American Dialectic, Vol. 1, No. 1, 2011 (137-166)

Kant's Early Ethics

Michael Rohlf

The *Groundwork of the Metaphysics of Morals* is by far Kant's best known work in moral philosophy and has attracted the most scholarly attention, followed at some distance by the *Critique of Practical Reason*. By themselves, however, these works do not present a complete picture of Kant's ethical thought, as recent scholarship on Kant's later ethical writings is beginning to reflect. But Kant's ethical thought before the *Groundwork* continues to receive little scholarly attention.[1] This is not altogether surprising, since Kant did not publish any works primarily about moral philosophy before the *Groundwork*, which was published near his sixty-first birthday in 1785; and there are only a few scattered comments related to moral philosophy in Kant's earlier works, all of which are devoted primarily to other topics. We do, however, have many unpublished notes that Kant wrote as early as the mid-1760's, in the form either of loose sheets or comments he wrote in books, which Kant often had interleaved with blank sheets so that he had

[1] Some notable exceptions include older studies such as Paul A. Schilpp, *Kant's Pre-Critical Ethics* (Evanston, Illinois: Northwestern University Press, 1938); Josef Schmucker, *Die Ursprünge der Ethik Kants* (Meisenheim: Verlag Anton Hain KG, 1961); Dieter Henrich, "Über Kants früheste Ethik," *Kant-Studien* 54:4 (1963), 404-31; and Keith Ward, *The Development of Kant's Views of Ethics* (Oxford: Basil Blackwell, 1972); as well as newer studies such as Richard L. Velkley, *Freedom and the End of Reason: On the Moral Foundation of Kant's Critical Philosophy* (Chicago and London: University of Chicago Press, 1989); Paul Guyer, "Freedom as the Inner Value of the World" in his *Kant on Freedom, Law, and Happiness* (Cambridge and New York: Cambridge University Press, 2000), chapter three; Tom Rockmore, ed., *New Essays on the Precritical Kant* (Amherst, New York: Humanity Books, 2001); and Susan Meld Shell, *Kant and the Limits of Autonomy* (Cambridge and London: Harvard University Press, 2009).

more room for comments. These notes, together with the scattered remarks in Kant's early published writings, reveal a great deal about the development of his views on moral philosophy before the *Groundwork*.[2]

I. The Prize Essay

The first significant indications of Kant's early views on moral philosophy appear in an essay he submitted in 1762 to the Prussian Royal Academy for a prize competition. It is often called the "Prize Essay" because Kant was awarded second prize for his submission, behind that of Moses Mendelssohn, who won first prize. (Both essays were finally published in 1764).

In the final section of the Prize Essay, Kant emphasizes that moral philosophy is still in its infancy:

> It is clear [...] that, although it must be possible to attain the highest degree of philosophical certainty in the fundamental principles of morality, nonetheless the ultimate fundamental concepts of obligation need first of all to be determined more reliably. And in this respect, [...] it has yet to be determined whether it is merely the faculty of cognition, or whether it is feeling (the first inner ground of the faculty of desire) which decides its first principles. (2:300)[3]

[2] Another source for Kant's early ethical thought that is not addressed here are student notes from Kant's early lectures on ethics. See volume 27 of the *Akademie* edition (cited in the next note) and *Immanuel Kant: Lectures on Ethics*, edited by Peter Heath and J. B. Schneewind (Cambridge and New York: Cambridge University Press, 1997).

[3] References to Kant's works cite volume and page number in *Kants gesammelte Schriften*, edited by the Royal Prussian (later German) Academy of Sciences (Berlin: Georg Reimer, later Walter de Gruyter, 1900-); except for references to the *Critique of Pure Reason*, which cite page numbers from both the first (A) and second (B) editions. English translations of Kant's early published works are taken from *Immanuel Kant: Theoretical Philosophy, 1755-1770*, edited by David Walford in collaboration with Ralf Meerbote (Cambridge and New York: Cambridge University Press, 1992). English translations from the *Critique of Pure Reason* are taken from *Immanuel Kant: Critique of Pure Reason*, edited by Paul Guyer and Allen W. Wood (Cambridge and New York: Cambridge University Press, 1997). English translations of Kant's unpublished notes included in

Despite this admission of ignorance, however, Kant makes a number of bold claims about the concept of obligation in this section, some of which anticipate his later views in certain respects. First, Kant already associates the "ought" of obligation with practical necessity or "a necessity of action," and he already distinguishes between two types of practical necessity that resemble in some respects what he later calls categorical and hypothetical imperatives, although here Kant uses different terminology.[4] What Kant later calls the hypothetical imperative is here called "the necessity of means," which says that "I ought to do something (as a *means*) if I want something else (as an *end*)" (2:298). But this is not the necessity involved in moral obligation, Kant says, because the necessity of means is conditional upon my having some optional end, while moral obligation must "command the action as being immediately necessary and not conditional upon some end" (2:298-99). In other words, the moral ought is nonoptional or, in Kant's later terminology, categorical. In the Prize Essay, Kant calls the type of necessity involved in moral obligation "the necessity of the ends,"

Immanuel Kant: Notes and Fragments, edited by Paul Guyer (Cambridge and New York: Cambridge University Press, 2005) are taken from that volume; all other translations are by Michael Rohlf. References to Kant's unpublished remarks in the *Observations* cite the *Akademie* edition and, parenthetically, the number assigned to each remark in *Notes and Fragments* (which does not include *Akademie* references). References to unpublished *Reflexionen* (reflections) cite the numbers (preceded by R) assigned to them in volume 19 of the *Akademie* edition. References to *Reflexionen* that are quoted also cite the Akademie edition, and the conjectural dates assigned to them by Erich Adickes in the early twentieth century are indicated in footnotes.

[4] Kant continued to develop this distinction in unpublished notes written soon after the Prize Essay. In his 1764-65 remarks in the *Observations* (see section III below) he distinguishes between the categorical and conditional goodness or necessity of actions. See 20:149-50 (39) and 20:155-56 (42). The same distinction appears in *Reflexionen* from the 1760's and early 1770's, in some of which Kant uses the terms "hypothetical" and "imperative." For example, see R 6639, 6659, and 6725.

which he interprets as commanding that "I *ought immediately* to do something else (as an *end*) and make it actual" (2:298).

