Autonomy as the Ground of Morality

O'Neil Memorial Lectures

University of New Mexico

March. 1999

Allen W. Wood

Yale University

First Lecture: The Idea of Autonomy in Kant

Those of us who are sympathetic to Kantian ethics usually are so because we regard it as an ethics of *autonomy*, based on rational self-esteem and respect for the human capacity to direct one's own life according to rational principles. Kantian ethical theory is grounded on the idea that the moral law is binding on me only because it is a law proceeding from my own will. The ground of a law of autonomy lies in the very will which is to be subject to the law, and this leaves no room for any issue about why this will should obey the law. The idea of autonomy also identifies the authority of the law with the value constituting the content of the law, in that it bases the law on our esteem for the dignity of rational nature in ourselves, which makes every rational being an end in itself.

But the very feature which attracts us to Kant's principle of autonomy also raises troubling questions. A moral law which proceeds from *my* will would seem to be a law valid only for me, or even a law whose content appears to be subject to my whims and arbitrariness. How can a law bind me when I am its author, and therefore capable of changing or invalidating it at my own discretion? The self-esteem which grounds

Kantian morality can therefore begin to seem (as it does to some of Kant's critics), like a kind of arrogance or even a perverse self-deification, in which each person blasphemously usurps the traditional place of the Deity as the giver of moral laws.

Kant emphasizes, however, that the law of autonomy is not subject to my whims; I cannot loose myself from it at will, since it is not up to me to make or unmake the idea of a rational will, nor is the law of autonomy even my *particular* law any more than anyone else's, since I am the author of this law in exactly the same sense that every other rational being is. The moral law can be universally valid for all rational beings only because it proceeds from the will of *every* rational being, and in fact from this will in the form of an *idea,* from my will conceived in its rational perfection.

With this clarification, however, the autonomy which attracted us so much to Kantian ethics begins to look like nothing but a euphemism, or even a deception. If the will which gives the moral law is not *my* will, but an ideal rational will, then there seems no force left in the assertion that this will is *mine*. If the moral law is a law whose authority lies in the power of reason common to all people, then instead of saying that the authority of the law lies in my will, why shouldn't we say instead that its authority lies simply in the *rationality* of its content. Why shouldn't we admit that when we are following the law, we aren't following our own will at all, but merely doing what is rational (even if we really want to do something else)? But if we admit that, then we can still raise the question, which Kant's notion of autonomy claimed to have put to rest, namely, *why* we should follow the rational course if some other happens to appeal to us, or (to put it more pointedly) we can ask *what interest* binds us to follow principles of reason. Whatever answer we may give to this question, it will necessarily compromise the

supposedly categorical nature of moral obligation, since a categorical obligation is one which must bind us independently of any interest. An obligation is categorical if it is something we rationally ought to do not because we will something else, but because the thing we ought to do is itself rationally required.

The Kantian view, of course, is that reason obligates us apart from any interest. That view seems to be involved with the notion that the relation between who I am as an agent and my acting according to reason is much more intimate than any that could be represented by saying that I *have an interest* in so acting. Yet the obvious possibility of my acting according to interests which oppose reason is enough to show that I am capable of understanding myself not only as having those interests, but even as being someone for whom they might outweigh moral (or other rational) considerations. So the very possibility of my viewing moral considerations as decisive for me appears to be one which allows that I am either more or less than a being who acts for moral reasons, and hence that it cannot be true that the moral law is purely an expression of *my* will.

Similar paradoxes – or perhaps just the same ones, looked at from a different standpoint – assail us when we consider the way Kant proposes to establish the moral law by identifying it as a law of autonomy and then relating the idea of autonomy to the idea of *freedom*. Kant understands the practical proposition that the moral law is valid for the human will as distinct from the speculative proposition that the human will is free. Yet he thinks the two propositions reciprocally imply each other, so that if we grant for any reason that the will is free, then we are committed to the validity of the moral law. In the *Groundwork*, Kant even tries to use this reciprocity between freedom and the law to provide a sort of deduction of the law.

Like the idea of autonomy itself, this ingenious argument raises troubling questions. It seems to commit Kant to a theory of freedom many have found Byzantine in its metaphysical extravagance. For Kant denies that we can be free as members of the sensible world, and holds that freedom can be consistently thought only if we ascribe it to ourselves as members of an unknowable noumenal world lying beyond nature, beyond space, even beyond time. To lay such a theory at the ground of Kantian ethics has seemed to many to discredit the entire ethical theory. Further, it commits Kant's theory of autonomy to the acceptance of an idea whose very expression is necessarily self-contradictory or at least oxymoronic: the idea of a *law of freedom.* Kant himself admits that the function of practical laws, insofar as they are imperatives, is to *constrain* the will, to *limit* it to a certain set of acts which it does only with some reluctance. Without the law to constrain it, the will would have been *free* to choose different ones. This makes it hard to understand how the law itself can be seen as arising from freedom. For it seems that what is free is precisely that which is not subject to any law, and what is subject to law is, to precisely to that extent, not free. The very notions of freedom and law, therefore, appear to contradict each other on the most elementary level, and an ethical theory that requires their combination would therefore seem to be doomed before it starts.

