condition of universal validity.

or choice need not have any such constitutive aim beyond choosing what to do. It has no internal norm of its own, but is a purely elective capacity to follow the conclusions of objective practical reason that an agent does not always exercise. Sometimes it does, but sometimes it does not follow objective practical reason. While it is not disconnected from practical reasoning, it represents a further step beyond simply concluding that an action is supported by good and sufficient reasons.

I find this conception of volition unsatisfactory. If the conclusion of practical reasoning still leaves a further question to be decided—whether to follow that conclusion or not—then choice is disconnected from practical reason in a problematic way. Moreover, in Kant's philosophical system, we should expect a rational capacity to have its own internal norm that makes it what it is, and (I would argue) that tacitly guides all exercises of the faculty—or better that the faculty is self-consciously guided by its own awareness of its internal norms. (Therein lies the spontaneity of its operation.) This would be the positive conception of the faculty, what it is a power to do. The objection grants that *Wille* in the narrow sense is constitutively guided by a standard of universal validity, but denies that choice has any such internal norm. However while choice need not have the constitutive aim of satisfying a condition of universal validity, if it is the causal or executive dimension of volition in the broad sense, it will be charged with and will understand itself to be carrying out the conclusions of practical reason (which is constitutively guided by a principle of universal validity). So its constitutive aim ties it to the internal norms of *Wille*.

Kant's claim that choice cannot be defined as a capacity to choose for or against the moral law is relevant here. Kant tells us that the positive conception of free choice (*die Freiheit der Willkür*) is "the ability of pure reason to be of itself practical . . . through subjecting the maxim of every action to the condition of its qualifying as a universal law" (*MS* 6: 213–14). And although only choice (as opposed to will in the narrow sense) is properly free, "freedom of choice cannot be defined. . . as the ability to make a choice for or against the law (*libertas indifferentiae*), even though choice as a phenomenon provides frequent examples of this in experience . . . [F]reedom can never be located in a rational subject's being able to choose in opposition to his (lawgiving) reason, even though experience proves often enough that this happens." Our experience of how choice is exercised cannot give us "the expository principle [*Erklärungsprinzip*] (of the concept of free choice) and the universal feature for distinguishing it (*from arbitrio bruto s. servo*)" (*MS* 6: 226).

One point to make about this passage is that the positive conception of freedom of choice should be the positive conception of the *causal* dimension of practical reason. (That is, he is not concerned with the positive conception of *Wille* in the narrow sense.) The claim is that the positive conception of free choice is the ability of pure reason to be *practical*—that is, the ability of pure reason to *realize*, through its representations (of actions and ends as good), the objects of those representations independently of conditions of sensibility, or the ability to *act* from pure practical reasoning. But then, freedom of choice does have its own internal norm or constitutive aim,

that of realizing the conclusions of (pure) practical reason, and if that is its positive principle, its exercise does not float free of *Wille* in the narrow sense. A second point concerns how to understand the positive conception of a faculty. I'll suggest in section 3 that the positive conception of a faculty—what Kant here terms its "expository principle" (*Erklärungsprinzip*)—is the internal norm consciousness of which tacitly guides all exercises of the faculty. If so, then all practical reasoning understands itself to be guided by the condition of universal validity (as sketched above) and all rational choice understands itself to be implementing the conclusions or judgment of practical reasoning, or doing what practical reason judges good in this sense. This passage from the *Metaphysics of Morals* does not require this reading of the "positive conception," but it does offer a way of filling out what this claim amounts to.

This section of the chapter has shown that Kant's key discussions of the will are consistent with and point toward the rationalist thesis about the nature of rational volition. However, I don't think that these texts require this interpretation of volition. First, the passages from the *Groundwork* and related passages from the second *Critique* (*KpV* 5: 57–62 ff.) make it clear that Kant thinks that rational volition is guided by representations of actions as good, but such representations could include judgments that an action is good as a means, or that some action or end is good conditionally or on prudential grounds. It is not obvious that such practical judgments make a claim to their own universal validity in the sense needed for the rationalist thesis (in fact, many philosophers think that it is obvious that they do not . . .), and Kant never explicitly makes this claim. Second, the claim that an imperfectly rational will "does not always do something just because it is represented to it that it would be good to do" (*G* 4: 413) and similar passages certainly lend support to the conception of freedom of choice as an elective capacity that operates independently of practical reason in the sense that it often ignores or acts contrary to the conclusions of objective practical reason. This tends to be the default reading of free choice. My aim in this section has been to suggest alternative readings of such passages that are (at least) sufficiently compelling to supplant the elective reading as the default reading. I'll discuss some of these passages in the final section. Finally, the rationalist thesis relies on a certain take of what goes into the "positive conception" of free choice. I think that it is philosophically compelling and will take the point up in section 3, but I grant that it is not the only way to read the texts. In sum, the strong thesis is consistent with and supported by, but not required by key texts about the will.[22]