Second, Kant claims that "such an immediate supreme rule of all obligation must be absolutely indemonstrable" (2:299). We can demonstrate that an action is necessary only as a means to some end. But since moral obligations do not command us to perform actions as means to some other end, "it is impossible, by contemplating a thing or a concept of any kind whatever, to recognize or infer what one ought to do." However, although contemplation or the faculty of cognition alone does not tell us what specific moral obligations we have, Kant claims that "there is an unanalysable feeling of the good," which is "an immediate effect of the consciousness of the feeling of pleasure combined with the representation of the object." The role of the intellect in moral judgment is then two-fold. On the one hand, it enables us "to analyze and render distinct the compound and confused concept of the good by showing how it arises from simpler feelings of the good." In other words, our general concept of obligation arises from combining simple feelings of the good in a confused way, and we can show that we have specific obligations by analyzing this general but confused concept into the simpler feelings that underlie it. So there can be no *purely* intellectual demonstration that an action is obligatory, but nevertheless we can *apply* the intellect to the analysis of feeling in order to arrive at specific moral obligations. On the other hand, Kant apparently believes that pure intellect is the source of a formal principle of obligation that is perfectionist:

> The rule: perform the most perfect action in your power, is the *formal ground* of all obligation *to act*. Likewise, the proposition: abstain from doing that which will hinder the realization of the greatest possible perfection, is the first *formal ground* of the duty to *abstain from acting*. (2:299)

By itself, Kant claims, this formal perfectionist principle, for which Kant offers no argument, does not lead to any specific obligations. To yield specific obligations, this formal principle must be combined with the material principles that result from analyzing feelings of the good. Kant does not explain how these two types of principles should be combined or why the perfectionist principle is required as a supplement to the analysis of feeling.

Thus Kant professes uncertainty about whether the faculty of cognition or feeling decides the first principles of moral philosophy, but in fact he seems to hold that both play an essential role. Although he refers obliquely to the Irish moral sense theorist Francis Hutcheson, as having "provided us with a starting point from which to develop some excellent observations" (2:300), Kant's position at this stage is no closer to moral sense theory than to the rationalism of his German predecessors. Nothing in the Prize Essay suggests that the feeling of the good derives from a separate moral sense, rather than from our ordinary capacity for pleasure; and Kant maintains that analysis of this feeling yields specific obligations only when combined with a formal principle of perfection, which itself apparently does not derive from analyzing feeling.

II. Negative Magnitudes

Shortly after writing the Prize Essay, Kant wrote an essay on Negative Magnitudes, which was published in 1763. This essay also contains a few comments related to moral philosophy, all of which are consistent with those in the Prize Essay, but some of which add new information that again anticipates Kant's later views.

In the essay on Negative Magnitudes, Kant does not repeat the more precise account from the Prize Essay of the relation between intellect and feeling in moral obligation; but what he does

say is consistent with the account in the Prize Essay, although he uses slightly different terminology. In section two of Negative Magnitudes, Kant claims that both reason and moral feeling are required for virtue. Both virtue and vice are possible, Kant says, only "insofar as a being has within him an inner law (either simply conscience or consciousness of a positive law as well) [...]. This inner law is a positive reason for a good action" (2:182). That Kant still regards this inner law as a principle of perfection is suggested by his remark that omissions of moral actions are "instances of a lack of greater moral perfection," and that when someone omits to perform a moral action "[w]hat is missing is a certain more powerful ground of perfection" (2:184). That reason is required to grasp this inner law of perfection is clear from Kant's claim that "[a]n animal lacking reason does not practice any virtue. But this omission is not a vice (*demeritum*), for the animal has not contravened any inner law. It was not driven by inner moral feeling to a good action" (2:183). This last sentence suggests that, although reason is required to grasp the inner law of perfection, moral feeling is nevertheless what drives us to act virtuously. But since nonrational animals cannot be driven to act from moral feeling, reason is evidently required as well for virtuous action. This is consistent with Kant's later view that consciousness of reason's inner law is the ground of moral feeling. At the same time, it is also consistent with his view in the Prize Essay that moral feeling originates independently of reason (in the feeling of pleasure and displeasure), but that reason must apply the principle of perfection to this feeling in order to arrive at specific moral obligations.

The most important new information related to moral philosophy in Negative Magnitudes is that Kant regards the morality of an action as a function of the internal forces that move one to act, rather than of the external (physical) actions and their

consequences, from which Kant draws the conclusion that "it is impossible for us, with certainty, to infer from another person's actions the degree of that person's virtuous disposition" (2:200). Kant holds that virtuous actions are motivated by moral feeling, but moral feeling must overcome countervailing forces in order to produce action, and we cannot reliably infer the degree of such countervailing forces or of the moral effort required to overcome them in other people. Kant gives the following example:

> Suppose that someone has ten degrees of passion—miserliness, say—and that this is sufficient, under certain circumstances, to conflict with the rules of duty. Let him apply twelve degrees of effort, and let them be exercised in accordance with the principles of benevolence. The result will be two degrees, and that will be the extent to which he will be benevolent and beneficent. Imagine another person who has three degrees of miserliness and seven degrees of capacity to act in accordance with the principles of obligation. The action will be four degrees of magnitude, and that will be the extent to which he will benefit another person after the conflict of his desires. But what is indispensable is this: in so far as the passion in question can be regarded as natural and involuntary, the moral value of the action performed by the first person will be greater than that performed by the second, even though, if one were to assess the actions by reference to the *living* force, the consequence of the latter case exceeds that of the former. (2:200)

So the moral value of an action is a function of the amount of moral effort that actually produces it, which in turn is determined by the strength of the internal obstacles to be overcome. On this view, overcoming more internal obstacles leads to actions with greater moral value, assuming that these internal obstacles are not themselves culpable. (Kant does not explain how we should apply this model to internal obstacles for which we are culpable, such as desires that conflict with our obligation only because we voluntarily and habitually gratify and thus strengthen them). This does not, however, imply that to become virtuous we should cultivate internal obstacles to virtue and then struggle to overcome them, because such obstacles would not be "natural and involuntary." It does,

however, seem to imply that morality involves a certain amount of luck, if one's actions can possess a high moral value only if there are strong internal obstacles that one must overcome in order to perform those actions. But it is not all luck, of course, since one must still muster the effort to overcome those obstacles.