The first Kantian line of defense against these objections is to maintain that freedom of the will, properly speaking, is nothing but the capacity to act according to reason, since this capacity is what agency itself consists in. Hence for a finite or imperfect will, which does not necessarily act as reason directs, freedom requires the capacity to constrain itself so to act. This constraint, far from being the opposite of freedom, is a necessary condition for freedom to exist at all. This reply, however, raises some of the same questions we

encountered earlier. For it is committed to saying that we have a fundamental interest simply as agents in acting according to reason, without saying (and even apparently denying that it is possible to say) anything further about what this interest consists in. But if reasons (in particular, moral reasons) can be experienced as constraints – and far from denying that they can be so experienced, Kant is at pains to affirm that for us they often are – then we must have interests which oppose them, and we must be capable of identifying ourselves with these interests or we could not experience reasons as constraints. What could be more evident than that experiencing the moral law as a constraint is incompatible with experiencing it as an expression of freedom?

The two lectures I am about to give will attempt to reply to these worries about an ethics of autonomy. The first lecture will draw mainly on Kant's thought about autonomy, and will proceed by clarifying some of the concepts employed both in an ethical theory based on autonomy and in the above objections to such a theory – concepts such as obligation, law, reason and freedom. The second lecture will focus on Fichte's thought, and will go deeper – exploring Fichte's view that if we explore the nature of self-hood (or 'I-hood') itself, then we will see why an ethics of autonomy is implied by a correct conception of selfhood and self-awareness.

Theories of autonomy and the main objections to them derive from competing pictures of what morality is and what moral demands are about. Most of the objections start from the perception that moral obligations arise from demands made on us by others. We are first taught morality by our parents, and it takes the form of their giving us to understand that there are certain things that we must and must not do. When we are children, some of the do's and don'ts they teach us are for our own protection, but the

ones that lead most naturally into moral obligations are not of this kind. They are rather demands whose observance benefits other people, quite often at our expense. Where it is also in our interest to meet them, this is usually because if we don't, others will punish us in some way for violating our obligations or because they will not show the same good will toward us in the future which our compliance with moral obligations shows toward them. This way of looking at morality naturally leads to the thought that moral demands are made by society on individuals, and their basic purpose is not directly the good of those on whom they are made, but some other good, such as the well-being of people other than the obligated agent, or some larger good, such as the greatest happiness of all, in which I as an obligated agent participate not so much from fulfilling my obligations but from others fulfilling like obligations toward me. This is the conception of morality we find in a utilitarian moralist such as Mill, who represents obedience to moral rules as aimed at protecting the interests of others, and moral blame as an external sanction society attaches to these rules in order to secure our compliance with them. On this picture, just about the only way to represent moral obligations as proceeding from the legislation of a single will is to see them as demands God makes on us through the supreme authority he supposedly has over us in virtue of being our creator.

It is perhaps not well enough appreciated that as long as we are looking at morality as a historical and social phenomenon, this picture is one with which Kant agrees at least in part. His view differs from the utilitarian one we have just sketching chiefly by being less flattering to morality as regards its social origins. Kant maintains that morality arose from certain peculiar features of social life, which he designates by the terms *Sittsamkeit ('propriety')* and *guten Anstand ('*good behavior'*)* or *Anständigkeit*. ('decorum') (Ak

8:113, 7:152, 27:300).[1] Kant thinks our most basic natural impulse as sociable beings is to achieve a status of superiority over others. By the lights of Kantian ethics, this impulse is deeply irrational, since Kantian ethics holds that every rational being has dignity – a supreme and incomparable worth, which is therefore necessarily equal in all rational beings. From the standpoint of reason, then, real superiority over others is not to be had.