[22] Chapter 2 of the Analytic of the second *Critique* is one place where one might expect to find support for Kant's acceptance of a strong guise of the good thesis, but this chapter does not seem to me to address the issue clearly one way or the other. First, Kant does say that *good* and *evil* are the "only objects of a practical reason" and that they are "necessary objects" of the faculty of desire or aversion "both however in accordance with a principle of reason." The necessity in question appears to be that what is good "must be an object of the faculty of desire in the judgment of very reasonable human being, and evil an object of aversion in the eyes of everyone"—what is good or evil is what everyone can rationally judge to be so (*KpV* 5: 58, 61). But Kant does not clearly say that volition is necessarily directed at what one judges to be good or that the good is the formal object of rational volition. Second, Kant finds an ambiguity in the scholastic formula *nihil appetimus,*

3 Philosophical Support

In this section I will explore what the rationalist thesis that volition constitutively understands itself to satisfy a condition of universal validity would do for Kant philosophically. I will be focusing on two well-known ideas from the *Groundwork* and second *Critique*—the thesis of autonomy of the will and the claim that "a free will and a will under moral laws are one and the same" (*G* 4: 447). I will suggest, first, that the thesis about rational volition is required by Kant's conception of autonomy, and second, that it gives the best account of Kant's claim that "a free will and a will under moral laws are one and the same" (*G* 4: 447)—or as we might say, that the moral law is the basic principle of free rational agency. Furthermore, it offers a satisfying account of how bad willing can be genuine free volition, given the close connection that Kant draws between free agency and the moral law.

To begin with the first, Kant characterizes autonomy of the will as "the property of the will by which it is a law to itself (independently of any property of the objects of volition)" and formulates that law (that the will is to itself) as the Universal Law version of the categorical imperative (*G* 4: 440, 444, 447). The best reading of the thesis that the will is a law to itself "independently of any property of the objects of volition" is that the nature of rational volition by itself is the source of its own fundamental norm.[23] But what does that mean? First, rational volition must have a "nature" that is sufficiently rich to generate a fundamental practical principle that can guide volition, without the need to turn to any object or interest given to the will from outside. To say that a rational faculty has a nature is to say that it has a formal aim or constitutive principle that can guide its own exercise, and moreover that does tacitly guide all exercises of the faculty. We know that the moral law (categorical imperative) is the law given by its nature ("not to choose in any other way than that the maxims of one's choice are also comprised as universal law in the same willing" (*G* 4: 440)). So unpacking Kant's thesis of autonomy in this way leads directly to the thesis that rational volition constitutively understands itself to satisfy a condition of universal validity, viz. that the Universal Law version of the categorical imperative is the formal principle of all rational volition.[24]

Must this norm tacitly guide *all* instances of rational volition? Late in *Groundwork* II Kant claims:

nisi sub ratione boni . . ., depending on whether *bonum* refers to *das Gute* or *das Wohl*. He thinks the formula "indubitably certain" if *bonum* refers to *das Gute*, in which case it can be rendered: "we will nothing under the direction of reason [*wir wollen nach Anweisung der Vernunft nichts*] except insofar as we hold it to be good or evil" (*KpV* 5: 60). When reason properly directs volition, it is of course *sub ratione boni*. But does this phrasing allow that we can sometimes will though not under the direction of reason, or that choice can ignore reason's judgments? That would be the alternative elective conception of the will. I need Kant to hold that all volition involves the exercise of reason and takes itself to conform to reason's standard, but it seems to me that this passage could be read either way.

[23] I defend the points in this paragraph and the next in more detail in Reath (2013: section III).

[24] Here note also *G* 4: 433, that the will is a "will giving universal law according to its natural end."

An absolutely good will, whose principle must be a categorical imperative, will therefore, indeterminate with regard to all objects, contain merely the form of willing as such [*bloß die Form des Willens überhaupt*], and indeed as autonomy; i.e. the fitness of the maxim of every good will to make itself into a universal law is the sole law that the will of every rational being imposes on itself, without underpinning it with any incentive or interest as its foundation. (*G* 4: 444)

That the principle of a good will "contains merely the form of willing as such" is easily read as the claim that the Formula of Universal Law expresses a formal feature of *any* exercise of the will. That would make it the formal principle that tacitly guides all rational volition. Furthermore, the role that the thesis of autonomy plays in Kant's foundational argument is to support the unconditional authority of the moral law. Kant does that by demonstrating that the moral law is the formal principle of rational volition to which one is committed simply in exercising the will.