III. Remarks in the Observations

Our most important source of insight into Kant's thinking about moral philosophy in the 1760's are not his published texts, however, but unpublished remarks that Kant wrote in his own interleaved copy of a book he published in 1764, entitled *Observations on the Feeling of the Beautiful and the Sublime*.[5] The most important of these remarks do not relate directly to the content of the *Observations* itself, which is of marginal interest for the development of Kant's moral philosophy because it deals primarily with the different tastes of men and women and of people from different cultures. The more important remarks seem rather to have been prompted by Kant's reading of Rousseau, whose novel *Julie* was published in 1761, followed by *On the Social Contract* and *Émile* in 1762. Kant probably wrote his remarks in 1764-65, and they show that reading Rousseau had a major impact on his thinking around this time. With one important exception that Kant emphasizes, Rousseau's influence on him seems mainly to have been positive rather than corrective. The remarks in the *Observations* do not mark a clear break with Kant's earlier thinking as reflected in the published texts examined briefly above. Instead Rousseau mostly

[5] Marie Rischmüller, ed., *Bemerkungen in den "Beobachtungen über das Gefühl des Schönen und Erhabenen,"* (Hamburg: Felix Meiner, 1991) is a more recent and scholarly edition of these notes than volume 20 of the *Akademie* edition. References to the Rischmüller edition are included in Guyer, ed., *Notes and Fragments*.

stimulated Kant to develop his thinking further in certain directions. The result was that many of the central ideas of Kant's later moral philosophy make their first appearance in these unpublished remarks.

The one respect in which reading Rousseau led Kant to actually change his mind about one of his earlier views concerns the sense in which the moral law is a principle of perfection. Kant describes this change in a famous and rare autobiographical remark:

> I am myself by inclination an investigator. I feel a complete thirst for knowledge and an eager unrest to go further in it as well as satisfaction at every acquisition. There was a time when I believed that this alone could constitute the honor of mankind, and I had contempt for the rabble who know nothing. *Rousseau* brought me around. This blinding superiority disappeared, I learned to honor human beings, and I would find myself far more useless than the common laborer if I did not believe that this consideration could impart to all others a value in establishing the rights of humanity. [20:44 (13)]

In other words, Kant represents himself as having been a sort of elitist. Under the influence of German rationalism, he regarded morality as enjoining us to perfect ourselves in all respects, and this view may have led Kant to suppose that "the rabble" were morally inferior to him because of their undeveloped intellectual capacities. But Rousseau convinced Kant that all human beings are inherently worthy of honor. In fact, Kant imbibed from Rousseau a special esteem for the natural or common man, which is reflected in the moral authority Kant later ascribes to "common human reason." The tables are now turned: "the most learned philosopher with all his knowledge [...] is as upright and no better than the common man" and would even be "more useless than the common laborer" if he could not contribute to "establishing the rights of humanity" [20:176 (50). See also 20:175 (49)].

Rousseau's influence led Kant to change his mind about perfectionism in two important respects. First, although Kant

continued to believe that morality enjoins us to perfect ourselves, Rousseau convinced him that the source of the honor or moral worth of human beings is not the degree of perfection that they have already achieved, but rather their perfectibility or their capacity to perfect themselves. Second, Rousseau changed Kant's mind about what human perfectibility consists in, hence about what morality enjoins us to do. Kant represents himself as having formerly believed that developing all of one's capacities to the fullest is an end in itself, whose achievement alone confers honor on a human being. But after reading Rousseau Kant writes, for example, that "the human being is perfect insofar as he can do without but yet has much power left over to promote the needs and happiness of others; thus he has a feeling of a will that is active in behalf of a good outside of himself" [20:146 (37)].

There are a number of important new claims here. Kant now believes that the *goal* of perfecting oneself is ultimately to promote the greatest good not just in oneself but in general, understood in terms of happiness or the fulfillment of needs. This utilitarian-sounding claim is tempered by Kant's view—which he emphasizes in these remarks perhaps more than anything else—that freedom is the supreme moral value, not only because promoting freedom is the most effective *means* to promoting general happiness, but also because happiness itself is valuable to us only on condition that we are also free. "Nothing," Kant writes, "can be more appalling than that the action of one human stand under the will of another" [20:88 (22)]; "freedom [...] is the supreme *principium* of all virtue and of all happiness" [20:31 (9)]; and "the greatest inner perfection and the perfection that arises from that consists in the subordination of all of our capacities and receptivities to the free capacity for choice" [20:145 (36)]. The importance Kant ascribes to freedom here is consistent with what we may anachronistically call a *rule utilitarian*

form of justification, which explicitly appears in some of these remarks. For example, Kant says that "[t]he will is perfect insofar as in accordance with the laws of freedom it is the greatest ground of the good in general" [20:136-37 (33)]. This suggests that the criterion for moral perfection is that one's will accord with the laws of freedom, but that the reason why this is the criterion of moral perfection is ultimately that it best promotes universal happiness.

What does it mean to conform one's will to laws of freedom? In these remarks Kant holds, again under the influence of Rousseau, that it involves acting according to the general or universal will, rather than merely according to one's individual will: "[i]n case of conflict, the universal will is more important than the individual will" [20:161 (44)]. But Kant anticipates an important theme of his later moral philosophy when he adds that laws of freedom require acting according to the universal will without contradicting oneself: "An action considered from the point of view of the universal will, if it contradicts itself, is morally impossible (impermissible). [...] The will of human beings would contradict itself if it willed that it abhor the universal will" (ibid.). "That will must be good which does not cancel itself out if it is taken universally and reciprocally" [20:67 (19)]. Here we see the obvious ancestor of Kant's formula of universal law and contradiction test in the *Groundwork*. As in the *Groundwork*, Kant also claims that the moral requirement to act according to the noncontradictory universal will is categorical, rather than being conditional upon our having some optional ends. We have seen that this develops a distinction Kant first introduced in the Prize Essay. But, unlike in the *Groundwork*, in 1764-65 Kant makes feeling the ultimate judge of whether actions conform to these requirements. Actions displease if they lead to "opposition and contrariety" and please "if there arises harmony and consensus"

[20:156 (42)].[6] Also unlike in the *Groundwork*, Kant again suggests a rule utilitarian form of justification by construing the test for universality and noncontradiction here as "a heuristic means to morality":