In its place, what people in fact seek is the good opinion of others, and they especially try to avoid being despised by others as inferior to them. Through its collective sensitivity to these matters, each society develops a set of customs *(Sitten*), conformity to which is a condition of maintaining the respect of others, and violation of which will cause the violator to be treated with contempt. It is significant, however, that Kant does not see this system of customs as achieving (or even aiming at) any larger good, such as the general happiness.(On the contrary, its basis is the comparative and competitive impulse, which is not only irrational but even the fundamental source of the radical evil in human nature.) Kant focuses instead on the fact that one of the main things I achieve by observing such customs is the successful concealment of things that would lower the opinion others have of me, or even the successful promotion of a false image of myself, which successfully deceives others about what I think, want and habitually do (Ak 7:149-153, 8:113-114). To the extent that morality is placed in the service of religion through the machinations of priestcraft, Kant thinks its rules become statutory observances whose superstitious aim is to win the special favor of supernatural beings by means of hypocritical flattery and self-abasement (Ak 6:167-180). The constant theme here is that the original aim of moral conduct is to win someone's favor through conformity to their

wishes, and that the attempt to do this typically involves some kind of dissimulation or falsehood.

It clashes with the popular image of Kant that he should be seen as a *critic* of morality for its inherent hypocrisy and for representing the subjugation of individuals to mindless social customs. But that shows only that the popular image of Kant is seriously wrong in this, as in many other respects. As soon as we cease to neglect Kant's writings on anthropology and history and begin to appreciate the crucial importance of his theory of human nature for his ethical thought, we will be forced to revise that image quite radically. Kant's view of morality as a social and historical phenomenon follows closely his view of other important social institutions, such as religion and the state. In all these cases, Kant views the institution in question as having a vital rational purpose in human life, but also as arising from something that directly contradicts this rational purpose. The state, whose purpose is to protect external freedom under universal law, arises from the unbridled power of military despots. The church, which is Kant's model for a free ethical community seeking a universal realm of ends, has its origins in the degrading spiritual slavery of priestcraft, which gains power over people by exploiting their superstitious delusions. In these cases, as in the case of morality, it is our vocation to reform the social institution through enlightened thinking so that it might eventually come to serve its proper rational end. The point of Kantian ethics is that the vocation of the human race is to refashion itself, opposing its natural-social impulses that lead to competition and discord, and fostering instead a realm of ends, in which rational beings act according to those laws which bring their ends into necessary agreement in an organic system.

The main point here for our purpose is simple: Kant's theory of morality as based on autonomy was never intended as an empirical sociological or psychological account of morality as a social or historical phenomenon. It aims from the start at telling us instead under what conditions something like moral obligation might have a rational basis. The same is true, of course, of utilitarian reconstructions of moral obligation, such as Mill's. It was never plausible to maintain that morality, as it actually functions in existing societies, effectively promotes the general happiness through protecting individuals from harm by others. The utilitarian's claim, analogous to Kant's, is rather that this is the sole legitimate function that anything like moral obligation might be given.

Kant rejects the utilitarian's rational reconstruction, of course. He does so mainly because it cannot make rational sense of the idea of moral obligation itself. The utilitarian account fastens on the point that moral obligations seem to involve demands made on us by others, with the implicit support of society, which benefit the others, often at our expense. But it neglects the even more important point that it is part of the concept of moral obligation that there are *good reasons* for us to comply with morality's demands even in the absence of coercive threats from society. This feature of obligation, Kant thinks, can be properly accounted for only by a theory of autonomy. Look at it this way: If the demands moral obligations impose on us are reasonable ones – reasonable not only for others to make, but also reasonable for us to fulfill – then there exist good reasons why we should do what morality requires. If such reasons are necessarily binding on us all, and strong enough to override our immediate desires and even our self-interest, then the principles of morality must belong to the fundamental principles of self-government

for every rational being. This is equivalent to saying that moral principles must be principles of autonomy.

In the *Groundwork for a Metaphysics of Morals*, Kant's search for a principle of morality begins with the idea of a categorical imperative, which yields the Formula of Universal Law and its variant, the Formula of the Law of Nature (see Handout). He then inquires after the motive, or the substantive value, which could ground obedience to a categorical imperative. This provides the second main formula of the principle of morality, the Formula of Humanity as End in Itself. The worth of every rational being as an end in itself also explains why the principles of every rational being's self-government should include demands made on the rational being by other rational beings. Kant then puts together the two thoughts behind these principles, that of a categorically binding universal law and that of the will of every rational being as having value as an end in itself, and derives a third formula, the Formula of Autonomy, "the idea of the will of every rational being as a will giving universal law" (Ak 4:431).