Turning now to the second cluster of ideas, early in *Groundwork* III, Kant argues that from the negative explication of free agency as independence from determination by alien causes there flows a positive conception—a positive specification of what free agency qua capacity is that includes a statement of the principle according to which free agency operates.[25] Freedom of the will can only be "autonomy, i.e. the property of the will of being a law to itself," and *Groundwork* II has argued that the Universal Law version of the categorical imperative is the law that the will, as practical reason, gives to itself. More precisely: "the proposition: the will is in all actions a law to itself, designates only the principle of acting on no maxim other that that which can also have itself as object as a universal law." Thus "a free will and a will under moral laws are one and the same" (*G* 4: 447)—and by that I take it Kant means that the basic principle of free agency is that of acting from maxims that can be willed as universal law or that satisfy a condition of universal validity. Likewise in the second *Critique* Kant argues that a will for which "the mere law-giving form of maxims alone is the sufficient determining ground" is a free will (*KpV* 5: 28–9).[26]

Setting aside the details of the argument, how should we understand the claim that "a free will and a will under moral laws are one and the same"? A standard reading of this claim finds two general points in it. The first is a conception of free agency: free agency is the capacity to determine choice by moral principles, and free and imputable actions are those chosen by a free agent in whom this capacity is unimpeded or undiminished, whether or not it is fully exercised. The second general point is to unpack the claim that

[25] Here compare also *KpV* 5: 105–6, where Kant holds that the reality of free agency is established by identifying the principle of its operation. This is "an objective principle of causality . . . in which reason . . . already itself contains this determining ground by that principle, and in which it is as pure reason itself practical." Thus freedom is "not merely thought indeterminately and problematically . . . but is even *determined with respect to the law* of its causality and *cognized* assertorically . . . [O]ur reason itself, by means of the supreme and unconditional law cognizes itself . . . and indeed even with a determination of the way in which, as such, it can be active." These passages are naturally read as implying that all exercises of free agency are in some sense guided by the moral law (or in our case the categorical imperative).

[26] I assume that the will in these passages is the faculty of rational volition in the broad sense (including both *Wille* in the narrow sense and *Willkür*).

the moral law is the basic principle of free agency in terms of the basic commitments of free rational agents.[27] The Universal Law version of the categorical imperative is reason's own principle, and action from that principle is free. And since it is a necessary component of the self-consciousness of rational agency that one acts under the idea of freedom and identifies with this capacity (the capacity that makes one an agent), one is rationally committed to acting on the principle by which it is defined. When one does, one fully realizes the capacity, and that is a good for the agent. (It is the good that the opening of *Groundwork* I presents as unconditional and without limitation, now understood in a way that shows how it can be a good for a finite rational agent.)

According to this conception, a free agent has the capacity to judge what is required by and to act from universally valid principles, and is committed to acting on such principles by one's self-conception as free. When one does act from maxims that satisfy this standard, one realizes one's free agency and the good that it represents. But sometimes free agents act badly—on maxims or practical judgments that fail to satisfy this normative standard, indeed sometimes knowing full well that one's maxim falls short of accepted moral standards. In such cases, though practical reason shows that one's maxim is not universally valid, the agent makes the elective choice to act on it all the same. Here the agent fails to exercise the capacity to act from universally valid principles, and in this respect bad choice represents a failure to exercise one's free agency. But

[27] Hill (1992) has developed one of the more interesting reconstructions of the argument that a free agent is committed to the fundamental moral principle, which has not been given sufficient recognition in the literature. Hill takes the argument to show that if one grants (plausibly) that the negative conception of freedom includes a capacity for desire-independent motivation, any free rational agent is committed to "some principle or principles which are rational and yet not hypothetical imperatives," and thus must acknowledge some rational principle beyond those recognized by instrumentalists such as Hobbes and Hume. Practical reason in free rational agents is not simply instrumental (112). His reconstruction of Kant's argument from negative freedom to autonomy goes roughly as follows (111–12):