> The goodness of the will is derived from the effects of private or public utility and from the immediate pleasure in them, and the former has its basis in need, the latter in the power for the good; the former is related to one's own utility, the latter to general utility; both feelings conform to natural simplicity. But the goodness of the will as a free principle is recognized not insofar as such forms of utility arise from it, but rather it is possible to cognize it in itself. And the happiness of others in accordance with reason. [20:156-57 (42)]

Here Kant distinguishes between a utilitarian *derivation* of what goodness of will consists in, and using the heuristic device of a free universal will to *recognize* goodness of will. In other words, in judging whether a given or proposed action is moral we need only ask whether it would contradict itself from the perspective of the universal will, for which the criterion is pleasure or moral feeling. But we can also ask the further question: why is *that* the test for a good will? Here our answer or "derivation" will appeal to "general utility." A morally good will is one that does not contradict itself from the perspective of the universal will *because* only such a will promotes universal happiness under conditions of freedom, without which that happiness would not be of value to us.

Finally, a brief look at Kant's view of happiness in these remarks suggests why he may have found this form of justification appealing in 1764-65. Kant was much exercised with finding a reliable method for obtaining happiness. The following remark is characteristic:

[6] In the same remark, Kant calls the capacity to judge actions according to these feelings of pleasure and displeasure "the sense of justice," "[t]he common sense for the true and the false," "the sense of good and evil," and "human reason" of the heart rather than the head.

> A person's contentment arises either from satisfying many inclinations with many agreeable things, or from not letting many inclinations sprout, and thus by being satisfied with fewer fulfilled needs. The state of him who is satisfied because he is not familiar with agreeable things is simple sufficiency, that of him who is familiar with them but who voluntarily does without them because he fears the unrest that arises from them is wise sufficiency. The former requires no self-compulsion and deprivation, the latter however demands this; the former is easily seduced, while the latter has been seduced and is therefore more secure for the future. [20:77 (21)]

Once again following Rousseau, Kant regards what he here calls wise sufficiency as the ideal appropriate for modern human beings. Even though happiness consists in the satisfaction of inclinations, it is wisest to limit one's inclinations to those one can reliably satisfy. So the best strategy involves "seeking to be free of [acquired] inclinations and thus learning to do without them gladly. It does not consist in conflict with the natural inclinations, but rather in making it the case that one has none except for the natural ones, because these can be easily satisfied" [20:77-78 (21)].[7] But if the best strategy for securing one's own happiness involves remaining as free as possible from subjection to unnecessary desires, then this suggests that the best strategy for securing universal happiness, insofar as it is in our power, would be to promote the freedom of others and to avoid subjecting others to one's own will.

Kant is skeptical, however, about the ability of human beings to act from purely moral motives without help from religion, whether moral motives are understood in terms of promoting universal happiness or conforming one's will to laws of freedom:

> It must be asked how far internal moral grounds can bring a person. They can perhaps bring him to be good if, in a condition of freedom, he does not have great temptations, but if the injustice of others or the force of mania does him violence, then this internal morality will not have sufficient power. He must have religion and be encouraged by the rewards of a future life; human nature is not capable of an immediate

[7] In this remark, Kant simply calls such a strategy "virtue," but elsewhere he is clear that "one must first eliminate injustice before one can be virtuous" [20:151 (40)].

> moral purity. But if purity were somehow supernaturally brought about in him, then the future rewards would no longer have the property of being motivating grounds. [20:28 (7)]

In other words, morality requires us to do good because it is right rather than because it (also) benefits us, but *at least initially* this motive is not strong enough to move any of us without our first being enticed by hope for future rewards and fear of punishment. But these impure motives must somehow (Kant is not sure how) give way to an immediate feeling of pleasure and joy in acting morally that is strong enough to overpower contrary temptations: "[t]he common duties do not need as their motivating ground the hope of another life, rather great sacrifice and self-denial have an inner beauty; but" even in this case, Kant continues,

> our feeling of pleasure in [acting morally] can never be so strong in itself that it will outweigh the oppression of discomfort, unless the representation of a future condition of the duration of such a moral beauty and of the happiness that will thereby be increased comes to its assistance, so that one will thereby find oneself more capable of so acting. [20:12 (2). See also 20:153 (41)]

So even pure moral motives will falter unless we believe in a "future condition" in which our sacrifice and self-denial have their intended effects of improving our character and promoting general happiness. Thus Kant holds that morality needs religion, both as a bridge to developing purely moral motives in the first place, and to assure us that the intended consequences of acting from moral motives will indeed come to pass. This view again suggests that for Kant in 1764-65 moral laws are justified only if they would lead to a certain consequence if universally followed, namely the greatest happiness under conditions of freedom.

IV. Writings of 1765-66

The influence of Kant's reading of Rousseau, as evidenced in his remarks in the *Observations*, is the most important development in Kant's moral thought in the 1760's. Two published works from 1765-66, shortly after Kant wrote these remarks, also contain brief comments related to moral philosophy, and these published comments show the continuing influence of Rousseau on Kant's thinking.

It was customary for university lecturers in Kant's day to publish announcements describing, in somewhat more detail than we do today, the contents and approach of their upcoming lectures. Kant published an announcement for his lectures during the winter semester of 1765-66, which includes discussion of a course on ethics. In this announcement, Kant does not mention Rousseau by name. Instead he mentions only Baumgarten, as the author of the textbook on which his lectures will be based, and the British moralists Shaftesbury, Hutcheson, and Hume, who, Kant says, "have penetrated furthest in the search for the fundamental principles of all morality" (2:311). This may give the impression that Kant had passed more seriously under the influence of the British moralists, especially since he also claims that "[t]he distinction between good and evil in actions, and the judgment of moral rightness, can be known, easily and accurately, by the human heart through what is called sentiment, and that without the elaborate necessity of proofs" (ibid.). But in fact this is nothing new. Kant had long thought feeling or sentiment to be essential for moral judgment, and we should recall that he also mentioned Hutcheson in the Prize Essay, which clearly does not reflect wholesale agreement with moral sense theory. Indeed, Kant adds in the *Announcement* that the attempts of these British moralists are "incomplete and defective," and then he goes on to describe an approach to his lectures that above all reflects the influence of Rousseau. Kant says that his goal will be to

distinguish between the perfections that are appropriate to human beings "in the state of *primitive* innocence" and "in the state of *wise* innocence" (2:312). This reflects Rousseau's emphasis on the plasticity of human nature, and it again suggests that the starting point of Kant's thinking about morality during this period is the perfection of which human beings are capable in their current state. Kant does not explicitly connect perfection with happiness here. But the resemblance that his distinction here between primitive and wise innocence bears to his distinction in the remarks in the *Observations* between simple and wise sufficiency, suggests that his lectures might plausibly follow the same train of thought that he sketched there: namely, as I interpreted it above, that since the best strategy for modern human beings to secure their own happiness involves remaining as free as possible from subjection to unnecessary desires, the best strategy for securing universal happiness, insofar as it is in our power, would accordingly be to promote the freedom of others and to avoid subjecting others to one's own will.