It is significant that Kant states the Formula of Autonomy using the word 'idea' (*Idee*). An *idea* is a concept of reason to which no object in the world of appearance can be adequate (KrV A310-320/B366-377). Thus to ground morality on the *idea* of the will of every rational being as legislative is precisely *not* to ground it on what particular rational beings arbitrarily decree. On the contrary, we regard ourselves as *categorically bound* by norms only to the extent that we see them as proceeding from reason, which has the *critical* capacity to recognize its errors and correct them. The volition which is author of categorical obligations is thus the will toward that (unattainable) *idea*, which is the same for every rational being. Kantian autonomy is therefore badly misunderstood when

it is equated with the notion that moral laws are made by the arbitrary will of fallible beings. On the contrary, to ground the moral law on the *idea* of the will is therefore to distinguish moral *truth* from what any finite rational being (or all such beings) might believe. (Since Kant holds that moral truth is irreducible either to what people think or to the results of any verification procedures, he is a moral *realist* in the most agreed upon sense that term has in contemporary metaphysics and metaethics.)

The nub of the argument as Kant presents it in the *Groundwork* is that no law whose bindingness rests on an external interest can be truly universal, that is, valid categorically for every rational will simply as such. For the interest which grounds it applies to the will only through a contingent inclination grounding the interest. Hence such a law "would itself need yet another law that would limit the interest of its self-love to the condition of a validity for universal law" (Ak 4:432). But a law so limited could not command unconditionally or categorically. The only way to conceive of a law which does command in that way is to suppose that its ground is the supreme worth of the rational will itself which obeys the law. This worth is present as much in others as in myself, and requires respect for them as much as for myself. It is objectively valid, moreover, only if it accords with the idea of the will, and not merely with the *fiat* of some fallible being such as myself. To the extent that I esteem myself as a rational being, a law conceived in this way is given by my will too, the very will that is to obey it. Thus it is possible to regard this same law as categorically obligatory by viewing it as proceeding from my own will.

A law proceeding from a self-legislating rational will obligates us only through *respect*. Since it is the rational will that is the author of this law, it is in a deeper sense *the rational will* which is the object of respect. Rational nature, that is, can be seen not

only to be an end in itself (with fundamental objective worth), but to have *dignity* (absolute or incomparable worth).

> "Nothing can have a worth other than that which the law determines for it. But the lawgiving itself, which determines all worth, must for that reason have a dignity, that is, an unconditional, incomparable worth; and the word *respect* alone provides a becoming expression for the estimate of it that a rational being must give. *Autonomy* is therefore the ground of the dignity of human nature and of every rational nature" (Ak 4:436).

Every other source of the law would have to bind the rational will to it by some *other* volition, thus grounding it contingently on a value different from that of the law (or of the rational will which gives the law). A law grounded on happiness, for instance, would have to appeal to our will to be happy. A law grounded on the will of God would have to appeal either to our love of God's perfections or our fear of his power. These further volitions would turn the categorical demand of the law into a merely hypothetical demand, by referring it to some other volition as its ground. This line of thinking convinces Kant that the principle of autonomy is the only possible solution to the problem of obligation, and that all other principles of obligation must fail to solve it because they must be grounded on heteronomy of the will (Ak 4:441-445; 5:34-41).

There are two possible kinds of rejoinders to this argument. The first kind denies that there really is such a thing as categorical obligation, and insists that moral volition can be grounded only on something contingent and subjective, such as a moral feeling or a desire for happiness. This says in effect that moral obligation, as Kant has been depicting it, is (in his words) nothing but a "high flown fantasy," "chimerical idea" or "cobweb of the brain" (G 4:394, 407, 445). It tries to depict the idea of a categorical imperative as something strange and fantastic, a metaphysical invention of philosophers rather than what it is, namely an essential feature of our ordinary notion of moral obligation. After this bit of rhetorical hand-waving, the objectors then dress up their favored deflationary alternative in its Sunday best, and through the pressure of skeptical arguments against the real thing, they try to blackmail us into accepting their sad imitation in its place. The

second line of objection agrees that moral obligation as Kant presents it is real, but tries to account for it in some way other than autonomy of the rational will – usually by appealing to some objective value external to that of our own autonomous wills. In the Second Section of the *Groundwork*, Kant discusses both sorts of positions under the heading of 'principles of heteronomy'. His most systematic presentation of the alternatives is in the *Critique of Practical Reason*, in a table given at the bottom of the first page of the Handout (Ak 4:442, 5:40).

Kant's examples of the first (or "deflationary") strategy are the conventionalist theories of Montaigne and Mandeville, the hedonism of Epicurus, and the moral sense theory of Hutcheson (which was taken up by Hume and Adam Smith, and even tempted Kant himself for a while in the early 1760s) (Ak 2:298-300). In the *Groundwork* Kant seems to dismiss such views peremptorily as principles of heteronomy, but his real strategy is to postpone his response to them. For throughout the First and Second Sections, he is *provisionally assuming* that morality (as he conceives it) is not an illusion. After inquiring into the conditions of its possibility, he will present a positive argument for his account only in the Third Section, where he connects the moral law with the rational presupposition of freedom. We will turn to that argument presently.