> A will with autonomy is committed to at least one practical principle that does not simply prescribe the means to the agent's desired ends and is "one's own" in the sense that it is deeply rooted in one's rational nature.
>
> A negatively free will can act for reasons that are not based on desires or hypothetical imperatives.
>
> Since a lawless will is absurd, such agents must act from some principle (when they act on non-desire-based reasons), thus must be committed to or accept some rational principle that is not a hypothetical imperative.
>
> The previous step rules out various substantive principles that are clearly hypothetical imperatives (such as desire-based principles, principles adopted because accepted by some external authority or convention, etc.). It also appears to rule out the principle of assigning prima facie weight as reasons to informed preferences that have survived critical reflection. (115)
>
> The only remaining candidate principles are those "which reflect some necessary features of rational agency independently of its special contexts." (112)
>
> Since such principles satisfy the definition of autonomy in Step (1), an agent who is negatively free has autonomy.

This argument is "non-constructive" in the sense that it does not clearly identify the principle(s) to which free agents are rationally committed, but Hill thinks that it points towards the principle that any rational agent is committed to valuing his or her practical rational capacities, which is the basis of the Formula of Humanity.

the action is free and imputable because chosen by a free agent in whom the capacity is undiminished.

While there is nothing decisively unsatisfactory with this conception, here are two worries that one might raise. The first, due to its reliance on the elective conception of the will mentioned in section 2, is whether it can count bad choice as a genuine exercise of volition, thus as genuinely bad willing. The issue here is whether a failure to exercise the capacity for volition counts as genuinely bad willing. This conception appears to treat bad choice as an elective choice that is divorced from, or at least sets aside, practical reasoning and judgment, and that makes it seem not just perverse or weak, but arbitrary. In good willing, choice follows the judgment of practical reason, but what happens in bad willing? Presumably the agent chooses an end that initiates action, but not through any representation of it as good. (And if not, what grounds the choice of the end?) Since bad willing so conceived does not engage the capacity to make judgments of good, it represents a volitional failure—a failure to exercise the capacity for rational volition. But then bad choice is not really volition—though it is still imputable because it is the act of an agent with a capacity for rational volition that he has failed to exercise.

Second, it is part of this conception that when one does act from maxims that genuinely satisfy the normative standard of universal validity, one "realizes one's nature" as free and rational. It is certainly an asset of a moral conception to be able to show that acting from moral principles secures a "higher good" of self-realization for the agent. But the *Groundwork*, at any rate, connects freedom and morality to establish the unconditional authority of morality. This conception appears to locate the reasons to give deliberative priority to the moral in the good of the complete realization of one's nature. But can the good of self-realization ground the conclusion that morality is required of us? Fully realizing one's nature is a good, but must one care about it above all other goods?[28]

I am proposing a different way to understand the claim that the moral law is the basic principle of free agency. I read it as the claim that the Universal Law version of the categorical imperative is the formal principle of free rational agency, to be unpacked through the rationalist thesis that rational volition constitutively understands itself to satisfy a condition of universal validity. As we've seen, this thesis holds that volition constitutively understands itself to move correctly from ends or principles taken to be good to judgment of an action as good—that is, it has the formal aim of acting on maxims that satisfy a condition of universal validity. To say that rational volition *constitutively* understands itself to satisfy a condition of universal validity is to say that this self-consciousness is a necessary feature that guides all volition; expressed as a principle, it is the Formula of Universal Law (FUL). The will "imposes" this principle on itself

[28] Several years ago Henry Allison pointed out to me that many interpretations of Kant's attempt to ground the authority of morality in free agency make it, wrongly in his view, an ethic of self-realization or perfection. See Allison (1996: 117–18).

through the self-consciousness that guides volition, thus is normatively committed to complying with this principle by its own self-understanding.[29]

The proposed conception makes volition an act of the (rational) faculty of desire that understands itself to satisfy a condition of universal validity. This mark applies as much to bad as to good willing, and that makes bad willing free, a genuine exercise of volition on all fours with good willing. What makes it bad is that it fails to satisfy the condition of universal validity that is the internal standard of volition because of some defect in the rational commitments or practical reasoning that guides volition. (For example, the agent takes a bad end or principle to be universally valid, or assigns unrestricted value to a conditionally good end, or reasons badly from the initiating end or principle, and so on, through perversity or weakness, and so on.)