The second text is the 1766 *Dreams of a spirit-seer elucidated by dreams of metaphysics*. Interpreting this text is tricky, since Kant is playfully investigating the view of the Swedish spiritualist Immanuel Swedenborg, and it is not always clear whether and to what extent Kant is speaking in his own voice. Swedenborg held that a spiritual world exists alongside and interacts with the physical world, and much of Kant's discussion deals with epistemological grounds for accepting or rejecting such a metaphysical picture. But at one point Kant suggests that it would at least seem to explain some of the phenomenology of moral experience, because it seems that moral impulses arise from an "alien will" outside ourselves:

> These impulses often incline us to act against the dictates of self-interest. I refer to the strong law of obligation and the weaker law of benevolence.

> Each of these laws extort from us many a sacrifice, and although self-interested inclinations from time to time overrule them both, these two laws, nonetheless, never fail to assert their reality in human nature. As a result, we recognize that, in our most secret motives, we are dependent upon the *rule of the general will*. It is this rule which confers upon the world of all thinking beings its *moral unity* and invests it with a systematic constitution, drawn up in accordance with purely spiritual laws. We sense within ourselves a constraining of our will to harmonize with the general will. To call this sensed constraining '*moral feeling*' is to speak of it merely as a manifestation of that which takes place within us, without establishing its causes. (2:335)

It becomes clear that Kant is not endorsing Swedenborg's view that we can somehow have insight into the existence and workings of such a spiritual realm, and in any case such insight would be unnecessary for moral purposes (2:372). But even if we cannot have *knowledge* of such a spiritual realm, Kant may be suggesting that we can adequately describe our moral experience (only?) by *thinking* in these terms. In any case, Kant obviously borrows his notion of "the general will" here from Rousseau, although he writes as if it were reified in the manner of Swedenborg's spirits. It is unclear how seriously Kant intended his readers to take this at the time, but it seems to be an ancestor of the noumenal realm or moral world of Kant's later works. If, however, Kant does not unambiguously claim that we ought to think of ourselves *now* as affected by (as he says in *Dreams*) or as inhabiting (he says later) a spiritual and moral realm distinct from the physical world—in fact, at times he appears to deny this (see 2:373)[8]—he does unambiguously claim that we must think of a *future* world with such a "moral unity":

[8] Whatever Kant's position is in *Dreams*, he certainly affirms this view by the early 1770's. See, for example, R1171:

"The moral feeling can only be set into motion by the image of a world full of order, if we place ourselves in this world in thought. This is the intellectual world, whose bond is God.

We are in part really in this world, insofar as human beings really judge in accordance with moral principles" (15:518, 1772-75).

> [T]here has never existed, I suppose, an upright soul which was capable of supporting the thought that with death everything was at an end, and whose noble disposition has not aspired to the hope that there would be a future. For this reason, it seems more consonant with human nature and moral purity to base the expectation of a future world on the sentiments of a nobly constituted soul than, conversely, to base its noble conduct on the hope of another world. (2:373)

So Kant repeats his view, first expressed in his remarks in the *Observations* [20:153 (41)], that morality needs religion even though pure moral motives cannot be based on religious hopes and fears. *Dreams* is the first published text in which Kant expresses this view.

V. The Inaugural Dissertation

Kant finally became a regular professor at the University of Königsberg in 1770, and the custom was to present a dissertation inaugurating his career in this new post. Kant's *Inaugural Dissertation* deals almost exclusively with metaphysics and epistemology, but one passage marks an important departure from his earlier views on moral philosophy:

> [T]he general principles of pure understanding [...] lead to some paradigm, which can only be conceived by the pure understanding and which is a common measure for all other things in so far as they are realities. This paradigm is NOUMENAL PERFECTION. This, however, is perfection either in the theoretical sense* or in the practical sense. [... I]n the latter sense, it is MORAL PERFECTION. *Moral philosophy*, therefore, in so far as it furnishes the first *principles of judgment*, is only cognized by the pure understanding and itself belongs to pure philosophy. Epicurus, who reduced its criteria to the sense of pleasure or pain, is very rightly blamed, together with certain moderns, who have followed him to a certain extent from afar, such as Shaftesbury and his supporters.
>
> *[Kant's Footnote] We consider something theoretically in so far as we attend only to those things which belong to being, whereas we consider it practically if we look at those things which ought to be in it in virtue of freedom. (2:396)

As we have seen, this passage is consistent with Kant's view in the *Prize Essay* that the intellect, understanding, or reason enables us to grasp the supreme law of obligation, which is a principle of perfection. Nothing in Kant's published or unpublished writings between the *Prize Essay* and the *Inaugural Dissertation* indicates that he surrendered this view, although we have also seen that the influence of Rousseau led Kant to modify his understanding of what it means for morality to command perfection. What is new here is that, prior to the *Inaugural Dissertation*, Kant never held that *pure* understanding or the intellect *alone* is sufficient for moral judgment. He always held that feeling or sentiment is *also* necessary for judging that we have specific obligations. Kant never followed Epicurus or Shaftesbury, as Kant represents them here, in reducing moral criteria solely to the sense of pleasure and pain. But prior to 1770 he did hold that it is necessary, although not sufficient, to base moral judgments on feelings of pleasure and pain. Now Kant rejects this part of his earlier view on the Platonist ground that moral perfection, the criterion of right and wrong, is a paradigm or ideal that we can grasp only through the pure understanding. Although this type of Platonism will be short-lived—it soon becomes a casualty of Kant's critical turn or the Copernican revolution in philosophy—Kant never again held that moral judgment is based on feeling, even in part. Though other important aspects of his thinking about moral philosophy continue to be in flux, from 1770 onwards Kant always held that moral judgment is based on reason alone.