The second (or "objectivist") way of disagreeing with Kant accepts the reality of obligation but tries to account for its categorical bindingness through the idea of an objective good external to the rational will. These attempts include divine command theory (familiar to Kant from the writings of Christian August Crusius) and the theory of perfection (which Kant found in Christian Wolff and the Stoics) (Ak 4:443, 5:40-41). Although Kant's official objection is that these positions involve principles of heteronomy, a closer look shows that his real argument against them takes the form of a dilemma: Either these positions are committed to principles of heteronomy, and are therefore unsatisfactory because they cannot account for the categorical character of moral obligation, or else their account of objective goodness itself must remain opaque

unless it is seen to be grounded on the dignity of self-legislating reason, which it requires to explain or complete it.

Kant has no objection to our thinking of ourselves as obeying God when we do what morality requires, but he denies that this thought provides a satisfactory account of moral obligation. Kant distinguishes the *legislator* of a law, the one who issues a command and attaches positive or negative sanctions to it, from the law's *author*, the one whose will actually imposes the categorical *obligation* to obey it. Kant has no objection to regarding God's will as the legislator of the moral law, but thinks only the rational will of the person obligated can be the author of the law (Ak 4:443). For if I regard a will other than my own as author of the law, then the law obligates me only through some interest (such as love or fear) that I have in obeying that other will. This interest would undermine the categorical nature of moral obligation.

If my motive for obeying God's will is fear, then I seem to be representing God as a cosmic despot, motivated by a desire for glory. The authority of his commands to us -- his groveling minions – rests on our dread of divine vengefulness or our hope of gaining divine favor for our own aims (Ak 4:443). Such a picture demeans both the Deity and ourselves, and degrades virtue into mere hired service (Ak 8:339; 28:1115, 1118). We don't do much better by representing our obedience to God as grounded on love of God if that love is merely another volition on which our reason for obedience depends, since that equally destroys categorical obligation. We avoid the problem only if we hold that God's will itself is inherently worthy of obedience because what God commands is in itself right (i.e. categorically obligatory). But this means we are still faced with our original problem of determining what makes something categorically obligatory, to which we now see that appeals to the divine will can contribute no solution.

Kant regards the principle of perfection as the alternative to autonomy that comes closest to a correct account of obligation (Ak 4:443). But problems arise when we ask what is meant by 'perfection'. We get a thoroughly unsatisfactory theory of obligation if

'perfection' is understood as the fitness of an object to some (arbitrarily chosen) end, e.g. a perfect fruit knife is one that can successfully cut up fruit. For that directly undermines the categorical character of duty. We do no better by understanding 'perfection' as relative to a concept of the general kind of thing (e.g. a perfect fruit knife as one which is sharp, safe, easy to use, and so on; a perfect human being is one which behaves in this or that way). That makes the value of perfection conditional on our interest in that kind of thing, which – even if the kind is "human"-- is still an interest independent of the moral law (Ak 4:441-442, cf. 5:22-26).

Of course in the case of a human being, "perfection" might be used in such a way that it refers simply to the goodness of the morally good will itself. One version of such a view, found in Aristotle and the Stoics, would be a form of eudaimonism (or theory of happiness) that identifies happiness not with subjective satisfaction but with a person's objective good, and equates this good (or its dominant component) with moral virtue or the exercise of practical reason. Kant's objection here is not that perfection is a principle of heteronomy, but that this concept of perfection is too indeterminate and empty to provide a satisfactory account of moral obligation. Perfectionists might say at this point that they do exactly what Kant does -- they rest obligation on our rational esteem for the objective worth of something. For Kant this value is the dignity of rational nature, for the perfectionist it is simply perfection or objective goodness wherever it might be found. Why does Kant think there is greater clarity in conceiving the ground of obligation is the dignity of rational will than as objective perfection or goodness? (Theological moralists might make an analogous objection here, saying that they stop with the transcendent goodness of the Deity.) Kant's reply is that the recognition of a law as categorically binding presupposes the unconditional and incomparable worth of the source of the legislation, which in relation to practical reason is adequately conceived not as perfection or divinity only as rational self-legislation. For only this has such a worth to the rational will originally rather than derivatively, making its commands truly categorical.