The thesis that the FUL is the formal principle of free agency (that volition constitutively understands itself to satisfy a condition of universal validity) needs some qualification. The claim is not that all rational volition does satisfy a condition of universal validity, but rather that it understands itself to satisfy this condition and proceeds under this self-conception. Thus as I frame the thesis, it allows the formal norm of universal validity to come apart from genuine moral principles in an agent's practice—that is, it is not claiming that agents necessarily take their maxims to have moral justification or to accord with commonly accepted moral principles. An agent who denies the authority or universal validity of ordinary moral norms (e.g. by taking himself to be exempt from them) can still take his maxims to satisfy an indeterminate condition of universal validity. Thus, the formal principle can be part of the self-consciousness of an agent with a very attenuated moral sense. In this respect, my reading departs from Kant's wording at *Groundwork* 4: 447: the thesis is not that the moral law, properly speaking, but that the principle of universal validity is the formal principle of free volition (where an agent's understanding of universal validity might fall short of moral justifiability).

Finally, by holding that the mark of free volition is that it *understands itself* to satisfy a condition of universal validity, one avoids the objection that bad willing does not really count as willing. If, to count as volition, an act of the faculty of desire must be guided by in the sense of *satisfying* the condition of universal validity, action on maxims that are not fully universalizable would not represent genuine free volition, thus would not count as bad willing. But the idea that volition constitutively understands itself to satisfy a condition of universal validity implies that the essential mark of volition is not that it does, but that it *takes itself* to satisfy a condition of universal validity. What makes an act of the faculty of desire an act of rational volition is this self-conception. But an act could be guided by this self-consciousness and fail to satisfy this normative standard. In that case, it is still volition—genuine free, though bad, willing.

[29] For elaboration of this last point, see Reath (2013: 47–9).

What one wants here is a sense of being guided by a principle that allows for genuine violations, or failures to conform the principle. I take that form of normative guidance to be supplied by the self-understanding or self-consciousness of a kind of cognitive or rational activity. The thought is that what makes an act the kind of rational activity it is—in this case what makes an act of the faculty of desire an act of rational volition—is that it understands itself to have a certain formal aim. This self-consciousness normatively guides the activity and gives the norms internal to that activity a grip on the agent: through the self-consciousness that one is engaging in a certain kind of rational activity, one takes oneself to be following and correctly applying the relevant internal norms (in this case, that one is acting from maxims that satisfy a standard of universal validity). But one can take oneself to be guided by a set of norms and still violate or fail to conform to them.

As a final note, this reading of the idea that the FUL is the formal principle of free agency includes the conception set aside earlier in this section and does full justice to Kant's claim that a free will is "subject to moral laws" (*unter sittlichen Gesetzen*). A faculty of desire that is constitutively guided by the condition of universal validity certainly has the capacity to act from genuine moral principles and fully realizes its nature when it does. Further, it is subject to genuine moral principles in the sense that it is committed to accepting the authority of morality by its own self-consciousness. In willing it understands itself to be acting from maxims that satisfy a condition of universal validity, it imposes this requirement on itself through this self-conception, and it is thus in self-contradiction when its maxims fail to satisfy the standard of universal validity.

I'll conclude this section by contrasting the main features of my own proposal with the elective conception of volition that I want to set aside. According to the elective conception, *Willkür* is a capacity for choice that is independent of and floats free of *Wille* (in the narrow sense) in that it involves the further decision whether or not to follow the judgment of practical reason. *Willkür* is the capacity to act from universalizable maxims, but (though it ought to follow pure practical reason) it has no constitutive aim or formal principle and thus does not necessarily operate *sub ratione boni*. Its freedom is its capacity to act from genuine moral principles, which an agent sometimes, but does not always exercise. According to the elective conception, the essential elements of Kant's idea that the moral law is the basic principle of free agency are that free agency is the capacity to act from moral principles, that free agents are rationally committed to acting from the moral law, and perhaps that doing so secures the good of self-realization. One problem with this conception is that it is not clear that it can count bad choice as genuine willing, rather than simply a failure to exercise the capacity for rational volition. (Still it is free choice and imputable because it is an act of an agent with the capacity for free choice.)

The contrasting conception that I wish to defend takes *Willkür* to be the causal moment of a complex faculty of volition whose constitutive aim—to realize what practical reason represents as good—links it to the constitutive aim of *Wille*. *Wille* (the

cognitive aspect of volition) understands itself to move correctly from principles taken to satisfy a conception of universal validity to representations of actions as good, and *Willkür* is constitutively aimed at realizing what practical reason represents as good. Rational volition as a whole thus proceeds *sub ratione boni* in the sense that it constitutively understands itself to satisfy a condition of universal validity. I read Kant's claim that the moral law is the basic principle of free agency as the claim that the Universal Law version of the categorical imperative is the formal principle of volition that tacitly guides all willing, as I've explained that thesis above. I've argued that this conception is required by Kant's doctrine that the will is a law to itself (the central plank of which is that the nature of the will is the source of its own formal principle). Further, the claim that FUL is the formal principle that tacitly guides all volition provides a natural reading of Kant's idea that the moral law is the basic principle of free agency. One clear strength of this reading is that it makes bad willing genuine free volition—free volition because it understands itself to satisfy a condition of universal validity, but bad because of defects in fundamental principles and value commitments and in reasoning. That is to say that bad willing is not just a failure to exercise the capacity distinctive of volition, but a misuse of that capacity.

