9

Kantian Rationalism: Inescapability, Authority, and Supremacy

DAVID O. BRINK

Kant appears to be the ultimate rationalist about moral psychology.[1] In claiming that moral requirements express categorical imperatives, he defends the existence of objective moral requirements that are part of practical reason and are supposed to have overriding authority. I want to examine and assess different strands in Kant's rationalism. In particular, I believe that in claiming that moral requirements are categorical imperatives Kant commits himself to three distinguishable claims. (*a*) If moral requirements are categorical imperatives, they are objective or inescapable; their application to an agent does not depend on the agent's own contingent inclinations or interests. Let us call this the *inescapability* thesis. (*b*) If moral requirements are categorical imperatives, they are requirements of reason; moral requirements have rational authority such that it is *pro tanto* irrational to fail to act in accordance with them, and this authority is independent of the agent's own aims or interests. Let us call this the *authority* thesis. (*c*) Kant also believes that the categorical character of moral requirements implies that their authority is always overriding. Let us call this the *supremacy* thesis.

Once we distinguish these three aspects of Kantian rationalism, we may not find them equally plausible. In her interesting and provocative article 'Morality as a System of Hypothetical Imperatives' Philippa Foot distinguishes, in effect, between the inescapability and authority theses and argues that only the inescapability thesis is defensible.[2] Though I take

[1] References to Kant are to the Prussian Academy pagination in the following works: *Kritik der reinen Vernunft* (cited as *KrV*) and trans. as *Immanuel Kant's Critique of Pure Reason*, by Norman Kemp Smith (New York: St Martin's, 1963); *Grundlegung der Metaphysik der Sitten* (cited as *G*) and trans. as *Grounding for the Metaphysics of Morals*, by J. Ellington (Indianapolis: Hackett, 1981); *Kritik der praktischen Vernunft* (cited as *KpV*) and trans. as *Critique of Practical Reason*, by L. W. Beck (Indianapolis: Library of Liberal Arts, 1956); *Metaphysik der Sitten* (cited as *M*) and trans. as *The Metaphysics of Morals* in *Kant's Ethical Philosophy*, by J. Ellington (Indianapolis: Hackett, 1983); *Kritik der Urteilskraft* (cited as *KU*) and trans. as *Critique of Judgment* by W. Pluhar (Indianapolis: Hackett, 1987).

[2] *Philosophical Review*, 81 (1972), 305–16; repr. with postscript in Philippa Foot, *Virtues and Vices* (Los Angeles: Univ. of California Press, 1978), 157–73.

Foot's claims seriously, I argue, by contrast, that Kant has a plausible argument from the inescapability of moral requirements to their authority. However, I express scepticism about Kant's arguments for the supremacy thesis. In fact, I believe that Kant may have to recognize a kind of dualism of practical reason between agent-centred and impartial imperatives. Unless this dualism can be resolved, the supremacy thesis must remain doubtful.

1. *The Rational Authority of Morality*

My interest in Kantian rationalism grows out of my attempt to understand and assess different conceptions of the rational authority of morality. It is common to think of morality as both impartial and objective, in particular, as containing various other-regarding duties of co-operation, forbearance, and aid that apply to agents independently of their own aims and interests. Most of us also regard moral obligations as authoritative practical considerations. But heeding these obligations appears sometimes to constrain the agent's pursuit of his own interest or aims. If we associate rationality with the agent's own point of view, we may wonder whether moral conduct is always rationally justifiable. We can capture this tension in common views in terms of a puzzle about the authority of morality.[3]

1. Moral requirements include impartial other-regarding obligations that do not apply to agents in virtue of their own aims or interests.
2. Moral requirements provide agents with overriding reasons for action; necessarily, it is on balance irrational to act contrary to moral requirements.
3. Rational action is action that achieves the agent's aims or promotes her interests.
4. Fulfilling other-regarding obligations need not advance the agent's aims or interests.

(1) articulates one conception of ethical objectivity, according to which moral requirements appear as impartial constraints on conduct that do not apply in virtue of the agent's own aims or interests. For instance, I do not defeat an ascription of obligation to me to help another by pointing out that doing so will serve no aim or interest that I have. (2) implies the weak rationalist thesis that there is always reason to be moral such that contra-moral behaviour is *pro tanto* irrational; but it also expresses the strong rationalist thesis that contra-moral behaviour is always on balance irrational.

[3] I have discussed the puzzle and various solutions elsewhere; see my 'A Puzzle about the Rational Authority of Morality', *Philosophical Perspectives*, 6 (1992), 1–26 and 'Objectivity, Motivation, and Authority in Ethics' (unpublished).

It is one way of attempting to understand the special authority moral considerations seem to have in practical deliberation. (3) expresses a common view of practical rationality, according to which it is instrumental or prudential. Though prudential and instrumental conceptions of rationality are different in significant ways, both represent the rationality of other-regarding conduct as *derivative*. Though no labels seem entirely satisfactory, we might describe this common assumption as the assumption that practical reason is *agent-centred*; by contrast, practical reason is *impartial* if it implies that there is non-derivative reason to engage in other-regarding conduct.[4] Finally, (4) reflects a common assumption about the independence of different people's interests and attitudes, which we might call the *independence assumption*. Though agents often do care about the welfare of others and there are often connections between an agent's own interests and those of others, neither connection holds either universally or necessarily. My aims could be largely self-confined, and my own good can be specified in terms that make no essential reference to the good of others, say, in terms of my own pleasure or the satisfaction of my desires.

Though each element of the puzzle might seem appealing and has appealed to some, not all four claims can be true. In fact, a number of

[4] (*a*) My contrast between agent-centred and impartial conceptions of rationality is different from the contrast, some have drawn, between agent-relative and agent-neutral reasons. Cf. Thomas Nagel, *The View from Nowhere* (New York: Oxford Univ. Press, 1986), 152–3. According to the latter distinction, reasons are agent-relative if their general form involves essential reference to the agent who has them; otherwise, reasons are agent-neutral. Agent-neutral theories are typically understood to be consequentialist, whereas agent-relative theories are quite varied. Prudential and instrumental conceptions of rationality are both agent-relative, though in different ways. A crucial issue as regards the authority of ethics is whether the justification of other-regarding moral conduct is derivative, as both prudential and instrumental conceptions of rationality must claim it is, or whether it is non-derivative. The distinction between agent-centred and impartial conceptions of rationality gets at this issue, whereas the distinction between agent-relative and agent-neutral conceptions of rationality does not. This can be illustrated by considering the view Broad called *self-referential altruism*, according to which an agent has non-derivative reason to benefit others, as well as herself, but the weight or strength of her reasons is a function of the nature of the relationship in which she stands to potential beneficiaries. Cf. C. D. Broad, 'Self and Others', repr. in *Broad's Critical Essays in Moral Psychology*, ed. D. Cheney (London: George Allen & Unwin, 1971), 279–80. Though self-referential altruism is agent-relative, its altruistic or impartial component makes its justification of other-regarding conduct non-derivative in a way that is alien to prudential and instrumental conceptions of rationality. I am here interested in the contrast between the way in which prudential and instrumental conceptions of rationality make the justification of other-regarding conduct derivative and the way in which agent-neutral theories and some agent-relative theories (e.g. self-referential altruism) do not. Though no labels seem entirely satisfactory, I refer to these two approaches as agent-centred and impartial conceptions, respectively. Notice that in so doing we leave it open whether impartiality should take an agent-neutral or agent-relative form. (*b*) We should also note that impartiality, in this sense, need not preclude some forms of partiality; it need not preclude greater concern for oneself and others to whom one stands in special relationships than to comparative strangers. Even agent-neutral interpretations of impartiality try to accommodate some kinds of partiality. And, as self-referential altruism makes plain, some theories that are impartial, in my sense, can recognize partiality at a fairly fundamental level. So the fact that Kant recognizes some kinds of partiality (*M* 451–2) is consistent with my claim that he accepts an impartial conception of practical reason.

influential historical and contemporary views can be seen as responses, perhaps tacit, to this puzzle that reject at least one element of the puzzle on the strength of others. Some *moral relativists* and *minimalists* appeal to (2)–(4) and reject the existence of impartial and objective moral norms asserted in (1); they claim that genuine moral requirements must be relativized to and further the agent's interests or aims in some way.[5] A weak rationalist might resist the strong rationalist thesis in (2). But those who appeal to (1), (3), and (4) to reject (2) typically reject even the weak rationalist claim; *anti-rationalists* deny (2) and claim that failure to act on moral requirements is not necessarily irrational. Others reject the agent-centred assumptions about practical rationality in (3) and defend the existence of *impartial practical reason.*[6] Finally, *metaphysical egoists* reject the independence assumption in (4) and resolve the puzzle by arguing that, properly understood, people's interests are interdependent such that acting on other-regarding moral requirements is a counterfactually reliable way of promoting the agent's own interests.[7]

Kant accepts (1), (2), and (4) and denies (3); he claims that practical reason can be impartial. Foot also accepts (1), but because she accepts (3) and (4), she rejects (2). She is an anti-rationalist; immoral action need not be irrational.[8]

My aim is to understand and assess the Kantian solution to the puzzle about the rational authority of morality. I am interested in a careful and sympathetic interpretation of Kant's texts, especially the *Groundwork*. But because my main interest in Kant derives from my systematic concerns with the authority of morality, I am more interested in the themes and resources of Kantian rationalism than in scholarship, especially those themes and resources that do not presuppose transcendental idealism, in particular, transcendental freedom.

[5] This view is represented by Callicles' claims about natural justice in Plato's *Gorgias* and by that strand of social contract theory---including Epicurus, Hobbes, and Gauthier—that understands the scope, content, and authority of morality in terms of rational agreement. See *Gorgias* 482de, 483ab, 488b–490a; Epicurus, *Kuriai Doxa* 31–8; Thomas Hobbes, *Leviathan*, esp. chs. xiii–xv; and David Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986). A more clearly relativistic version of the view is Gilbert Harman, 'Moral Relativism Defended', *Philosophical Review*, 85 (1975), 3–22.

[6] An important contemporary defence of impartial practical reason is Thomas Nagel, *The Possibility of Altruism* (Princeton: Univ. Press, 1970).

[7] For one such view and a discussion of its classical roots, see my 'Self-Love and Altruism', *Social Philosophy and Policy*, 14 (1997), 122–57.