Let us try to state Kant's argument in a slightly different way. If my recognition of an obligation is supposed to be based on something over and above the dignity of my legislating reason (such as the value of objective perfection or divine goodness), then a further ground would be needed to explain why my will values that object. If that ground is distinct from my respect for rational nature as self-determining, it thereby renders my acceptance of the obligation conditional on some other volition of mine, and the categorical nature of the obligation has been forfeited. If it turns out to be the same ground as respect for my rational nature, then perfectionism or divine command morality becomes acceptable, but only because its account of obligation turns out to be parasitic on a covert appeal to the autonomy of reason. The point we have to come back to is that no property of any object could provide me with a reason for regarding any act as categorically obligatory except insofar as that property plays a reason-giving role in following some principle legislated by my rational will. Kant does not deny that there is such a thing as objective goodness, but he does maintain that it can offer us objective reasons for acting only if it operates through our respect for our own rational capacity for self-legislation.

Now let us return to the more radical (or "deflationist") line of objection. As we have said, it is really answered only in Section Three of the *Groundwork,* by linking the moral law to the presupposition of freedom. Kant's argument here is not easy to summarize. Between 1781 and 1788 he gave at least three distinct accounts of the relation between morality and freedom -- in the first Critique, the *Groundwork* and the second Critique. My account of his argument will be closer to the *Groundwork* than to the other two works, but it is an unashamed reconstruction that does not precisely follow any of the three texts. Its aim is chiefly to call attention to those features of Kant's account that (in my view) have the greatest lasting philosophical interest.

The basis of Kant's deduction of the moral law is what Henry Allison has called the "Reciprocity Thesis", presented on the back page of your Handout. The reciprocity thesis is an alleged mutual entailment between the propositions F and M:

F: The rational will is free.

M: The moral law is unconditionally valid for the rational will.

The argument as I will present it attempts to ground the moral law on F↔M by using the first half of the entailment (F→M), discharging the antecedent of the conditional by arguing that F as an indispensable presupposition of all rational judgment (theoretical or practical).

Kant distinguishes several senses of 'freedom': *Transcendental* freedom is the capacity of a cause to produce a state spontaneously or "from itself" (*von selbst*) (KrV A533/B561). A transcendentally free cause, in other words, is a "first cause", one which can be effective independently of any prior cause. This is distinguished from *practical* freedom, which we attribute to ourselves as agents. Kant's metaphysical contention is that the will can be practically free only if it is transcendentally free, and transcendental freedom could exist only in a noumenal world, not in the empirical world. But his argument for the moral law is really concerned only with practical freedom – which even Kant himself sometimes thinks can be treated independently of speculative issues about transcendental freedom (KrV A800-802/B829-831). Practical freedom, in turn, is taken in two distinct senses: In the "negative" sense, a will is practically free if it acts independently of external causes determining how it acts; in the "positive" sense, it is

practically free if it has the power to determine itself in accordance with its own law (KrV A534/B562, Ak 4:446, 5:33).

The key to understanding the Reciprocity Thesis is Kant's view that freedom is causality, but causality "of a special kind" (Ak 4:446). A *natural* cause is a state of a substance upon which another state of some substance follows in accordance with a necessary rule; this rule is the pertinent causal law (KrV A189/B232, A534/B562.). But since a will acts not only according to laws but according to their *representation*, the "law" of a *free* cause must be one it *represents* to itself (Ak 4:412). This cannot mean merely that a free will is *aware* of the law it follows. For the law is one under which it considers its actions *from a practical standpoint*. In the case of an imperfectly rational will, which does not always act as reason directs, the law is represented as a principle according to which it *ought* to act. I will refer to such a law, in contrast to a natural law, as a *normative* law.

The notion of a cause acting according to normative laws may strike us as bizarre, but it is not. For we often, or even typically, explain human actions by reference to norms the agent recognizes. A chess player moves the bishop only diagonally because that is the rule in chess. In constructing the sentences they speak or write, people choose words that accord with the rules of grammar. We use these rules to explain why the sentences are formed as they are. An appeal to norms also explains why composers avoid parallel fifths and why a batter keeps his weight on his back foot as long as possible. Explanations of actions according to the agent's intentions are all normative law explanations. It is only because intentions are norms that people can bungle their intended actions or fail to carry them through. Yet despite such cases, we do not regard explanations by reference to

intentions as defective, pointless or merely bad substitutes for natural law explanations. We even use normative laws as part of explanations of actions that *contravene* them, by describing the actions as *failed attempts* to comply with the norm. Normative law explanations are uniquely appropriate to voluntary, rational actions, simply because rational actions are in their very concept freely chosen and norm-guided.