That the rationalist thesis connects in these ways with these central Kantian ideas provides powerful (even if not decisive) reasons to think that Kant is committed to it. Kant thinks that it is a necessary feature of practical self-consciousness that rational agents act under the Idea of Freedom. If the Universal Law version of the categorical imperative is indeed the formal principle of rational agency, then acting *sub ratione boni* in this sense is just a different aspect of this same Idea.

4 Bad Willing?

In this section, I consider briefly how the rationalist thesis that FUL is the formal principle of rational volition fits the phenomenon of bad willing. Does the thesis fit what Kant says about bad willing? And in making a case for ascribing this thesis to Kant, am I saddling him with a highly implausible conception of volition and action?

Kant clearly thinks that we can knowingly and intentionally act contrary to moral principle. To cite a few instances:[30] a finite will "does not always do something just because it is represented to it as something that it would be good to do," which presumably implies that such agents sometimes do things that they know to be bad or contrary to reason (*G* 4: 413). The scoundrel is conscious of the good will in him that "constitutes the law, the repute [or authority] of which he recognizes as he transgresses it [*dessen Ansehen er kennet, indem er es übertritt*]" (*G* 4: 455). Human beings are "unholy enough

[30] Thanks to Jens Timmermann for these passages and for discussion of this issue generally. The first three passages suggest cases of weakness, but I think that all can be read either as weakness or as "perversity" or "true vice" (bad willing). Bad (rather than weak) willing is easier for my account to handle, but I'll comment briefly on weakness at the end of this section.

that pleasure can induce them to break the moral law, even though they recognize its authority" (*MS* 6: 379). "Now through experience we can indeed notice unlawful actions, and also notice (at least within ourselves) that they are consciously contrary to law [*das sie mit Bewußtein gesetztwidrig sind*]" (*Rel* 6: 20).

The rationalist thesis about the will does not preclude the possibility of action that is knowingly contrary to moral principle. As I said in section 3, it does not claim that rational volition necessarily takes itself to satisfy standards of moral justification. In general, the thesis traces bad willing back to some part of the cognitive aspect of volition, and here we've mentioned two general possibilities. One is that practical reasoning and judgment are initiated by defective or unsustainable principles or value commitments that are taken to satisfy a condition of universal validity. For example, an agent might reason from an unrestricted principle of own happiness that is (wrongly) taken to be universally valid. Here the agent reasons from a non- or counter-moral principle and does not accept the deliberative priority of morality. A second is bad or defective reasoning from acceptable principles that is incorrectly taken to lead to a universally valid conclusion about action—e.g. as when an agent who professes acceptance of the conditional value of happiness (wrongly) ignores those conditions in his reasoning to action. While it is perhaps clearest in the first kind of case, an agent who consciously rejects the deliberative priority of morality can take his or her practical reasoning or maxim to satisfy a condition of universal validity, while knowing that it does not conform to accepted moral principles. What this shows is that the condition of universal validity can be understood by an agent in such a way that it falls short of genuine moral justifiability, so that the formal principle of universal validity can come apart from morality in the agent's reasoning. This allows for genuine volition that is consciously contrary to morality—'*mit Bewußtsein gesetzwidrig*'.

Kant's discussion of self-love and self-conceit in the second *Critique* provides a model for tracing bad willing to defective practical principles that are taken to satisfy a condition of universal validity. Indeed, "self-conceit" appears to elevate a counter-moral principle to the status of "law." To cite a well-known passage:

> Now, however, we find our nature as sensible beings so constituted that the matter of the faculty of desire . . . first forces itself upon us, and we find our pathologically determinable self, even though it is unfit to give universal laws through its maxims, nevertheless striving to make its claims primary and originally valid, just as if it constituted our entire self. This propensity to make oneself, on subjective determining grounds of choice [*nach subjectiven Bestimmungsgründen seiner Willkür*] into the objective determining ground of the will [*des Willen*] in general can be called self-love; and if *self-love* makes itself lawgiving and the unconditional practical principle, it can be called *self-conceit*. (*KpV* 5: 74)