[8] This is a fair characterization of Foot's view in 'Morality as a System of Hypothetical Imperatives'. However, recently she has changed her view; see Philippa Foot, 'Does Moral Subjectivism Rest on a Mistake?', *Oxford Journal of Legal Studies*, 15 (1995), 1–14. There, in direct opposition to her view in 'Morality as a System of Hypothetical Imperatives', Foot understands moral requirements as requirements of practical reason and rejects agent-centred assumptions about practical reason. She attempts to explain why familiar other-regarding demands are requirements of practical reason by representing them as 'Aristotelian necessities', without the general observance of which social life and its benefits would be difficult if not impossible. I won't explore this suggestion or its adequacy here.

2. *Inescapability without Authority*

Kant, of course, distinguishes between hypothetical and categorical imperatives. He writes

> Now all imperatives command either hypothetically or categorically. The former represent the practical necessity of a possible action as a means of attaining something else that one wants (*will*) (or may possibly want) (*wolle*). The categorical imperative would be one which represented an action as objectively necessary in itself, without reference to another end. (*G* 414)

Here and elsewhere (*KpV* 20–1) Kant claims that hypothetical imperatives are conditional on what an agent wants (or wills).[9] If so, instrumental imperatives are hypothetical imperatives. But he must also think that prudential imperatives are hypothetical.[10] For prudential imperatives presumably represent action as necessary to achieve a distinct end, namely, the agent's happiness or interest. And Kant clearly regards Greek eudaemonist theories as heteronomous and, hence, as containing only hypothetical imperatives (*KpV* 24, 64–5, 109, 111–13). If so, we can understand hypothetical imperatives to be conditional on whether the conduct enjoined promotes the agent's antecedent aims or interests, whereas categorical imperatives are not.

Following Foot, we might identify two distinguishable senses in which imperatives might be categorical. In one sense, imperatives are categorical just in case they *apply* to people independently of their aims or interests; if so, we might say they express *categorical norms*. Imperatives are categorical in another sense just in case they provide those to whom they apply with *reasons for action* independently of their aims or interests; if so, we might say they generate *categorical reasons*.

Famously, Kant claims that moral requirements express categorical, rather than hypothetical, imperatives (*G* 416, 425). Presumably, he thinks moral requirements are categorical imperatives in both senses; they express categorical norms that generate categorical reasons. But once we distinguish clearly between inescapability and authority, we might accept inescapability without authority; we might agree that moral requirements express categorical norms but deny that they generate categorical reasons.

Various systems of norms appear to express categorical norms whose authority, however, is not (obviously) categorical. For instance, it is natural,

[9] *Wollen* can be translated as 'to want' or as 'to will'. Kant does not think that every object of one's desire is an object of one's will; to will something is to have one's choice in some way determined by practical reason (*G* 412, 427, 446). If so, it is possible to read this passage (*G* 414) as saying that hypothetical imperatives represent as practically necessary actions that secure means or necessary conditions to what the agent wills, and not merely to what she wants. I will discuss the significance of this interpretation of Kant's remarks about hypothetical imperatives later (sect. 10).

[10] Indeed, Kant appears to equate all empirical motivation with self-love (*KpV* 22, 34).

and I think plausible, to view legal and occupational requirements this way. But legal and occupational requirements are often morally or prudentially important. It is, I think, because she wants to examine morality's relation to something agreed to be fairly unimportant that Foot explains her assessment of Kant with an analogy between morality and etiquette. Indeed, rules of etiquette often overlap with requirements of morality or prudence. The focus on etiquette must be on those rules of etiquette that seem especially unimportant morally or prudentially, for instance, rules requiring that invitations addressed in the third person be answered in the third person. She invites us to compare morality and *mere* etiquette.

According to Foot, both rules of (mere) etiquette and moral requirements are inescapable; they express categorical norms. The moral duty to help others in distress, when you can do so at little cost to yourself, does not fail to apply to you—we do not withdraw our ascription of obligation to you—just because you are indifferent to your neighbour's suffering and in a hurry to read your mail, as would be the case if it was a hypothetical norm. In the same way, rules against replying to a third-person invitation in the first person don't fail to apply to you—we don't take back our ascriptions of duties of etiquette to you—just because you think etiquette is silly or you have a desire to annoy your host, as would be the case if rules of etiquette stated hypothetical norms.

But rules of etiquette seem to lack *authority*; they appear to generate hypothetical, not categorical reasons. On this view, rules of etiquette may state categorical norms, but failure to observe these norms does not seem irrational unless this in some way undermines the agent's interests or aims. Here too moral requirements may seem on a par with requirements of etiquette. If the independence assumption is correct, obligations of forbearance, mutual aid, and justice need further no aims or interests of the agent. Though we do not need to withdraw the ascription of obligation in such cases, perhaps we should allow that immoral conduct in such a case is not irrational. This is Foot's view.

> [I]t is supposed [by Kant and others] that moral considerations necessarily give reasons for acting to any man. The difficulty is, of course, to defend this proposition which is more often repeated than explained. . . . The fact is that the man who rejects morality because he sees no reason to obey its rules can be convicted of villainy but not of inconsistency. Nor will his action necessarily be irrational. Irrational actions are those in which a man in some way defeats his own purposes, doing what is calculated to be disadvantageous or to frustrate his ends. Immorality does not *necessarily* involve any such thing.[11]

[11] 'Morality as a System of Hypothetical Imperatives', 161–2.

So Foot accepts the inescapability thesis but rejects the authority thesis. Because she assumes that practical reason is agent-centred, she finds the authority thesis mysterious. In fact, she thinks that Kantians mistakenly appeal to the inescapability thesis to support the authority thesis.[12]

We can now see the sense in which Foot thinks morality is a system of hypothetical imperatives. For whereas she does think that moral requirements, like requirements of etiquette, express categorical norms, she thinks that they, also like requirements of etiquette, generate hypothetical, rather than categorical reasons. Because Kant would not want to regard requirements of etiquette as categorical imperatives, this shows that the basic sense of categoricity is that in which, on her view, moral requirements are not categorical imperatives.

3. *Authority*

On Foot's version of anti-rationalism, the authority, but not the scope or content, of morality depends on the aims or interests of agents. But the analogy between morals and manners as yet provides no explanation of the common belief that morality has a special authority. On one reading of her claims, Foot seems to say that the special authority of morality is just an illusion—an artefact of moral education. But she also claims that the authority of morality does not require categorical imperatives. The part of morality most obviously threatened by agent-centred rationality is other-regarding morality, for it is obligations of forbearance, mutual aid, and justice that are most likely to frustrate the agent's own interests and desires. But Foot thinks that people can be and are committed to the interests of other people and common causes, as morality requires, and that these social interests and sentiments ensure that they do act as morality requires and that they have (hypothetical) reason to do so.

> This conclusion may, as I said, appear dangerous and subversive of morality. We are apt to panic at the thought that we ourselves, or other people, might stop caring about the things we do care about, and we feel that the categorical imperative gives us some control over the situation. But it is interesting that the people of Leningrad were not struck by the thought that only the *contingent* fact that other citizens shared their loyalty and devotion to the city stood between them and the Germans during the terrible years of the siege. Perhaps we should be less troubled than we are by fear of defection from the moral cause . . .[13]

If we rely on purely instrumental assumptions about rationality, we can establish the authority of other-regarding moral requirements to those who

[12] Ibid. 162. [13] Ibid. 167.

have suitable other-regarding attitudes. Especially if such attitudes are strong and widespread, this may seem an adequate account of the authority of other-regarding morality.

But the instrumental justification of morality appeals to other-regarding attitudes without grounding them; as a result, it seems unable to explain why those who lack these attitudes should cultivate them or why those who do have them should maintain them. This is presumably part of what Kant has in mind when he objects to accounts of moral motivation that make it dependent on contingent and variable inclination; he concludes that the authority of morality must depend on features of rational agents as such (*G* 389–90, 397–400, 427, 442–3; *KpV* 21, 24–6, 36).

A more traditional defence of morality is to argue that the demands of morality and enlightened self-interest coincide. The main lines of this story are familiar enough. Much of impartial other-regarding morality involves norms of co-operation (e.g. fidelity and fair play), forbearance, and aid. Each individual has an interest in the fruits of interaction conducted according to these norms. Though it might be desirable to reap the benefits of other people's compliance with norms of forbearance and co-operation without incurring the burdens of one's own, the opportunities to do this are infrequent. Non-compliance is generally detectable, and others won't be forbearing and co-operative toward those who are known to be non-compliant. For this reason, compliance is typically necessary to enjoy the benefits of others' continued compliance. Moreover, because each has an interest in others' co-operation and restraint, communities will tend to reinforce compliant behaviour and discourage non-compliant behaviour. If so, compliance is often necessary to avoid such social sanctions. Whereas non-compliance secures short-term benefits that compliance does not, compliance typically secures greater long-term benefits than non-compliance. In this way, compliance with other-regarding norms of co-operation, forbearance, and aid might be claimed to further the agent's interests. In so far as this is true, the rational egoist can ground other-regarding sentiments and explain why those who do not have them should cultivate them and those who do have them should maintain them.

However, as long as we rely on pre-theoretical understandings of self-interest, the coincidence between other-regarding morality and enlightened self-interest, on this view, must remain imperfect. Sometimes non-compliance would go undetected; and even where non-compliance is detected, the benefits of non-compliance sometimes outweigh the costs of being excluded from future co-operative interaction. Moreover, even if the coincidence between morality and self-interest were extensionally adequate, it would be counterfactually fragile. On this justification of compliance with other-regarding norms, compliance involves costs, as well as benefits; it must

remain a second-best option, behind undetected non-compliance, in which one enjoys the benefits of others' compliance without the costs of one's own. But then if one had some way of ensuring that one's own non-compliance would go undetected—for instance, one had sole access to the ring of Gyges—one could enjoy the benefits of the compliance of others without the burdens of one's own, and one would have no reason to be compliant. The imperfect coincidence of morality and self-interest, which the independence assumption ensures, implies that immorality need not always be irrational. And this is presumably part of what troubles Kant about accounts of moral motivation that make it dependent on the agent's own happiness (*G* 425–7, 442–3; *KpV* 20–1, 24, 64–5, 109, 111–13).

None the less, anti-rationalists may find this acceptable. It allows us to explain why everyone has some stake in morality, and why people generally have reason to behave morally, but it insists that immoral action is not always irrational. As long as we have not tied the scope and content of morality to its rationality, we can reproach the immoralist with immorality. What is lost if we cannot also reproach him with irrationality?