Kant argues for F→ M on the ground that freedom is a kind of causality, together with an analysis of what 'causality' must mean in the case of freedom. It must mean being subject to an unconditional and self-given normative law. (If the will is perfect or holy, the normative law tells us what its self-determined volitions necessarily *are*; if it is finite and imperfect rather than holy, then this law is a categorical *imperative*, determining what its volitions *ought* to be.) The argument of the Second Section of the *Groundwork* has shown that the moral law, as most fully developed in the Formula of Autonomy, is exactly such a law for any rational will. Therefore, if there is a free will, then the moral law is valid for it (in other words, F→ M). To complete the argument, Kant now needs only to discharge the antecedent by showing that we have reason to assert F.

In the *Groundwork* Kant claims that "freedom must be presupposed as a property of the will of all rational beings" (Ak 4:447). Freedom is not being proved theoretically, but it is claimed to be a *presupposition* of taking the practical standpoint at all -- which we unavoidably do, even when we are engaged in theoretical inquiry.

> "Now, one cannot possibly think of a reason that would consciously receive direction from another quarter with respect to its judgments, since the subject would then attribute the determination of his judgment not to reason but to an impulse. Reason must regard itself as the author of its principles independently of alien influences; consequently, as practical reason or as the will of a rational being it must be regarded of itself as free" (Ak 4:448).

Notice that these remarks, focusing not on actions but only on *judgments,* concern the way in which we must regard ourselves in making judgments of any sort, even wholly theoretical ones. If even there we must regard ourselves as free, then there is no room in any sort of understanding of ourselves for a conception of ourselves as other than free.

Kant holds that we must think of ourselves as free in all our rational judgments because we must regard our judgments as acts we perform because they are required by certain norms. Suppose I judge that *q* based on the evidence that *p or q* and *not-p*. Here I can regard this as a *rational judgment* on my part only if I am prepared to give it a normative explanation, by viewing it as proceeding from my correct application of the logical rule *modus tollens*, regarded as a *normative* principle which I, simply as a rational being, recognize as valid and therefore impose on my own judgments. To say that judgment is an exercise of free agency, in the sense we mean here, is therefore precisely *not* to say that I may judge any way I please. On the contrary, my judgment can go contrary to the norm only if it involves a failure or mistake. If I think of my judgment as prompted by some conscious *cause* external to my free and rational norm-guided activity (for example, if I see it as prompted by fear of what my logic teacher will do to me if I don't give the answer of which I know she approves), then to that extent I cease to regard it as a judgment which is *rational* by the standard of the relevant norms (logical rules of inference). If my fear of my logic teacher leads to my giving the right answers, that will be because those happen to be the ones the teacher wants; but the rightness of my judgments would be only contingently the result of anyone's applying rational norms (and the rationality of my judgments would have to be ascribed to my logic teacher, not to me). The verdict would be the same if I came to regard my judgment as the result of some unconscious process (of neurotic compulsion or post-hypnotic suggestion) whose results accord only contingently with the rules of logical inference. (It is noteworthy, however, that even when Freudian explanations undermine the normative law explanation I consciously give for what I do, they are not natural law explanations but normative law

explanations, insofar as they typically appeal to unconscious *intentions*.) Not all my reasoning processes need be entirely conscious and explicit, but to regard them as successful processes of reasoning, they must be regarded as the result of my freely (though perhaps habitually and unreflectively) following rational norms. Even mistakes in reasoning are regarded as rational processes only to the extent that I see them as falling under such norms and as failing to comply with them.

For this argument to be relevant to *moral* freedom, Kant must maintain that the norms of theoretical reasoning, like those of morality, are both self-given and unconditional. And he does maintain this. But Kant's argument does *not* require him to say that logical rules are a species of moral rule, or that moral rules are merely rules of logic. What he needs to claim is only that the capacity we ascribe to ourselves in regarding ourselves as subject to moral obligations is of exactly the same *kind* as that we ascribe to ourselves in thinking of ourselves as judging according to rational norms. Thus if we cannot intelligibly doubt that we have such a capacity in one case, we have no good ground for doubting that we have it in the other. And this he *can* claim. For we do not accept logical rules only conditionally -- because, for example, we think that reasoning according to *modus tollens* will be advantageous to us. On the contrary, following *modus tollens* is unconditionally necessary simply in order to preserve the truth of our judgments in making inferences. But seeing myself as following the norms required to judge truly is not an independent interest. For seeing myself as trying to judge truly is no different from merely seeing myself as rationally judging.