I take the "pathologically determinable self" to be the (rational) self as susceptible to the influence of sensible inclination. It experiences sensible inclinations as incentives (including the natural inclination to distinguish oneself in comparison with and to be well regarded by others; cf. *MS* 6: 465). But as a rational subject it experiences these

incentives by taking them up into practical self-consciousness in a way that makes them available for use in practical reasoning and judgment. Desires that originate independently of one's rational capacities are taken up in rational form, as potential ends or principles from which one can reason and that can be the basis of "claims." Simplifying, in self-love an agent (wrongly) "claims" an unrestricted value for own happiness so that it is the basis of sufficient reasons for action, while in self-conceit an agent (wrongly) claims for one's own person an unconditional standing not reciprocally accorded to others.

This passage makes it clear, first, that these forms of bad willing rest on defective fundamental principles and value commitments. What is corrupted, as it were, are the basic premises of practical reasoning, in this case due to a defective conception of the self: the "pathologically determinable self" strives to make the claims of self-love and self-conceit valid "just as if it constituted our entire self." Further, the claims of self-love and self-conceit are treated as universally valid. Though the pathologically determined self is "unfit to give universal laws through its maxims," it puts its claims forward as "primary and originally valid," as "objective determining grounds of the will" or as "law-giving." These remarks suggest that it is the nature of the rational self, even as sensibly affected, to operate through principles that are understood to make a claim to universal validity. Finally, if contrary to fact the sensibly affected self were the entire self—if the self were exhausted by incentives of self-concern (*Selbstsucht*)—then by default the principles underlying self-love and self-conceit would be universally valid principles. But this conception of universal validity would support only a weak notion of justification that falls short of moral justifiability, because these principles cannot ground genuinely sharable or reciprocally recognized normative claims.[31]

The thesis that rational volition constitutively understands itself to satisfy a condition of universal validity does not locate bad willing (willing on bad principles) in an act of choice (*Willkür*) that ignores or proceeds independently of the judgments of practical reasoning. Rather, it traces it back either to defective rational principles and value commitments or to bad reasoning from (possibly innocent) principles, that, as exercises of practical reason, understand themselves to satisfy a condition of universal validity. This conception has strong interpretive support from Kant's discussion of self-love and self-conceit (as well as the discussion of "impurity" and "depravity" or "perversity" in the *Religion* (*Rel* 6: 30, 36)). Further, the thesis can accommodate bad willing that is knowingly contrary to morality because volition can be guided by non- or counter-moral principles adopted on the supposition of their universal validity, in full awareness that they violate common moral principles. (As I have said, the formal standard of universal validity can be understood in a way that falls short of moral justifiability.)

[31] For further discussion, see the Appendix of ch. 1 of Reath (2006).

A complete account must also show that this conception of volition can accommodate weakness—volition that fails to conform to principles that the agent actually accepts. Indeed weakness may pose the greater challenge, and I don't have the space to address it adequately here. But the preferred strategy will be to locate weakness in the cognitive dimension of volition—in the fundamental principles or reasoning and judgment that lead to action, rather than in a simple failure of choice (*Willkür*) to follow practical reason (*Wille*). One possibility is to trace weakness to unclarity in the content of an agent's principles—for example, if an agent is unclear about what a principle demands in some situation, or understands it so as to permit exceptions in its application that it does not strictly license.[32] Another is to trace weakness to defective reasoning from or application of principles that are otherwise sound. Such cases would be genuine but weak willing.

A back-up possibility for handling certain kinds of weakness is to treat them as intentional action that is not the result of volition. In this respect, the thesis is less strong than it might initially appear since it need not imply that all intentional behavior understands itself to satisfy a condition of universal validity. Such choices would be intentional in the sense that they issue from a faculty of desire—a capacity by means of one's representations to realize the objects of these representations. But they would not be rationally willed since they do not engage the distinctive capacity for rational volition, which initiates action from a principle understood to satisfy a condition of universal validity.[33] (Perhaps the action in this case would be the work of a sub-agential system in an agent of which the agent is conscious, and which the agent has the capacity but fails to control.) Instances of weakness understood in this way would not, strictly speaking, be "weak willing," but volitional failure—failures to exercise the will. But the resulting action can still be regarded as free and imputable, since it is performed by an agent who possesses but fails to exercise the relevant rational capacities.