Anti-rationalism would be more satisfactory if morality and rationality were two independent but co-ordinate perspectives. For then it might seem to be an open question whether an agent should side with morality or rationality when they conflict. But in the present context, practical rationality is not just one standard or perspective among others, with no obviously privileged position; it should be understood to concern whatever fundamentally matters in practical deliberation or whatever it is ultimately reasonable to do. So, for example, if I have doubts about whether I have reason to act on a particular norm, I should be interpreted as having doubts about whether that is a norm of practical rationality, rather than as having doubts about rationality. But then anti-rationalism has the potentially unsettling consequence that morality need not always have authority in our deliberations.

We might ask why Foot and other anti-rationalists assume that practical reason must be agent-centred. One reason appeals to apparent connections between practical rationality and motivation. It seems plausible that judgements of practical rationality normally give rise to motivation. If recognition of reasons for action normally motivates, this may seem to require that reasons for actions be grounded in antecedently motivational facts or states of the agent, such as her interests or desires.[14] But we can respect this link between practical judgement and motivation without supposing that

[14] Cf. Bernard Williams, 'Internal and External Reasons', repr. in his *Moral Luck* (Cambridge: Univ. Press, 1981). Williams argues from a somewhat stronger assumption about the link between recognizing the truth of practical judgements and motivations to a kind of instrumental conception of rationality that grounds reasons for action in the agent's antecedent pro-attitudes.

rationality is constrained by what is antecedently motivational or that motivation might be produced by cognitive states alone. We can accept the common view that motivation requires a desire or pro-attitude. On this view, intentional action is the product of representational states, such as belief, which aim to conform to the world, and practical states or pro-attitudes, such as desires, which aim to make the world conform to them. As such, normative motivation, like all motivation, requires pro-attitudes. But, other things being equal, our motivational states track our beliefs about what we have reason to do. Given that practical reason concerns whatever fundamentally matters in practical reasoning, we should expect results of practical deliberation normally to affect one's motivational set.[15] Believing it is best that things be a certain way normally produces a desire or pro-attitude to make things be that way.[16] If so, motivation can be consequential on practical rationality, not the other way around (*G* 460–1). So if there are good arguments for thinking that practical rationality can be impartial, the connection between rationality and motivation is no obstacle to rationalism.[17]

4. *Kantian Inescapability*

Foot complains that Kantians appeal to the inescapability thesis to support the authority thesis. If these are independent theses, this is a mistake. But even if they are distinct theses, they need not be independent. In fact, Kant believes that the way in which moral requirements are inescapable explains their authority. To explain Kant's argument from inescapability to authority, we will need to examine his views about the Categorical Imperative at some length.

Kant often claims that common-sense morality presupposes that moral requirements, or at least their foundations, must be justifiable a priori and not on the basis of experience (*G* 388–9, 410). There are at least two different claims here.

[15] Action based on *moral* feeling is not heteronomous; moral feelings are consequential on recognizing the authority of pure practical reason (*G* 401 n.; *KpV* 24–5, 75–82).

[16] I defend this as a systematic claim at greater length in 'Objectivity, Motivation, and Authority in Ethics', sects. 1–4. Allison says that Kant explicitly dismisses a similar claim about how practical reason might motivate; see Henry Allison, *Kant's Theory of Freedom* (New York: Cambridge Univ. Press, 1990), 122–3. Allison does not say enough about the view he thinks Kant dismisses or his grounds for dismissing it for me to evaluate his (Allison's) claim. I do not think Kant needs to or even does reject the picture I have sketched of the relation between judgements of practical reason and motivation.

[17] My answer to this motivational challenge to the possibility of impartial practical reason is similar to some claims made by Christine Korsgaard in 'Skepticism about Practical Reason', *Journal of Philosophy*, 83 (1986), esp. 21–3. But, whereas she seems to think that the motivational capacity of judgements of practical reason requires a prior desire to be rational, I think that such a desire need play no role in the production of motivation.

One claim concerns particular duties (e.g. Sam's duty to fulfil his contractual obligation to sell his widgets to Ben). Though an agent's particular duties do depend upon certain contingent circumstances, such as his past actions (e.g. the fact that Sam signed a contract to sell his widgets to Ben), they do not depend upon contingent facts about the agent's interests and desires at the time of action. In particular, an agent cannot defeat a claim that he has a duty simply by pleading disinclination or disinterest. If so, we can know an agent's particular duties independently of knowing these empirical facts about him. We have already accepted this idea in accepting the inescapability thesis.

But this does not yet establish the strong claim that morality is justifiable a priori; for this to be true, morality must in some way be independent of all contingent empirical facts about agents. Kant supposes that particular, concrete duties are established by the application of quite general moral principles, such as the requirement to treat others as ends and not merely as means, to contingent empirical circumstances (e.g. the circumstances of Sam's promise to Ben) (*M* 217). Moreover, he believes that these more abstract principles must be independent not only of the agent's particular interests and desires at the time of acting but independent of all contingent facts about the agent and his circumstances; they must depend upon general features of moral agents (*G* 408). Indeed, this is presumably the difference between the Categorical Imperative, in its various formulations, and particular categorical imperatives. Whereas the Categorical Imperative is supposed to be justified independently of empirical facts, particular categorical imperatives result from the application of the Categorical Imperative to particular circumstances. If these abstract principles are treated as the *ground* of the more concrete, particular duties, then we can understand why Kant would believe that there is a sense in which even these more particular duties apply to agents independently of contingent facts about themselves and their circumstances; their ground is so independent and is, therefore, knowable independently of knowledge of these contingent facts (*G* 389).[18]

But what would it be for moral duties to apply to agents in virtue of general features of moral agents? To be a moral agent is presumably to be *responsible*; only responsible agents are properly praised and blamed, because only they can be held accountable for their actions. Non-responsible agents, such as brutes and small children, appear to act on their strongest desires or, if they deliberate, to deliberate only about the instrumental means to the satisfaction of their desires. By contrast, responsible agents, we assume,

[18] Cf. Allen Buchanan, 'Categorical Imperatives and Moral Principles', *Philosophical Studies*, 31 (1977), 249–60.

can distinguish between the *intensity* and *authority* of their desires and deliberate about the appropriateness of their desires and aims (*G* 396, 437, 448, 452; *KrV* A534/B562, A553–4/B581–2, A802/B830; *KpV* 61–2, 87; *M* 213, 391–2; *KU* 442–3).[19] Whether consciously or merely implicitly, a responsible agent can and does assess the desirability of her impulses, and her choices reflect these deliberations about her desires. If so, capacities for practical deliberation—formulating, assessing, revising, choosing, and implementing projects and goals—are essential to being an agent. Because moral agents are essentially reasoning and deliberative creatures, moral requirements must apply to rational agents as such if they are to apply to moral agents as such (*G* 408, 412, 423, 425–7; *KpV* 20–1, 29–30).[20] As Kant writes in the *Metaphysics of Morals*, 'They [requirements of morality] command everyone without regard to his inclinations, solely because and insofar as he is free and has practical reason' (*M* 216). If moral requirements apply to people in so far as they are rational beings and not in so far as they have contingent inclinations and interests, then we can see why they must be expressed by categorical, rather than hypothetical, imperatives. For hypothetical imperatives do, and categorical imperatives do not, apply to us in virtue of our contingent interests and inclinations.

Of course, Kant thinks we cannot know whether there are, in fact, any moral requirements until we can show that moral agents are free and responsible, a task he attempts to complete, among other places, in Section 3 of the *Groundwork*. There he argues that (*a*) freedom requires the capacity for determination by reasons, not one's kinaesthetically strongest desires (*G* 446–8, 457, 459–60; cf. *KrV* A534/B562, A553–4/B581–2, A802/B830; *G* 396, 437; *KpV* 61–2, 72, 87; *M* 213, 216, 391–2; *KU* 442–3), that (*b*) this capacity requires transcendental freedom (*G* 450–3, 455–7; cf. *KrV* A534/B562; *KpV* 3–4, 43, 46, 94–106), and that (*c*) transcendental freedom is compatible with what we can and do know (*G* 450–3, 455–7; cf. *KrV* A538–58/B566–86; *KpV* 3–6, 47–9, 54, 95–106, 114, 133). These claims raise complex issues that I cannot address properly here. But my belief is that whereas (*a*) is plausible, (*b*) is not; responsibility requires deliberative self-government, but deliberative self-government does not require

[19] Cf. the use Irwin makes of self-consciousness in order to reconstruct Kant's views of rational agency; see Terence Irwin, 'Morality and Personality: Kant and Green', in A. Wood (ed.), *Self and Nature in Kant's Philosophy* (Ithaca, NY: Cornell Univ. Press, 1984), 31–56. Of course, the importance of such capacities of practical deliberation to agency is not peculiar to Kant. Cf. Plato, *Republic* 437e–442c; Aristotle, *De Anima* 2. 2 and *Nicomachean Ethics* 1102b13–1103a3, 1111b5–1113a14; Cicero, *De Officiis* 1. 11; Bishop Butler, *Fifteen Sermons*, ii. 13; Thomas Reid, *Essays on the Active Powers of the Human Mind*, ii. 2; T. H. Green, *Prolegomena to Ethics*, sects. 85–158.

[20] Morality applies to rational beings as such, that is, to beings in so far as they are rational. So moral duties apply to agents who have only a rational nature (e.g. gods) and to rational agents who also have an empirical nature (e.g. humans); but Kant thinks that moral duties appear *as imperatives* only to the latter class of agents (*G* 414, 455 and *KpV* 20, 32, 82; but see *KpV* 81).

transcendental freedom.[21] Transcendental freedom seems neither necessary nor sufficient for responsibility. It seems unnecessary, because responsible actions do not require choices that lie outside a causal nexus; they require only that choices not be determined by the agent's inclinations, independently of his deliberations. It seems insufficient, because an action's determination by aspects of an agent that are in principle unknowable (an agent *qua noumenon*) cannot explain why the agent is responsible for the action. If so, a defence of Kantian rationalism requires an account of our capacities for deliberative self-government that does not presuppose libertarianism or noumenal determination of the will. Any such account must explain our ability to recognize and respond to practical reasons in naturalistic terms. Certainly, such an account is required if we are to develop a Kantian moral psychology that does not presuppose transcendental idealism, in particular, transcendental freedom. In what follows I will assume that some naturalistic account of deliberative self-governance is possible.[22]

5. *The Categorical Imperative*

This understanding of why moral requirements must be represented by categorical imperatives leads Kant to the first of his three main formulations of the Categorical Imperative—the formula of Universality. If moral requirements are not to be based on empirical conditions, it seems they must be universal or universalizable. For an agent's action to be morally permissible, Kant argues, it must be possible for her to will that her maxims, or the subjective principles of her action (*G* 401 n., 421 n.), become a universal law of nature.