Kant's argument may also be regarded as a practical *reductio* or as showing that the denial of practical freedom is in a way self-refuting. Let a *fatalis*t be someone who denies practical freedom (where this freedom is understood in the Kantian way, as causality according to norms). The fatalist, therefore, holds that ~F. She must regard her own acts of judgment *solely* as the necessary effects of natural laws, denying that they can be correctly explained by reference to reasons or normative rules of inference (such as

modus tollens). If fatalism is to be an interesting position, then the fatalist must be prepared to give *arguments* to for ~F, assert ~F on the basis of those arguments and expect those to whom she gives the arguments to be convinced that ~F on the basis of them. Yet fatalism itself says that all judgments (including the fatalist's judgment that ~F and the judgments of those she hopes to convince), are to be explained solely by reference to natural laws and can never be correctly explained by reference to norms of reasoning. Fatalism itself, therefore, undermines the fatalist's claim that she, and those she tries to persuade of fatalism, can hold fatalism on any rational grounds.

In 1783 Kant reviewed a book on moral philosophy by Johann Henrich Schulz, whose position he described as a "universal fatalism, which…turns all human conduct into a mere puppet show and thereby does away altogether with the concept of obligation" (Ak 8:13). In the review, Kant stated his argument against fatalism quite explicitly:

> "Although he would not himself admit it, [Schulz] has assumed in the depths of his soul that understanding is able to determine his judgment in accordance with objective grounds that are always valid and he is not subject to the mechanism of merely subjective determining causes, which could subsequently change; hence he always admits freedom to think, without which there is no reason" (Ak 8:14).

This argument makes the point that we can doubt the reality of freedom only if we also doubt our capacity to judge rationally, including even our capacity to judge whether to entertain those very doubts. A fatalist might still assert fatalism and even present arguments for it. But she would be unable to represent herself or those to whom she offers the arguments as holding fatalism rationally on the basis of those arguments.

Classical compatibilist (or "soft determinist") approaches to the free will problem would accept the fatalist's idea that rational judgment is a natural causal process, but try show, contrary to the fatalist, that it could be *at the same time* a case of free action, or conformity to rational norms. Here Kant sides with the fatalist, holding the compatibilist

project to be impossible. Note, however, that the argument for freedom we have just seen is even more basic than Kant's arguments for incompatibilism. For they say that whatever we may or may not hold about the compatibility of freedom and natural causality, we must presuppose our own freedom, as the capacity to act under norms of reason, in order even to represent ourselves as competent to decide on rational grounds whether fatalism or compatibilism are true. Our agreement or disagreement with Kant's incompatibilism therefore should make no difference to our acceptance of his argument that F is a necessary presupposition of all rational judgment. For the same reason, accepting Kant's argument from freedom for the validity of the moral law does not by itself require us to appeal to Kant's theory of noumenal causality. Of course, if Kant and the fatalist are right that understanding ourselves as rational judgers and agents is incompatible with regarding ourselves as beings belonging to a natural causal order, then Kant's desperate theory of noumenal causality might unfortunately turn out to afford the only possible solution to the problems that raises. But Kant's argument for the validity of moral obligation does not by itself raise those problems or require Kant's solution to them.

Kant's way of vindicating the principle of morality is thoroughly consistent with the transcendental strategy Kant employs throughout the critical philosophy. Kant's argument attacks skepticism about morality by showing that the skeptical doubts undermine the very conditions of their own intelligibility. It is grounded on the typically Enlightenment appeal to that critical self-confidence in reason without which it would be impossible even to acknowledge the limits and fallibility of reason.

In this lecture I still haven't replied to all the objections I raised at the beginning. In particular, I have said very little explicitly about those objections that might be said to

turn on the question of *who I am*. We saw that Kantian ethics apparently holds that I must regard myself simultaneously in two incompatible ways. In order to regard the moral law as a law of autonomy, I must identify myself with the rational will (or even the idea of a rational will in general). But in order to see it as a source of obligation on me, I must simultaneously regard my will as one which resists reason and needs to be constrained to obey its law. Kant underestimates the problem when he says such things as that I must regard myself from two standpoints, or as belonging to two worlds, or as having inclinations which conflict with the moral law. For all these formulations still permit me to say unproblematically that I am both the giver of the law and the one who obeys it. But the problem is that it looks like I cannot say both these things. For if the law is to be a law of autonomy, then I must regard the moral law as *my* law, given by *my* will. I must therefore identify *my* will as the will which gives the law, and any volition which resists obeying it must be regarded as other than mine, as alien to me. If the law is an expression of my freedom, then *my* will cannot be constrained by it; if it were, then acting according to the law would no be freedom for me, but constraint. Yet this relation of being constrained by the law (of being unfree in relation to it) is precisely the relation in which I must stand to the law if I am to experience it as a *law* for me at all.

… (end of excerpt)

Notes

[1] Citations of Kant's writings will be by volume:page number from *Kants Schriften* in the Akademie Ausgabe (1902-) now published in Berlin by W. deGruyter, abbreviated 'Ak'. The Critique of Pure Reason, however, will be cited as "KrV" by A/B page numbers.