VI. Reflexionen prior to the Groundwork

After the *Inaugural Dissertation* of 1770, Kant published nothing of significance until the *Critique of Pure Reason* of 1781. For evidence of how Kant's thinking developed during this crucial

period, we must rely mainly on unpublished notes or *Reflexionen*, which were assigned approximate dates by Erich Adickes in the early twentieth century. Since we cannot be sure about the dating of these notes, any attempt to reconstruct the development of Kant's thought during this period is somewhat speculative. Accordingly, I limit myself here to identifying distinct lines of argument that Kant develops in these notes and do not pretend to establish a chronology that charts when Kant abandoned one line of argument in favor of another. In any case, it may well be that Kant tried to work out multiple argumentative strategies at the same time, either because he didn't realize that they pulled in different directions or because he realized this but was searching for a reason to prefer one strategy over the other. The only dates that we can more reliably use to distinguish different periods in the development of Kant's ethical thought after 1770 are 1781, when the first edition of the *Critique of Pure Reason* appeared, and 1785, when the *Groundwork* appeared. I regard the *Groundwork* as marking the beginning of Kant's mature moral philosophy. Some passages on moral philosophy in the first *Critique* and some of the notes that Adickes could date only to the period 1780-89 express views that, on my reading, Kant probably had abandoned before 1785.

Nearly all of the issues and themes that appear in Kant's unpublished notes on moral philosophy between 1770-85 made their first appearance earlier, especially in Kant's 1764-65 remarks in the *Observations*. I focus here only on one set of issues that seems especially to have exercised Kant in these notes: namely, the relationship between the moral incentive and the fundamental principle of morality. Kant seems to have been asking himself over and over again: how should the incentive to act morally be characterized, such that acting on that incentive can lead one to do all and only what the fundamental principle of morality requires?

And conversely: how should the fundamental principle of morality be characterized, such that all human beings can have a sufficient incentive to obey it? I suggest that during this period Kant entertained three main strategies for answering these questions.

VI.1 According to the first strategy, the moral incentive and the moral principle itself come apart, because the incentive to do what morality requires is simply rational self-interest: in other words, we should act morally because it is the best way to secure the best happiness for ourselves. Kant sketches this strategy in R7097: "Moral laws do not have in themselves obligating force, but contain only the norm. They contain the objective conditions of judging, but not the subjective conditions of execution. The latter consist in agreement with our longing for happiness" (19:248).[9] Often, as in the rest of this note, Kant links this strategy with belief in God and the hope that God will reward moral behavior in a future life.[10] It is unclear whether Kant ever considered endorsing such a crude version of this strategy, since we have seen that already well before this period he regarded hope for rewards in a future life as an impure moral motive, which at most can serve as a bridge toward developing pure moral motives.

But there is another, more sophisticated version of this strategy which Kant entertains more seriously, though again we cannot know whether he intended to endorse or reject it at any given time. This involves claiming that acting morally is the best way to

[9] 1776-78.

[10] Sometimes Kant alludes to nonreligious reasons for believing that acting morally is the only way to happiness that one desires for oneself. For example: "The satisfaction in the happiness of the whole is really a longing in accordance with the conditions of reason for one's own happiness. For I cannot hope to be happy if I were to have something special and fate were to have a special relation to me" [R6965, 19:215, 1776-78? (1770-71? 1773-75?)]. Perhaps Kant means here that it's implausible to hope that fate will reward me alone with happiness, but he may be suggesting that psychologically I simply could not enjoy happiness unless others were happy as well.

obtain happiness for oneself, not because God rewards moral behavior, but rather because my happiness depends on the cooperation of other human beings. So Kant argues perhaps most clearly in R7199:

> The first and most important observation that a human being makes about himself is that, determined through nature, he is to be the author of his happiness and even of his own inclinations and aptitudes, which thus make this happiness possible. He concludes from this that he had to order his actions not in accordance with instinct but in accordance with concepts of his happiness which he himself makes [...]. As a freely acting being, indeed in accordance with this independence and self-rule, he will thus have as his foremost object that his desires agree with one another and with his concept of happiness, and not with instincts; and the conduct befitting the freedom of a rational being consists in this form. [...]. Thus the motivating ground of a rational being should not be empirical self-love, because this proceeds from the individuals to all, but rational self-love, which obtains the rule for the individual from and through the universal. In this way he becomes aware that his happiness depends on the freedom of other rational beings, and that it would not agree with self-love for each and everyone to have just himself as his object, thus his own happiness [must come] from concepts and be restricted through the conditions that he be the author of universal happiness or at least not contradict others being the authors of their own happiness. (19:272-73)[11]

In other words, rather than deriving morality from antecedent principles of self-interest, moral rules in fact provide the best strategy for obtaining one's own happiness, since this is necessarily bound up with universal happiness, and the goal of morality is to produce universal happiness.

On the sophisticated version of this strategy, then, the moral incentive is rational self-interest, and the moral norm directs us to promote universal happiness. Thus Kant claims, for example in R6714, that morality and general utility coincide: "Morality is in agreement with universal and general utility and hence meets with necessary approval. This also seems to be the true cause of its preeminent goodness" (19:139).[12] Usually Kant tries to introduce

[11] 1780-89? 1776-79??
[12] 1772? (1771?).

freedom in this context as a means to happiness, or as a condition of our enjoying it, or both. For example:

> Do the good gladly. Seek your happiness [*crossed out*: through freedom] under the universal conditions [*crossed out*: of freedom] thereof, i.e., those that tend toward [*crossed out*: are valid for] the happiness and the freedom of everyone, and that are also valid for the essential ends of nature. (R6989, 19:221)[13]

Sometimes this leads to strikingly rule utilitarian formulations of the moral norm. For example:

> The rule of actions through which, if everyone were to act in accordance with it, nature and the human power of choice would universally concur for happiness, is a law of reason and as such signifies morality. (6958, 19:213)[14]

But it should be emphasized that the rules Kant probably has in mind are again laws of freedom, which restrict each person's will to conditions that harmonize with everyone else's will, or with the universal will. So they are not rules that, if followed by everyone, would positively produce universal happiness. Instead they are rules that, if followed by everyone, would give each of us the freedom to produce our own happiness, though others may perhaps contribute positively to our happiness as well. This, however, still seems to be a basically rule utilitarian view.