F1 Act only on those maxims that you can at the same time will to be a universal law (*G* 421; *KpV* 30, 69; *M* 225–6).

This may sound like a hopelessly abstract claim, as Foot and others seem to think. But Kant offers two kinds of help in understanding F1. First, he offers examples of moral issues to which he then applies F1. He also links F1 with two other main formulations of the Categorical Imperative.[23]

[21] Contrast Allison, *Kant's Theory of Freedom*, esp. 35–41.

[22] It is instructive to see the role that capacities for deliberative self-government play in interesting versions of compatibilism; see e.g. Harry Frankfurt, 'Freedom of the Will and the Concept of a Person', *Journal of Philosophy*, 68 (1971), 5–20 and Gary Watson, 'Free Agency', *Journal of Philosophy*, 82 (1975), 205–20.

[23] In recognizing three main formulations of the Categorical Imperative, I do not distinguish as many formulations as others have. See e.g. the commentary by Paton in Immanuel Kant, *Groundwork for the Metaphysics of Morals*, trans. H. J. Paton (New York: Harper & Row, 1956) and Bruce Aune, *Kant's Theory of Morals* (Princeton: Univ. Press, 1979), 111–20. My more coarse-grained division is not unfamiliar and seems adequate for present purposes.

He claims that F1 implies a second formulation—the formula of Humanity (*G* 437).

F2 Treat humanity, whether yourself or any other rational agent, always as an end in itself and never merely as a means (*G* 429; *KpV* 87, 131; *M* 462).

And F2 is supposed to imply a third formulation—the formula of Autonomy (*G* 438).

F3 Every rational being should be regarded as an autonomous legislator in a kingdom of ends (*G* 431–3, 438).

Of course, F2 and F3 themselves require interpretation, but the fact that Kant identifies F1, F2, and F3 (*G* 436) may help in interpreting any one of them.

6. *The Formula of Universality*

What does it mean to say that an agent must be able to will his maxim to become a universal law?[24] What sort of universality or universalizability does F1 require? Kant claims that actions can violate F1 in one of two ways:

> Some actions are so constituted that their maxims cannot without contradiction even be thought as a universal law of nature, much less be willed as what should become one. In the case of others this internal impossibility is indeed not found, but there is still no possibility of willing that their maxims should be raised to the universality of a law of nature, because such a will would contradict itself. (*G* 424)

Thus, an action violates F1 if its maxim is such that (*a*) it is impossible or inconceivable for everyone to act on it, or (*b*) its universalization, though conceivable, would reveal some contradiction in the agent's will.

Kant thinks that the case of false promises involves maxims whose universalization is inconceivable (*G* 403, 422; cf. *KpV* 27). I cannot will that my maxim of keeping promises only when it suits my interests be universal, because if everyone acted on this maxim, promises would often not be kept and the general level of trust necessary to sustain the practice of promising would not obtain. Thus, a general practice of false promising would prove self-defeating.

[24] Though quite different, my own reading of F1 has benefited from the discussions by Onora Nell (now O'Neill), *Acting on Principle: An Essay in Kantian Ethics* (New York: Columbia Univ. Press, 1975); Onora O'Neill, 'Consistency in Action', repr. in her *Constructions of Reason* (New York: Cambridge Univ. Press, 1989), 81–104; and Christine Korsgaard, 'Kant's Formula of Universal Law', repr. in her *Creating the Kingdom of Ends* (Cambridge: Univ. Press, 1996), 77–105.

However, this seems not to show that the universalization of a maxim of false promising is inconceivable. What it shows is that the practice of promising could not be sustained if everyone were to make false promises. But this just shows a certain consequence of universal false promising; there is nothing inconceivable about the resulting state of affairs. Moreover, this is a consequence not of universal false promising but of universal false promising *only if* each recognizes the promises of others as false. But then there seems nothing *self*-contradictory about universal false promising.

Moreover, this kind of inconceivability, if that is what it is, had better not be a sufficient condition of violating F1, because there appear to be many perfectly innocent activities that are not universalizable in this sense. No one could will to perform any activity that is part of some larger division of labour—for instance, practising philosophy or selling but not producing widgets—because, if everyone performed that one activity, no one would perform the other activities in the division of labour necessary to produce the products that sustain the division of labour.[25]

Fortunately, the conceivability interpretation of F1 appears not to be basic.[26] Some maxims whose universalization is conceivable (and presumably some Kant thinks are not) cannot be *willed* to be a universal law.

[25] Nell seems to think that this worry does not apply to Kant's conception test (*Acting on Principle*, 78–9). It may be that I am assuming, as she is not at this point, that the activity is part of a larger division of labour that must be sustained if the agent is to act on her maxim (cf. ibid. 68, 79). Kant might avoid this problem if he were to claim that the universalization of maxims could only be assessed *jointly*. But I'm not quite sure how this would go, and Kant appears to think the universalization of maxims can be assessed individually.

[26] This conclusion would have revisionary implications if we accepted, as many commentators do, Kant's suggestion that maxims whose universalization is inconceivable violate *perfect* duties whereas those whose universalization involves a contradiction in the will violate only *imperfect* duties (*G* 424). However, this suggestion had better not be Kant's considered view. Violation of the conceivability test is neither necessary nor sufficient for breach of a perfect duty. As I've claimed, some perfectly innocent activities that are components of larger divisions of labour—that violate neither perfect nor imperfect duties—are such that the universalization of their maxims appears to be, in Kant's sense, inconceivable. Moreover, many maxims whose universalization is conceivable but would, on Kant's view, involve a contradiction in the agent's will do violate perfect duties. For example, duties of mutual aid can only be established by the contradiction in the will test; universalization of the kind of resolute self-reliance that denies duties of mutual aid is perfectly conceivable (*G* 423). But, even if some duties of mutual aid, such as giving to charity, involve imperfect duties, others involve perfect duties, such as the duty to rescue a drowning child when this can be done with little cost or risk to the agent. Similar remarks could be made about perfect duties of forbearance, such as duties not to torture the infirm. The difference between perfect and imperfect duties, therefore, had better be picked out by something other than these two interpretations of F1. Fortunately, there appears to be a more straightforward way to distinguish between perfect and imperfect duties. We should locate this distinction not in different kinds or grounds of duty but rather in the *content* of one's duties and maxims. On the contradiction in the will test, the question is whether maxims can be willed to be a universal law. If not, it is impermissible to act on the maxim ($\sim U(M) \rightarrow \sim P(M)$); if so, it is permissible to act on the maxim ($U(M) \rightarrow P(M)$). As long as we accept a common correlativity principle, according to which a course of action is impermissible just in case it is obligatory not to do it ($\sim P(a) \equiv O(\sim a)$), it follows that if the contradictory of one's maxim cannot be universalized, then acting on it is obligatory ($\sim U(\sim M) \rightarrow O(M)$). Here O(M) specifies an obligation or duty to act on one's maxim. When M specifies that a

On one interpretation, a contradiction in the will would involve willing both P and ~P; this would make F1 a kind of practical analogue of the principle of non-contradiction. But this isn't involved, even in the case of false promises. In that case the agent wills that he take advantage of others' good faith. It's true that his ability to realize the object of his will presupposes others' good faith, and this presupposes that promises are in general kept. If, as Kant claims, he who wills the end, in so far as he is rational, also wills means and necessary conditions to the attainment of his end (*G* 417), then the agent also wills that the practice of promise-keeping continue. And this aim is undermined if everyone acts on his maxim and each recognizes that others are breaking their promises.[27] But the universalization is not part of his will; it's a constraint on acceptable willings that Kant introduces. So we don't have any formal contradiction in his will. He does not will both P and ~P. He wills P (that the practice of promise-keeping continue so that he may take advantage of it) and it's true that *if everyone observes his maxim and recognizes that others do . . . then ~P* (if everyone observes his maxim . . . then the practice of promise-keeping discontinues). For there to be a contradiction in his will of this sort, he would have to will the consequent of this conditional. Whereas he may believe that the conditional is true, I don't see any reason to suppose that he wills the consequent or, for that matter, the conditional or its antecedent.

A more common interpretation of F1 and consistency in one's will results if we ask if we can consistently accept the consequences of everyone acting with our motives. Kant suggests this reading of F1 in a preliminary discussion of false promising.

> The most direct and infallible way, however, to answer the question as to whether a lying promise accords with duty is to ask myself whether I would really be *content* if my maxim (of extricating myself from difficulty by means of a false promise) were to hold as a universal law for myself as well as for others . . . (*G* 403; emphasis added)

certain sort of action always be done, O(M) expresses a perfect obligation or duty, and when M specifies of a certain sort of action that it need not always be done but that it must sometimes be done (when being at the discretion of the agent), then O(M) expresses an imperfect obligation. This suggestion about how Kant can and should draw the distinction between perfect and imperfect duties is compatible, I believe, with his suggestion in *The Metaphysics of Morals* that we understand the distinction as one between required actions and required ends (*M* 390). At another point in *The Metaphysics of Morals*, Kant suggests that we understand the distinction as one between duties that can or should be enforced by external sanction and those that cannot (*M* 383). This last account of the distinction between perfect and imperfect duties appears to be orthogonal to the others.

[27] This is somewhat similar to Korsgaard's favoured interpretation of the contradiction in conception test, which she calls 'the practical contradiction interpretation'; see Korsgaard, 'Kant's Formula of Universal Law'. But in addition to (other) problems it faces, which I discuss in the text, it seems clearly to involve a contradiction in the will. If so, it's hard to see how this could be a good interpretation of the contradiction in conception test, if only because it would make it unclear how Kant draws the distinction between contradictions in conception and contradictions in the will.