One advantage of the rationalist conception of volition that I have been trying to articulate on Kant's behalf is that it counts bad or perverse willing as genuine volition—not just a failure to exercise the capacity for volition, but a misuse of it. Above I suggested that the alternative elective conception has to view bad (and weak) willing across the board as volitional failure—as based on an arbitrary act of choice that, because it does not engage the capacity for practical judgment, is not genuine volition but a failure to exercise the will. Bad action would still be imputable because it is the choice of an agent with the capacity for rational volition that he has failed to exercise. In resorting to this model of imputability for some instances of weakness, the advantage of my account runs out at this point. But since my conception uses this model only

[32] Here see Hill (2012: section 5). Hill suggests a Kantian account of weakness that locates it either in the content of a person's principle or in how resolutely a person wills a principle. Both are ways of locating weakness in a person's *Wille* rather than *Willkür*.

[33] Engstrom seems to allow for the possibility of intention that falls short of volition—for instance in practical (efficacious) thought whose efficacy is not based on a judgment of good. See Engstrom (2009: 44 ff.).

in certain cases of weakness, rather than using it across the board for all bad and weak willing, the advantage is retained.

Obviously more needs to be said about both perverse and weak willing than I have room for here, and I do not underestimate the challenges that these phenomena pose for the conception of volition that I have tried to articulate. What I hope to have done in this chapter is to lay out a case for ascribing this conception to Kant that focuses on his central doctrines of autonomy and the connection between morality and free agency.

References

Allison, Henry A. 1990. *Kant's Theory of Freedom*. Cambridge/New York: Cambridge University Press.

Allison, Henry A. 1996. "Kant on Freedom: A Reply to my Critics," in Allison, *Idealism and Freedom*. Cambridge/New York: Cambridge University Press.

Beck, Lewis White. 1960. *A Commentary on Kant's Critique of Practical Reason*. Chicago: University of Chicago Press.

Engstrom, Stephen. 2009. *The Form of Practical Knowledge*. Cambridge, MA: Harvard University Press.

Engstrom, Stephen. 2010. "Reason, Desire and the Will," in Lara Denis, ed., *Kant's Metaphysics of Morals: A Critical Guide.* Cambridge/New York: Cambridge University Press.

Herman, Barbara. 2006. "Reasoning to Obligation." *Inquiry*, 49(1): 44–61.

Herman, Barbara. 2007. *Moral Literacy*. Cambridge, MA: Harvard University Press.

Hill, Thomas E., Jr. 1992. "Kant's Argument for the Rationality of Moral Conduct," in Hill, *Dignity and Practical Reason*. Ithaca, NY: Cornell University Press.

Hill, Thomas E., Jr. 2002. "Personal Values and Setting Oneself Ends," in Hill, *Human Welfare and Moral Worth*. Oxford/New York: Oxford University Press.

Hill, Thomas E., Jr. 2012. "Kant on Weakness of Will," in Hill, *Virtue, Rules, and Justice: Kantian Aspirations*. Oxford/New York: Oxford University Press.

Johnson, Robert. 1988. "Weakness Incorporated." *History of Philosophy Quarterly*, 15(3): 349–67.

Kant, Immanuel. 1996a. *Kant: Practical Philosophy*, trans. and ed. Mary J. Gregor. Cambridge/New York: Cambridge University Press.

Kant, Immanuel. 1996b. *Religion and Rational Theology*, trans. Allen W. Wood and George di Giovanni. Cambridge/New York: Cambridge University Press.

Kant, Immanuel. 2000. *Critique of the Power of Judgment*, trans. Paul Guyer and Eric Matthews. Cambridge/New York: Cambridge University Press.

Korsgaard, Christine M. 1996a. *Creating the Kingdom of Ends*. Cambridge/New York: Cambridge University Press.

Korsgaard, Christine M. 1996b. *The Sources of Normativity*. Cambridge/New York: Cambridge University Press.

Korsgaard, Christine M. 2009. *Self-Constitution: Agency, Identity and Integrity.* Oxford/New York: Oxford University Press, 2009.

Reath, Andrews. 2006. *Agency and Autonomy in Kant's Moral Theory*. Oxford/New York: Oxford University Press.

Reath, Andrews. 2010. "Formal Principles and the Form of a Law," in Andrews Reath and Jens Timmermann, eds., *Kant's Critique of Practical Reason: A Critical Guide*. Cambridge/New York: Cambridge University Press.

Reath, Andrews. 2013. "Kant's Conception of Autonomy of the Will," in Oliver Sensen, ed., *Kant on Moral Autonomy*. Cambridge/New York: Cambridge University Press.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.