Kant recognized that this first strategy of coupling the motive of rational self-love with a rule-utilitarian moral principle is seriously flawed. The main problem is that such rules will lead to universal happiness, in which my happiness is included, only if everyone follows them. But obviously not everyone does act morally all of the time. Morality, however, seems to obligate me even if others act immorally. Yet on this strategy I would have no incentive to act morally if others were not doing the same, because rational

[13] 1776-78.
[14] 1776-78? (1770-71? 1773-75?).

self-love would recommend promoting only my own happiness rather than universal happiness in that situation. So it seems that rational self-love cannot be the moral incentive. Kant raises this problem in R7204:

> The foremost problem of morals is this: Reason shows that the [*crossed out*: universal] thoroughgoing unity of all ends of a rational being with regard both to himself as well as to others, hence formal unity in the use of our freedom, i.e., morality, would, if it were practiced by everyone, produce happiness through freedom and would derive the particular from the universal, and, conversely, that should the universal power of choice determine every particular one, it could act in accordance with none but moral principles. At the same time it is clear, however, that if only one were to subject himself to this rule without being certain that others would also do likewise, his happiness would not be obtained in this way. Now the question arises, what is left to determine the will of every (right-thinking) person to subject himself to this rule as inviolable [?] (19:283)[15]

Kant assumes that a pure moral incentive should be sufficient to motivate anyone to act morally even when others are not doing the same. Instead of trying to argue that rational self-love would so motivate us, Kant entertains a second strategy that introduces a different moral incentive.

VI.2 Kant's second strategy replaces rational self-love with the incentive to be worthy of happiness, but it uses the same rule utilitarian moral norm as the sophisticated version of the first strategy. So, on this second strategy, the starting point of my practical reasoning is my interest in obtaining happiness for myself, but I limit my pursuit of my own happiness to rules or conditions that make it possible for others to participate in happiness as well. Why do I limit myself in this way? Because I cannot enjoy happiness unless I am also worthy of it. At least this is true of any virtuous person, because the motive to be worthy of happiness is the moral incentive. But the moral norm still directs me to act according to laws of freedom since these alone make possible universal

[15] 1780-89? 1776-78? (I return to R7204 in section VI.3 below.)

happiness. Only by doing my share to promote universal happiness do I become worthy of enjoying happiness myself.

There are many notes from the 1770's in which Kant develops this second strategy. For example:

> The concept of morality consists of the worthiness to be happy (the satisfaction of one's will in general). This worthiness rests on correspondence with the laws under which, were they universally observed, everyone would partake of happiness to the highest degree, as can occur only through freedom. (R6892, 19:195)[16]

But this is also the strategy that predominates in the *Critique of Pure Reason* of 1781. This book includes a sketch of Kant's moral philosophy in a section entitled the Canon of Pure Reason, where Kant writes the following:

> The practical law from the motive of **happiness** I call pragmatic (rule of prudence); but that which is such that it has no other motive than the **worthiness to be happy** I call moral (moral law). The first advises us what to do if we want to partake of happiness; the second commands how we should behave in order even to be worthy of happiness. The first is grounded on empirical principles; for except by means of experience I can know neither which inclinations there are that would be satisfied nor what the natural causes are that could satisfy them. The second abstracts from inclinations and natural means of satisfying them, and considers only the freedom of a rational being in general and the necessary conditions under which alone it is in agreement with the distribution of happiness in accordance with principles, and thus it at least **can** rest on mere ideas of pure reason and be cognized *a priori*. (A806/B834)

One important feature of this strategy is that it fits nicely with Kant's long-held view that, although morality is not based on religion, it nevertheless needs religion. If my motive to act morally is that I want to be happy, but only under the condition that I am worthy of happiness, then my motive to act morally will vanish unless I can conceive of some way in which it is possible for me to achieve this goal. But since I become worthy of happiness myself only by following laws that make universal happiness possible, it is

[16] 1776-78. See also, for example, R1187, 4612, 6856, 6857, 6910, 6965, and 6971.

this universal happiness that I must be able to conceive as possible in order to sustain my moral motives. Is universal happiness possible? Even if everyone acted morally, would universal happiness be the result? It seems that the answer would not necessarily be "yes," unless nature itself were designed to cooperate with our collective efforts. Hence for this incentive to motivate us, we need to believe in a God who created nature such that it necessarily cooperates with our collective efforts to promote universal happiness. So Kant argues in R6876: "[T]he **nature** of things [...] **contains no necessary connection** between good conduct and well-being, and thus the **highest good** is **a mere thought-entity**. [... R]eligion alone can prove the reality of this *summi boni* with regard to human beings" (19:188).[17] By the highest good, Kant means a state of affairs in which everyone would be both happy and worthy of it, which is to say, virtuous. In other words, the highest good is the end that justifies the laws of freedom as means according to the rule utilitarian moral principle endorsed by this strategy. In the *Critique of Pure Reason* Kant makes clear that this strategy depends on belief in God and the possibility of the highest good:

> Thus without a God and a world that is not now visible to us but is hoped for, the majestic ideas of morality are, to be sure, objects of approbation and admiration but not incentives for resolve and realization, because they would not fulfill the whole end that is natural for every rational being and determined *a priori* and necessarily through the very same pure reason. (A813/B841)[18]

I will not chart Kant's rejection of this strategy in detail here, since he addresses this in the Dialectic of the *Critique of Practical Reason*, but it is obvious that this strategy really does base morality on religion in a way that Kant regards as problematic. We need independent reasons to believe in a God if our incentive to be moral

[17] 1776-78.
[18] See also R6858, A468/B496, A589/B617, and A815/B843.

depends on such belief, but Kant wants to hold that belief in God is a product of moral incentives rather than their basis. That is, he wants to hold that we believe in God because we are independently committed to acting morally, which leads us to hope that the goal of morality one day can be realized. Our belief in God is the expression of this hope. But this is obviously circular. Belief in God cannot be both the basis and a product of moral incentives. So something is wrong with the view that the incentive to act morally is the interest in becoming worthy of happiness.