And, later, in discussing the fourth example involving the duty of mutual aid, Kant writes

> A fourth man finds things going well for himself but sees others (whom he could help) struggling with great hardships; and he thinks: what does it matter to me? Let everyone be as happy as Heaven wills or as he can make himself; I shall take nothing from him nor even envy him; but I have no desire to contribute anything to his well-being or to his assistance when in need. . . . [E]ven though it is possible that a universal law of nature could subsist in accordance with that maxim, still it is impossible to will that such a principle should hold everywhere as a law of nature. For a will that resolved in this way would contradict itself, inasmuch as cases might often arise in which one would have *need* of the love and sympathy of others and in which he would deprive himself, by such a law of nature springing from his own will, of all hope of the aid he *wants* for himself. (423; emphasis added)

These passages suggest an understanding of the universalization required by F1 that makes it out to be very much like the golden rule. Can you accept the consequences of everyone acting on your principles? If so, you may act on them; if not, you may not; and if the contradictory of your maxim cannot be universalized, acting on it is obligatory.

There are stronger and weaker interpretations of F1 depending on the range of consequences one must consider in universalizing. On one reading, which we might call *empirical* universalization, I must ask whether I can accept what would be the actual or probable consequences of everyone's acting on my maxim. Whereas on the other, stronger reading, which we might call *counterfactual* universalization, I must ask whether I can accept the consequences of everyone's acting on my maxim in all (epistemically) possible circumstances or worlds.[28] The difference between the two readings is easily brought out in connection with Kant's example involving mutual aid. If my own talents and resources are secure (e.g. I have a large and diversified investment portfolio), then I may have no difficulty accepting the consequences of the empirical universalization of my individualist maxim, because it may well be safe to assume that I will never be in need of help from others. However, it's much harder for me to accept the consequences of the counterfactual universalization of my individualist maxim, for there surely are possible worlds in which I lose my talents and resources or never

[28] Just as the weaker reading asks me to consider the probable consequences of everyone's acting on my maxim, the stronger reading should perhaps ask me to consider the consequences of everyone's acting on my maxim in all epistemically possible worlds. Certain features (e.g. my gender or my race) may be essential to me—if I have that feature, I have it in all (metaphysically) possible worlds in which I exist—yet I can conceive of not having that feature—for instance, I can conceive of discovering that, despite appearances, I do not in fact have that feature. A stronger version of universalization would require me to assess the consequences of everyone's acting on my maxim even in (epistemically possible) worlds in which I exist without these essential features.

had them in the first place. In these worlds I may well want assistance; if so, I cannot accept the counterfactual consequences of the universalization of my individualist maxim.[29]

The fact that Kant thinks the individualist maxim cannot be universalized is some evidence that he is concerned with counterfactual, and not merely empirical, universalization. Moreover, counterfactual universalization better makes duty independent of empirical conditions than does empirical universalization. But even counterfactual universalization is too weak. For counterfactual universalization, like empirical universalization, requires only consistency in one's attitudes, even if it requires consistency across a larger range of possible worlds. Like the golden rule, counterfactual universalization asks what consequences one can accept, and this must ultimately be a contingent psychological matter. Perhaps few of us could accept the consequences of everyone's acting on our individualist maxim in those (perhaps merely possible) circumstances in which we are destitute. But surely it's possible for someone—the *resolute* individualist—to accept even these consequences.[30] If so, and if we interpret F1 as requiring only counterfactual universalization, then the resolute individualist has no duty of mutual aid.

It is, I think, for this sort of reason that many readers have found Kant's discussion of the fourth example unsatisfactory and have concluded that the Universality formula is a formal test of consistency that has no determinate content. This is Hegel's 'empty formalism' charge.[31] But we have good reason to wonder whether counterfactual universalization is the best interpretation of the Universality formula. Kant wants moral duty or its ground to be independent of *all* desires and interests; moral duty is supposed to depend only upon features of rational beings as such. In fact, this is why he contrasts F1 with the golden rule (430 n.). The golden rule says 'Do unto others as you would have them do unto you.' The most natural interpretation of this claim is that it requires only the sort of role reversal test that we saw counterfactual universalization represents ('How would you like it if someone did that to you?'). But then the golden rule, like

[29] There are interesting and difficult issues here about how to weigh and combine one's preferences among possible worlds similar to issues about how to weigh and combine the claims of different persons within a possible world. Should I act as if each world were equiprobable and maximize expected average value over worlds, should I restrict myself to pair-wise comparisons of worlds, or should I employ some other method?

[30] Wolff notes this objection to his interpretation of the universal law formula but appears to conclude that this is a problem for Kant, not reason to look for a better interpretation. See Robert Paul Wolff, *The Autonomy of Reason* (New York: Harper & Row, 1973), 170–1. Cf. Hare's discussion of the 'fanatic' in R. M. Hare, *Freedom and Reason* (New York: Oxford Univ. Press, 1963), ch. 9.

[31] G. W. F. Hegel, *The Philosophy of Right*, trans. T. M. Knox (Oxford: Clarendon Press, 1952), sect. 135; cf. J. S. Mill, *Utilitarianism* (Indianapolis: Hackett, 1979), i. 4; Henry Sidgwick, *The Methods of Ethics*, 7th edn. (Indianapolis: Hackett, 1981), 389 n.; and C. D. Broad, *Five Types of Ethical Theory* (London: Routledge & Kegan Paul, 1930), 130.

counterfactual universalization, makes one's moral duties hostage to one's antecedent desires in a way Kant clearly wants to avoid.

How then should we interpret F1? Kant thinks that our duties must be determined by features of us as moral and, hence, rational agents (*G* 408, 412, 425–7, 432, 442; *KpV* 32). So we should interpret F1 as asking what rational beings can consistently will. But this claim is ambiguous. It might be interpreted as asking what *rational beings*—that is, *someone who is rational*—can consistently will. This test can depend on the contingent interests and desires possessed by rational beings, and so counterfactual universalization is one way of articulating it. Alternatively, F1 might be interpreted as asking the different question about what *rational beings as such*—that is, *someone in so far as she is rational*—can will. On this interpretation, F1 asks what we can will, not in so far as we have particular, contingent wants and interests, but what we can will in so far as we are rational beings (*KpV* 29–32, 43). If we distinguish between the will of an *impurely* rational agent—an agent in so far as she has contingent interests and desires—and the will of a *purely* rational agent—an agent solely in so far as she is rational—we might say that this test appeals to the will of a purely rational agent.[32] This seems to be the correct way to interpret the idea that our duties should depend only on features of us as moral and, hence, rational agents (*G* 426–7).

This interpretation has some interesting implications. On this interpretation, I ask whether—in so far as I am a rational being—I can consistently will that my maxim be a universal law. Rational beings are different from one another in countless ways, but not just in so far as they are rational. Different maxims will survive counterfactual universalization depending on the contingent interests and desires of the rational agent who tries to universalize. Not so under this interpretation. Because all agents are alike in so far as they are rational, the results of this sort of test do not depend on who performs it (*G* 427; *KpV* 20–1). Nor is it clear that universalization, as distinct from universality, is essential to F1. Universalizability is a way of counteracting the influence of certain contingent factors in the determination of moral requirements (*G* 424). Our worries about counterfactual universalization suggest that it is an inadequate remedy; our interpretation

[32] This distinction between purely and impurely rational beings should not be confused with Kant's own distinction between infinitely and finitely rational beings; finitely rational beings are rational beings with an empirical nature, whereas infinitely rational beings (e.g. gods) do not have an empirical nature (*KpV* 32, 82). Whereas Kant's distinction separates rational beings into disjoint classes, my distinction does not; it views rational beings under two different aspects: in so far as they have an empirical nature and solely in so far as they are rational. Infinitely rational beings necessarily will things as purely rational beings, and only finitely rational beings can will things as impurely rational beings. But finitely rational beings, as well as infinitely rational beings, can will things as purely rational beings, because this is to will something in so far as one is rational. Indeed, much of my discussion focuses on finitely rational beings *qua* purely rational beings.

of F1 shows that it is unnecessary. Kant secures the independence of duty from the relevant contingent factors by focusing on the will of a purely rational being. We need only ask what a rational being would will, *qua* rational being; it shouldn't matter whether you or I ask the question, and we shouldn't need to ask what if everyone did that.[33]

But we may wonder whether there *is* anything that a purely rational being would will. I can understand what it is for a rational being to will or choose various actions and outcomes on the basis of her interests and preferences. But what would an agent stripped of all such interests and preferences will or choose? It may seem that there is no basis left on which to will or choose. This may be another ground for the 'empty formalism' charge.

7. *Connecting the Formulas*

Kant does not agree. Among other things, he thinks that we can get from the idea of what someone would want or will in so far as she was rational (F1) to F2 and F3 (*G* 429, 432–3, 436). Kant thinks that the one thing that a purely rational being would will or choose for its own sake is *rational agency* (*G* 427–9). It seems reasonable that in so far as one is a rational agent one will value the exercise of rational agency. To be a rational agent is to deliberate about what is best to do. But then in so far as one is a rational agent, one must want one's choices and actions, whatever they are, to be regulated by the exercise of one's deliberative or rational capacities. This is to value the realization of rational agency or to regard rational agency as good in itself. And Kant might argue that a purely rational agent has no basis for finding anything else intrinsically valuable. Moreover, if I choose rational agency solely in so far as I am a rational being—solely in virtue of properties common to all rational agents as such—then I choose to develop rational agency as such, and not the rational agency of this or that being—in particular, not just my rational agency (*G* 427; *KpV* 20–1). If so, then F1 directs me to be concerned about other rational agents, as rational agents, for their own sakes.[34] Kant concludes that in so far as we

[33] This removes worries about whether the universalization of maxims to pursue innocent components of larger divisions of labour (that are themselves innocent) is conceivable, in Kant's sense. Because universalization is not really essential to F1, the apparent non-universalizability of these innocent activities does not imply that these activities are impermissible. Moreover, the way in which universalizability plays no real role in the favoured interpretation of F1 might usefully be compared with the way in which justice as fairness represents a problem in individual decision theory under special circumstances, rather than a contract among several parties with conflicting interests. For the thickness of the veil of ignorance in the original position aims to abstract from those features of individuals that set them at odds and would otherwise require a contract among them to be represented as a bargaining problem. See John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard Univ. Press, 1971) (cited as *TJ*), 17, 119, 121, 138, 139.

[34] Cf. Christine Korsgaard, 'Kant's Formula of Humanity', *Kant-Studien*, 77 (1986), esp. 190–7.

are rational beings we would will that all rational agents be treated as ends in themselves and never merely as means (*G* 429). This is how he gets from F1 to F2.

The transition from F1 and F2 to F3 is more straightforward. If F1 represents a test for the permissibility of our maxims that we interpret in terms of the choice of a purely rational agent and, so interpreted, F1 equals or implies F2, then we get the following picture. We are free to act on maxims that we, as rational beings, can will to be universal and that treat other people as ends in themselves and never merely as means. This sounds very much like F3; every rational being should be regarded as an autonomous legislator in a kingdom of ends (*G* 432–3).

8. *The Content of the Categorical Imperative*

However, these claims about the relation among the three main formulations of the Categorical Imperative do not yet answer the 'empty formalism' charge. Moreover, if Kantian claims about the Categorical Imperative are to have a bearing on our puzzle about the rational authority of other-regarding morality, then we need some assurance that the Categorical Imperative will enjoin some familiar other-regarding duties. I'll briefly sketch two ways of articulating the content of the Categorical Imperative, though I regard them as complementary, rather than competing, strategies.

First, we might begin with F2. We get moral content by figuring out what it would be to treat someone as an end, and not merely as a means. To use something as a means is to treat it as an instrument or resource for one's own aims; to treat it merely as a means is to treat it only this way, in particular, not as something with interests or value of its own. Of course, this is acceptable where, as with tools, the means have no value of their own but only instrumental value. But with people and rational agents in general this is not true. To respect people as ends is, for Kant, to value them and recognize their worth as rational agents (*KpV* 87). If, as we've claimed, what it is to be a rational agent is to be able to distinguish between the intensity and authority of one's desires and to have capacities for deliberative self-governance, then F2 requires that we value rational agents as deliberative beings and not treat them as mere means to the satisfaction of our own aims.

F2 prohibits treating rational agents as mere means. This requires treating them as ends, whose deliberation and agency are valuable. This requires not simply that we refrain from doing things that would harm the agency of others but also that we do things to promote their rational agency. And this will involve a concern to promote or assist, where possible, the opportunities of others for deliberation and agency, the effectiveness of their

deliberations, and the execution of their choices and commitments (*M* 450, 452). Kant makes this clear in his discussion of the application of F2 to the example involving mutual aid.

> Now humanity might indeed subsist if nobody contributed anything to the happiness of others, provided he did not intentionally impair their happiness. But this, after all, would harmonize only negatively and not positively with humanity as an end in itself, if everyone does not strive, as much as he can, to further the ends of others. For the ends of any subject who is an end in himself must as far as possible be my ends also, if that conception of an end in itself is to have its full effect in me. (*G* 430)

Indeed, given the concern each must have for rational agents, it is reasonably clear how a maxim of complete indifference to the needs of others would represent a contradiction in the will of a purely rational being. For Kant believes that it is analytic that in so far as I will the end, I must, in so far as I am rational, will means and necessary conditions to that end (*G* 417). If this is analytic in Kant's sense (*KrV* A6–7/B10–11), then willing the means is part of willing the end. But various human needs are means or necessary conditions to the pursuit of rational agency. In so far as I am rational, I do will the pursuit of rational agency; but then I cannot in consistency fail to will that rational agents be supplied those things they need as means or necessary conditions to the exercise of their rational agency.

There appear to be two main limitations on one's duties to promote the rational agency of others. First, I am constrained in the ways I can promote the agency of others, in much the way that I am constrained in the ways that I can help you win a competitive race. I can help you train for the race, but I cannot run and win the race for you. It's like this with rational agency. The exercise of one's rational agency involves making one's fate dependent, so far as possible, on one's actions and making one's actions dependent, so far as possible, on one's deliberations. I can provide intellectual and material resources for your deliberations and the execution of your plans, but I cannot deliberate for you (*M* 386). I can promote your agency only in ways that engage your deliberative capacities.[35] Second, if we are to respect the constraint that F2 imposes, the agent's obligations to help

[35] Kant thinks that whereas each has a duty to promote her own perfection, each has a duty to promote the happiness, rather than the perfection, of others (*M* 385–8). In so far as this self/other asymmetry rests on this claim that I am constrained in the ways that I can promote the rational agency of another, it need not reflect a fundamental asymmetry. For I can promote the rational agency of others, by providing them with various intellectual and material resources for their practical deliberations, just not in ways that do not engage their own deliberative capacities. In other words, the issue is not so much about whether as about *how* to promote the rational agency of another. If so, such self/other asymmetry as there is is compatible with and, in fact, seems to depend upon a prior and deeper symmetrical concern for the rational agency of self and others. Indeed, some such deep symmetry seems to be needed if we are to square this asymmetry in *The Metaphysics of Morals* with the apparently symmetrical concern for self and others contained in F2.

others realize their agency cannot be so encompassing that she becomes a mere means to the realization of their ends; she must also treat herself as an end and recognize duties to herself (*G* 429–30).

The interesting and difficult issues concern what, if any, distributional constraints F2 imposes on concern for rational agency. What does F2 require when rational agents make competing claims on me? It is sometimes thought that F2 imposes a side-constraint on action, roughly, that I can and should act to promote rational agency only on the condition that I never harm or impede anyone's rational agency.[36] Suppose that only by causing harm to B's rational agency can A prevent individually comparable harms to the agency of C, D, and E. On this view, F2 forbids harming B's agency, even though so acting might better promote rational agency or at least minimize harms to rational agency. But it is not obvious that F2 requires such a side-constraint. F2 requires that one treat rational agents as ends and not merely as means. If A harms B's agency only in order to protect the agency of C, D, and E, perhaps A treats B as a means, but he does not treat her as a mere means. To do that would require viewing her as a mere instrument or tool, not as someone whose own agency is valuable. But A does not view her that way; A has taken her agency into account. A proceeds, but with great reluctance that derives from a concern with her agency; if A could have protected the agency of C, D, and E without harming her agency, he certainly would have. If A acts impermissibly in acting so as to minimize harm to rational agency, it is not because in so acting he must be treating those whose agency he harms as mere means.

It is natural to think that to treat every agent as an end is precisely to be impartial in a way that takes the agency of each affected party into account equally. I think that this is right and frames further reflection in a useful way. But it does not yet settle much, because there are alternative conceptions of impartiality and equality. On an *aggregative* interpretation of impartiality, we consider the interests of each affected party, *qua* rational agent, and balance benefits to some against harm to others, where necessary, so as to achieve that outcome that is on balance *best* from the perspective of rational agency. On this view, the claims of individual rational agents might be outvoted by a majority. By contrast, we might interpret impartiality to require *unanimity*. On this view, we require that benefits and harms be distributed in a way that is acceptable in a suitable sense to *each* affected agent. There is some reason to think that Kant favours the second interpretation of impartiality (*KpV* 87).[37] In discussing the application of

[36] Cf. Thomas E. Hill, Jr., 'Humanity as an End in Itself', repr. in his *Dignity and Practical Reason in Kant's Moral Theory* (Ithaca, NY: Cornell Univ. Press, 1992), 48–9, 52, 56.

[37] In so far as Kant endorses this interpretation of impartiality, partiality seems unlikely to enter into his moral theory at the most fundamental level (cf. n. 4 above).

F2 to the example of false promising, he writes 'For the man whom I want to use for my own purposes by such a promise cannot possibly concur with my way of acting toward him and hence cannot himself hold the end of this action' (*G* 429). But F2 cannot require the agreement of impurely rational agents. That interpretation of unanimity would impose an intolerable distributional constraint; for each could exercise a veto based on his contingent interests and inclinations. Moreover, this interpretation would make moral requirements depend on the contingent interests and inclinations of agents in a way Kant clearly eschews. Rather, Kant must mean that I am constrained to treat others in ways they could accept were their agreement to reflect only their rational nature. It is not entirely clear what distributional constraint this interpretation of unanimity imposes; in particular, it is not clear that it rules out interpersonal aggregation.[38] Moreover, this brings our interpretation of F2 back to our interpretation of F1.

We might try to determine the content of the Categorical Imperative by focusing on F1. On our interpretation, F1 asks what a rational being as such, independently of her contingent interests and inclinations, would will. We might model this as the problem of what terms of conduct one would choose—in so far as one was rational and valued rational agency—to govern a world of rational beings who have different, sometimes conflicting, contingent interests and desires in which resources are scarce. We might call these conditions the *circumstances of humanity*. These are not the circumstances of rational agency as such, and so, on Kant's view, they are not the circumstances of morality. But they are pervasive features of the human condition that help shape and characterize the kind of moral problems that we face. And we might get an idea of what moral requirements the Categorical Imperative generates for us by trying to model the choice that a purely rational being would make about the terms of social interaction for such circumstances. The natural way to do this is to represent the choice of the terms of conduct for the circumstances of humanity as one that must be made by someone subject to important motivational and informational constraints (*G* 427; *KpV* 21).

Our task might profitably be compared with Rawls's method for modelling the choice of principles of social justice in *A Theory of Justice*.[39] Our

[38] I explore some of these issues, though not especially with Kant in mind, in 'The Separateness of Persons, Distributive Norms, and Moral Theory', in R. Frey and C. Morris (eds.), *Value, Welfare, and Morality* (New York: Cambridge Univ. Press, 1993), 254–89.

[39] Rawls discusses the Kantian interpretation of justice as fairness in *TJ*, sect. 40 and in 'Kantian Constructivism in Moral Theory', *Journal of Philosophy*, 77 (1980), 515–72. He discusses various aspects of Kant's moral philosophy in 'Themes in Kant's Moral Philosophy', in E. Förster (ed.), *Kant's Transcendental Deductions* (Stanford: Univ. Press, 1989), 81–113. Whereas Rawls's explicit motivation for the conditions in the original position is an appeal to considerations of fairness in an agreement to terms of institutional design, the Kantian motivation for the special conditions from which the terms of interaction in the circumstances of humanity are chosen is an appeal to the idea of a rational agent

chooser knows that she will live with others in the contingent circumstances of humanity—that she will have particular characteristics, information, and preferences—but her choice of rules governing conduct in the circumstances of humanity is to be based on her concern for rational agency as such. So we place her behind a *veil of ignorance* that deprives her of knowledge about her various personal and social characteristics, such as her sex, talents, preferences, conception of the good, social position in society, society, and generation. In depriving her of this information, we make her choice independent not only of contingent facts about her interests and desires but also of knowledge as to which rational being she is. This is important if her choice is to reflect the will of a rational being as such and not a parochial concern for rational agency manifested here or there. Her positive motivation, of course, will be to choose principles that will most realize rational agency in the circumstances of humanity. Here she will be concerned with conditions that favour the development and exercise of deliberative capacities, where these include the capacities for forming, revising, assessing, choosing, and implementing structured plans and projects. It's plausible to suppose a rational being as such would favour certain principles of institutional design. Given her ignorance as to which projects and plans, talents, and resources she will have when the veil is lifted, this kind of motivation will obviously lead her to give priority to those goods and resources that serve as necessary conditions to exercising these deliberative capacities and as maximally flexible resources in pursuing her conception of the good once this is known. Following Rawls, we might call such goods and resources *primary goods* (*TJ*, sect. 15). They will include such things as the conditions of physical and mental well-being, education, personal and civic liberties, and economic resources. Her interest in rational agency suggests that above a certain minimum level of material resources, she will assign some kind of priority to personal and civic liberties necessary to exercise her capacities for practical deliberation.[40] And because of her ignorance as to which impurely rational agent she will be, when the veil is lifted, she will presumably assign some kind of presumption to principles that ensure the equal distribution of these conditions for pursuing rational agency.