VI.3 The rest of R7204 summarizes these two strategies and broaches a third one, which, I suggest, carries over into at least one major thread of Kant's argument in the *Groundwork*:

> Now the question arises, what is left to determine the will of every (right-thinking) person to subject himself to this rule [of morality] as inviolable [even when others act immorally]:* [1] happiness in accordance with the order of eternal Providence, or [2] the mere worthiness to be happy (in accordance with the judgment of all that he did as much as he could to contribute to the happiness of all, or [3] the mere idea of the unity of reason in the use of freedom [?] This last ground is not to be valued lightly.
>
> *[Kant's footnote] (How can this *a priori principium* of the universal agreement of freedom with itself interest me? Freedom in accordance with principles of empirical ends has no thoroughgoing consensus with itself; from this I cannot represent anything reliable with regard to myself. It is not a unity of my will. Hence restricting conditions on the use of the will are absolutely necessary. Morality from the *principio* of unity. From the principle of truth. That one complies with one's *principium* that one can publicly avow, which is thus valid for everyone. Perfection in regard to form: the [*crossed out*: universal] agreement of freedom with the essential conditions of all ends, i.e., *a priori* purposiveness.) (19:283-84)[19]

The third strategy broached obscurely in this note is explained more clearly in R7202:

> Freedom is in itself an ability to act and to refrain from action independently of empirical grounds. Thus there can be no grounds that

[19] 1780-89? 1776-78? (This quotation from R7204 picks up where the one at the end of section VI.1 above leaves off.)

> would have weight to determine us empirically in all such cases. The question is thus: how may I utilize my freedom in general? I am free, however, only from the coercion of sensibility, but I cannot at the same time be free from restricting laws of reason; for precisely because I am free from the former I must be subject to the latter, since otherwise I could not speak of my own will. Now this same unrestraint through which I can will what is itself contrary to my will, and because of which I have no secure basis to rely on myself, must be displeasing to me to the highest degree, and a law will have to become known as necessary *a priori*, in accordance with which freedom is restricted by conditions under which the will agrees with itself. I cannot renounce this law without contradicting my reason, which alone can establish practical unity of the will in accordance with principles. (19:281)[20]

Finally, consider Kant's summary of this third strategy in R7220:

> One represents freedom, i.e., a power of choice that is independent of instincts or in general of direction by nature. So freedom is in itself a rulelessness and the source of all ill and all disorder where it is not itself a rule. Freedom must accordingly stand under the condition of universal conformity to rules and must be an intelligent freedom, otherwise it is blind or wild.
>
> Whatever the *principium* of the rules for the use of freedom in general is, is moral. (19:289)[21]

As the latter two notes make clear, the starting point for this strategy is not happiness but freedom. In other words, the problem to be solved by morality is not: how do I become happy, or even how do I become worthy of happiness? The problem is rather that I am in some sense free, and that my freedom is lawless, the source of all manner of ills, unless I subject it to some law. Put differently, I do not properly speaking have a will at all, or at least a unified will, unless I act according to rules, because otherwise my behavior is entirely uncontrolled and incoherent. But to act according to rules is to subject my freedom to a law. So I need to subject my freedom to some law in order even to have a unified will or, in that sense, to be a self at all.

[20] 1780-89.
[21] 1780-89? (1776-79?)

Now Kant argues, as the final note indicates, that *whatever* law unifies my will and enables me to control my freedom would be a moral law; and it turns out, according to Kant, that there is only one such fundamental law. This is the view that Henry Allison calls "the reciprocity thesis" and that Kant first defends in print in the *Groundwork*: "a free will and a will under moral laws are one and the same" (4:447).[22] But if we set aside examining that controversial claim as a task for an analysis of the *Groundwork* itself, then we begin to see the outlines of this third strategy and how it differs from the other two.

The incentive to act morally is now, essentially, that we want to be free, or more strictly that we want to be autonomous, where this is understood as agency that is law-governed because we give the law to ourselves. The idea is that we can be free only by acting morally, and that all moral action is free action. So the law that makes us free is, on this view, the moral law. Notice that nothing requires such a law to make any reference to happiness. It may *turn out* that acting morally enables us to become happy, but this is *not required* by the way this strategy specifies the moral norm. A rule utilitarian form of justification is thus entirely out of place here. Even freedom is not to be understood as the end or goal of acting morally. Rather, acting morally *consists in* acting freely, and vice-versa, whatever the consequences may be. So this strategy not only introduces a moral incentive that differs from those of the other two strategies. It also relies on a different moral principle, which is fundamentally nonconsequentialist.

One more aspect of this third strategy should be mentioned. Happiness is something of which nonhuman animals are capable in a sense that is at least analogous to human happiness: nonhuman

[22] See Henry Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990), chapter 11.

animals avoid sources of pain and seek to satisfy their desires, even if their desires are more basic than ours and it is not the idea of desire satisfaction but rather desires themselves that move them to act. But Kant holds that nonrational animals are not free in any sense that is analogous to human freedom. We have seen that since 1770 Kant held that moral judgment depends on reason rather than feeling. On this strategy, this means that reason is the source of the law that makes us free: namely, the moral law. So only rational beings can be moral or free, and in fact morality comes into the world with free rational beings. Human beings are not the only possible rational beings, but as a matter of fact we are the only actual ones in the natural world. It is also an implication of Kant's third strategy, but not of the other two strategies, that human beings are not just the only moral agents but also the only objects or ends of moral action. If the moral norm constrains my behavior only toward beings capable of freedom, rather than toward all beings capable of happiness (which includes nonhuman animals), then I can have direct moral duties only toward human beings.

These are well-known but controversial features of Kant's mature ethical thought in the *Groundwork* and later works. Kant's early ethical writings, especially his unpublished notes, shows that his mature views developed through a much more extensive engagement with rule utilitarian forms of moral reasoning than his later ethical writings suggest. The *Groundwork's* heavy emphasis on purifying the foundations of moral philosophy, so that both the fundamental principle of morality and our motivation to obey it make no reference to happiness (discussion of which in later writings is relegated to a separate, nonfoundational doctrine of the highest good), may reflect Kant's rejection of the rule utilitarian strategies he entertained in these notes and his embrace of this third strategy or later descendents of it.