Whatever principles of just institutional design emerge from this sort of *ex ante* choice will frame and constrain principles of interpersonal morality.

as such. However, it is reasonable to think that these two different motivations converge on a common description of the initial circumstances of choice (*TJ* 251–5). Also, whereas Rawls's focus is on defending principles of justice for the basic structure of society (*TJ* 7, 17, 54), the Kantian has the more comprehensive aim of ascertaining principles of right conduct as well as institutional design. Whereas Rawls calls his project *justice as fairness*, we might call the Kantian project *rightness as rational agency*.

[40] However, I doubt the priority of personal and civic liberties over other primary goods should be lexical, as Rawls claims.

One possibility is that principles of interpersonal morality might be generated by a sequence of choices, the successive stages of which gradually lift the veil of ignorance (cf. *TJ*, sect. 31). At the first stage a choice is made, as we have described, behind a very thick veil of ignorance; at the next stage the veil is lifted so as to reveal what sort of society, with what sort of natural and social resources, the chooser will occupy; at the third stage, the veil is lifted so as to reveal everything about the society and its occupants except which person the chooser is. The idea would be that principles chosen at any stage would be constrained by principles accepted at earlier stages. It is reasonable to think that it is an open question whether someone choosing out of a concern for rational agency but in ignorance of whether he will be A, B, C, D, or E will choose to avoid harming rational agency or will choose instead to minimize harms to rational agency.

A natural worry about this strategy for interpreting F1 is that it may seem to reintroduce contingent facts into the determination of moral requirements in just the way we spent so much time weeding out. For it asks what a rational being as such would will for the circumstances of humanity, and these circumstances include contingent conditions of human need, interest, and desire. But this objection to modelling F1 in this way confuses the conditions or circumstances under which a choice is *made* and the conditions or circumstances that a choice is *for*. It is the former, not the latter, that must be free of contingent factors on the Kantian view. F1 requires that the choice be made by rational beings in so far as they are rational; but the choices certainly apply in circumstances in which agents have particular, contingent desires and needs. Indeed, Kant thinks that the choices of purely rational agents appear as imperatives *only* to impurely rational agents (*G* 414, 455; *KpV* 20, 32, 82). But then there should be no objection to modelling F1 as a choice made by purely rational beings for the circumstances of humanity.

These remarks about the interpretation of F1 and F2 merely outline strategies for developing a substantive moral theory. But this may be enough for present purposes. The fact that both strategies ground moral requirements in an impartial concern for rational agents in the circumstances of humanity makes the 'empty formalism' charge less compelling. It also assures us that Kant can recognize categorical imperatives that enjoin other-regarding action about whose rational authority we can inquire.

9. *From Inescapability to Authority*

We should now have some grip on the way Kant thinks that moral requirements express categorical norms. Does this account of morality and its inescapability help explain its authority? In claiming that moral requirements

express categorical imperatives, Kant claims that they apply to us in virtue simply of our being moral agents, not in virtue of our contingent circumstances and attributes. What makes us responsible agents is our ability to distinguish the intensity and authority of our desires, to deliberate about our actions, and to regulate our actions in accordance with these deliberations. These capacities for deliberative self-governance are the features that make us rational agents, and this is why moral requirements apply to us in so far as we are rational agents. But if some requirements apply to me in virtue of those very features that make me a responsible agent, capable of practical deliberation and subject to reasons for action, then these requirements presumably give me reason to act, such that failure to fulfil those requirements is *pro tanto* irrational. Because, according to Kant, moral requirements do apply to me in virtue of my being a rational agent and not in virtue of my contingent interests and aims, they must give me reason for action, independently of my interests and aims; they give me categorical reasons.

Notice that this route from inescapability to authority is not available for all categorical norms. Legal requirements and requirements of etiquette are categorical norms; they do not apply to someone, to whom they apply, in virtue of her aims or interests. We would not withdraw ascriptions of legal duties or duties of etiquette upon learning that performing her duties would further no aim or interest the agent has. In virtue of what features these requirements do apply is not entirely clear. Particular legal duties presumably apply to one in virtue of one's being a member of or falling within the jurisdiction of a certain kind of social system, defined perhaps by a set of first-order rules and second-order rules specifying the ways in which the first-order rules can be recognized, adjudicated, and changed.[41] Particular duties of etiquette presumably apply to one in virtue of one's belonging to a group in which certain social conventions and rituals, designed to grease the wheels of social interaction, are operative. Though requirements of law and etiquette are in one sense inescapable, they lack authority, because, unlike moral requirements, their inescapability is not grounded in facts about rational agents as such. It is not a condition of being a rational agent that one live by any particular standards of law or etiquette, and perhaps a rational agent need not live under the rule of law or etiquette at all. But moral requirements, according to Kant, apply to any rational agent in virtue of those very deliberative capacities that make her a responsible agent, capable of having reasons for action. If so, it is the way in which moral requirements are categorical norms that explains why they have special authority, not enjoyed by etiquette or law.

[41] See H. L. A. Hart, *The Concept of Law* (Oxford: Clarendon Press, 1961).

Hence, to say that moral requirements express categorical norms and that they provide categorical reasons is to say two distinct things. Though distinct, the two claims are not independent. For it is precisely the way in which moral requirements are categorical norms—they apply to anyone in so far as she is a rational agent—that explains why they provide reasons for action, independently of the agent's interests and aims. If so, Kant does not confuse morality's inescapability and authority, as Foot suggests; he argues from its inescapability to its authority.

Because the Categorical Imperative applies to rational agents as such, it enjoins impartial concern for any rational agent as such. If so, Kant can claim that practical reason can be impartial; I have non-derivative reason to be concerned about any rational agent as such. If practical reason can be impartial, then it is clear how Kant can defend the rational authority of impartial morality against an anti-rationalist threat.

10. *Authority without Supremacy?*

If this is right, Kant can defend a rationalist thesis about the authority of morality; necessarily, there is reason to fulfil other-regarding moral requirements, such that failure to do so is prima-facie or *pro tanto* irrational. Important as this (weak) rationalist thesis is, however, it does not deliver the strong rationalist thesis that contra-moral behaviour is always on balance irrational. A prima-facie or *pro tanto* reason to do something may be overridden or defeated by countervailing reasons. But Kant presumably accepts this stronger rationalist thesis, as well. For instance, he claims that a morally good will—a will that conforms to duty for the sake of duty (*G* 390, 397–8; *KpV* 71–2, 81, 151)—is incomparably good (*G* 434–6).

> This estimation, therefore, lets the worth of such a disposition [i.e.the morally good disposition] be recognized as dignity and puts it infinitely beyond all price, without which it cannot in the least be brought into competition or comparison without, as it were, violating its sanctity. (*G* 435)

But Kant's claim that a good will is incomparably better than other things is only a statement of the stronger rationalist thesis, not an argument for it. Is the stronger thesis plausible? The answer depends on whether there are competing reasons for action.

In the *Critique of Practical Reason* Kant identifies the highest good with the combination of virtue and happiness (*KpV* 110). If virtue and happiness were independent parts of the highest good, then there would appear to be room for a conflict between virtue and the agent's own happiness. But Kant does not understand virtue and happiness as independent elements of

the highest good. For Kant, happiness must always be *conditioned* by virtue; happiness or the satisfaction of desire (*KpV* 22, 34) has value only in a life lived in accordance with the moral law (*KpV* 110–11, 119).[42] Kant's claims about the highest good show that he does not recognize a conflict between moral requirements and agent-centred demands, but they do not themselves constitute an argument against the possibility of such conflicts.

One source of possible conflict is hypothetical imperatives. Unlike categorical imperatives, the necessity of hypothetical imperatives is conditional; they enjoin means necessary to furthering our empirical interests and aims (*G* 414; *KpV* 20–1). On one interpretation, where this condition is met—where the agent has the relevant empirical interest or aim—the hypothetical imperative applies. If the independence assumption of the puzzle about the authority of morality is true, then impartial moral requirements need not further the agent's interests or aims. If hypothetical imperatives generate (hypothetical) reasons, then it appears that there must be possible conflicts between hypothetical reasons and categorical reasons. Unless there is some reason to believe that hypothetical reasons are inferior reasons, the supremacy thesis must seem doubtful.

This doubt about supremacy depends on two assumptions about hypothetical imperatives—that they are conditional only on the agent having the relevant empirical interest or desire and that they supply reasons for action when this condition is met. These assumptions fit some things Kant says about hypothetical imperatives. The German allows us to read that on which hypothetical imperatives are conditional—*wollen* and its cognates—as what one wants or desires (*G* 414, 417; *KpV* 20–1). Moreover, in criticizing other moral systems that ground moral demands in human happiness or sentiment as resting on inclination and, hence, as heteronomous, Kant believes that they rest morality on a hypothetical imperative (*G* 432–3, 443–4; *KpV* 20–8, 35–6). In so far as Kant argues this way, he seems to assume that hypothetical imperatives are conditional on the agent's interests or desires.[43] He may also seem to assume that hypothetical imperatives provide reasons for action when the agent has the associated interest or desire. For in describing the justification for the Hypothetical Imperative, Kant claims that whoever wants or wills (*will*) the end must also, in so far as he is rational, want or will (*will*) the means to that end (*G* 417). If we read *will* in this passage as want or desire, then Kant seems to be saying something like this:

[42] Cf. the useful discussion in Stephen Engstrom, 'Happiness and the Highest Good in Aristotle and Kant', in S. Engstrom and J. Whiting (eds.), *Aristotle, Kant, and the Stoics: Rethinking Happiness and Duty* (New York: Cambridge Univ. Press, 1996), 102–38.

[43] Moreover, this is a common way of interpreting these and other remarks Kant makes about hypothetical imperatives. See e.g. Lewis White Beck, *A Commentary on Kant's Critique of Practical Reason* (Chicago: Univ. Press, 1960), 85 and Allison, *Kant's Theory of Freedom*, 89.

(*a*) If one wants to ϕ, then one has reason to produce means and necessary conditions to ϕ-ing.

With these two assumptions in place, supremacy is jeopardized, because wants or desires that would ground hypothetical imperatives and reasons can and do conflict with categorical reasons.

But we need not accept the interpretation of hypothetical imperatives on which this doubt about supremacy rests. It's not just that Kant thinks there is more to practical reason than prudential or instrumental reason; he denies, I think, that interests or desires automatically supply reasons to act, as this interpretation of the Hypothetical Imperative implies. It is not at all obvious that (*a*) is true. Why should one have reason to promote the satisfaction of one's desires regardless of the content of those desires? In so far as Kant regards hypothetical imperatives as conditional only on the agent's wants or desires, it's not clear that he supposes that having the relevant wants or desires automatically provides reason to act. In criticizing other moral theories that ground morality in happiness or sentiment as resting morality on hypothetical imperatives, Kant clearly thinks that they are incapable of representing their demands as duties. Nor is it clear that he thinks these theories even show that we have reason to act, so as to promote these interests or inclinations. Moreover, Kant's claims about the highest good should make us doubt that he accepts both assumptions about hypothetical imperatives. For, as we have seen, Kant claims there that happiness, which he understands to consist in the satisfaction of (empirical) desire (K*pV* 22, 34), has value only when it is conditioned by virtue, that is, when it occurs in a life lived in accord with the moral law (*KpV* 110–11, 119). But then he must think hypothetical imperatives are conditional on more than simple possession of an interest or desire; he must deny that hypothetical imperatives automatically provide reasons when the condition of their application is met; or both.

Indeed, many of Kant's claims about hypothetical imperatives can be interpreted as insisting that hypothetical imperatives are conditional on something more than the agent's (empirical) interest or desires. Though the German does allow us to read that on which hypothetical imperatives are conditional—*wollen* and its cognates—as what one wants or desires, it also allows us to represent hypothetical imperatives as conditional on what one *wills*. On this interpretation, hypothetical imperatives are conditional on what one wills (*G* 414; *KpV* 20), and the rationale for the Hypothetical Imperative is that whoever wills the end must also, in so far as he is rational, will the means to that end (*G* 417). To will something is, for Kant, not simply to desire it or have an interest in it; the will (*Wille*) is a faculty of choice in so far as the agent is rational (*G* 412, 427, 446). There are

different ways of trying to understand the significance of Kant's claim that hypothetical imperatives are conditional on what the agent wills.

A second interpretation holds that hypothetical imperatives are just conditional claims of practical reason; hypothetical imperatives instruct one to do those things that are means to or necessary conditions of doing those things that one already has reason to do. If so, we might interpret Kant's rationale for the Hypothetical Imperative as saying something more like this:

(*b*) If one has reason to ϕ, then one has reason to produce means and necessary conditions to ϕ-ing.

This claim grounds one reason in another reason, without grounding the first; as such, it does not fully ground hypothetical imperatives. This purely relational or conditional claim is quite plausible and, if true, arguably analytic. Though it clearly provides one way of understanding Kant's claims that hypothetical imperatives supply only conditional or relative reasons (*G* 420), it fails to identify any sense in which hypothetical imperatives depend on interest or desire. So this interpretation does not explain well why Kant thinks that moral systems grounded in happiness or sentiment reduce morality to a hypothetical imperative.[44] Nor does it explain well Kant's more general insistence that the requirements of happiness, at least when conditioned by the moral law, are hypothetical imperatives (*G* 389, 415–16, 433, 442–4; *KpV* 21, 24–6, 35–6, 64–5, 109).

This second interpretation does not really exploit Kant's idea that the will is a faculty of choice in so far as the agent is rational. This suggests we interpret the condition of a hypothetical imperative not simply as another reason but as something one has reason to pursue just in so far as one is rational. This also serves to ground the antecedent and, hence, the consequent reason in (*b*).

(*c*) If one would choose to ϕ just in so far as one was rational, then one has reason to produce means and necessary conditions to ϕ-ing.

But this third interpretation still fails to identify any sense in which hypothetical imperatives depend on interest or desire and so fails to explain his criticism of other moral systems as resting on hypothetical imperatives or his view that requirements of happiness are hypothetical imperatives. Indeed, on this interpretation it is hard to see how to distinguish hypothetical and categorical imperatives. For if hypothetical imperatives are simply

[44] A friend of the purely conditional reading might claim that moral theories grounded in happiness or sentiment are defective precisely because they represent as duties requirements of happiness or sentiment without establishing that these ends are reasonable (cf. *G* 444). This criticism does explain Kant's critical interest in moral theories that are conditioned, but is does not explain his evident critical interest in moral theories that are conditioned on happiness or sentiment (cf. *G* 425; *KpV* 20–8, 34–5, 41, 64–5).

requirements to secure means or necessary conditions to the ends of a purely rational being, and these conditions are part of the will of a purely rational being, then hypothetical imperatives appear to be just a special case of categorical imperatives. Recall that we must distinguish categorical imperatives and the Categorical Imperative and that the former are, at least in part, what the latter requires in particular circumstances and conditions (Section 4). But then hypothetical imperatives, on this purely conditional reading, must apparently be categorical imperatives.

A more attractive interpretation of Kant's considered view about hypothetical imperatives tries to preserve the insights and avoid the problems of the other interpretations. Unlike the first, it insists that hypothetical imperatives are conditional on the agent's will, and not simply her interests or desires; unlike the second and third, it insists that the agent's interests or desires are among the conditions of hypothetical imperatives. Recall Kant's claims about the role of happiness in the highest good: he thinks that happiness, which he understands to consist in the satisfaction of (empirical) desire (*KpV* 22, 34), has value only when it is conditioned by virtue, that is, when it occurs in a life lived in accord with the moral law (*KpV* 110–11, 119). This suggests another interpretation of many of Kant's claims about hypothetical imperatives. On this view, to say that hypothetical imperatives are conditional on what one wills is to say that they depend upon interests or desires that are conditioned by what one would choose just in so far as one is rational. In other words, hypothetical imperatives, on this view, are conditional on interests or desires that one has that are not ruled out or screened off by the moral law. Similarly, when Kant explains the Hypothetical Imperative by claiming that whoever wills the end, also, in so far as he is rational, wills the means to or necessary conditions of his ends (*G* 417), what he is claiming is analytic is not (*a*), (*b*), or (*c*) but something more like this:

(*d*) If one wants to ϕ and ϕ-ing is consistent with the demands of (pure) practical reason, then one has reason to produce means and necessary conditions to ϕ-ing.

I doubt that (*d*) is analytic, but it is more plausible than (*a*), and it secures dependence on interest or desire that (*b*) and (*c*) do not. To will the end, at least in these contexts, just is, I believe, to choose something based on one's desires in a way consistent with and regulated by those ends that a rational agent, as such, would endorse.[45] I don't know if we can understand

[45] Kant's later writings distinguish between *Wille* and *Wilkür*, though the two are not distinguished in the *Groundwork* and the *Critique of Practical Reason*. *Wille* in a narrow sense refers to a capacity of practical reason, whereas *Wilkür* refers to a capacity for choice on the basis of desire or inclination (*Triebfeder*); *Wille* is also used in a broad sense to refer to a will (*Wilkür*) determined by *Wille* in the narrow sense. (Cf. *M* 213–14; Beck, *A Commentary on Kant's Critique of Practical Reason*, 176–81; and

all of Kant's remarks about hypothetical imperatives as premissed on this interpretation of the way in which hypothetical imperatives are conditional on the agent's will. But it does provide an attractive view of hypothetical imperatives and reasons that is consistent with many things he says and that affords him a plausible reply to this doubt about morality's supremacy. For one cannot will, in this sense, ends excluded by the Categorical Imperative. If so, it's hard to see how there might be hypothetical reasons that conflict with the impartial demands of the Categorical Imperative.[46]

11. *A Dualism of Practical Reason*

However, another threat to the supremacy of impartial moral requirements is harder to dismiss. Moral requirements generate categorical reasons, because they apply to rational agents as such—that is, to an agent in so far as he has those capacities that are essential to responsibility and the possession of reasons for action. These categorical reasons are impartial, because they apply to the agent just in so far as he is one rational agent among others, and not because he is a particular rational being (*G* 427; *KpV* 20–1). But I am essentially not just a rational agent but also a *particular* rational agent, numerically distinct from other agents. The claim that I am a particular rational agent is *not* the claim that I am a finite rational being with an idiosyncratic set of empirical needs and desires; this may secure some kind of particularity, but it is not the particularity I am concerned with here. I am interested, instead, in the particularity of purely rational beings. Purely rational agents are still particular beings, as is clear from the fact that even gods (infinitely rational beings) who have no empirical natures would still be numerically distinct from one another.

One view about what distinguishes rational agents as such that seems promising and has some Kantian credentials is that the identity of rational

Allison, *Kant's Theory of Freedom*, 129–36.) Whereas some of Kant's central claims in the *Groundwork* about the will are concerned with *Wille* in the narrow sense (*G* 412, 427, 446), I suppose that it is something like this broad sense of *Wille* on which, I think, Kant here makes hypothetical imperatives conditional. Indeed, it might not be too far wrong to think that the (*a*)-reading makes hypothetical imperatives conditional on *Wilkür*, the (*b*)-reading and (*c*)-reading make them conditional on *Wille* in the narrow sense, and the (*d*)-reading makes them conditional on *Wille* in the broad sense. We might also note a parallel between the (*a*)-reading and the (*d*)-reading and Kant's distinction between self-conceit and self-love (*KpV* 73–7). Like the (*a*)-reading, self-conceit treats any desire as giving reason for its satisfaction; by contrast, rational self-love, like the (*d*)-reading, conditions the rationality of pursuing one's desires on their conformity with the moral law.

[46] For some different but related claims about the Hypothetical Imperative and its relation to the Categorical Imperative, see Stephen Darwall, *Impartial Reason* (Ithaca, NY: Cornell Univ. Press, 1983), 16, 79; Thomas E. Hill, Jr., 'The Hypothetical Imperative', repr. in *Dignity and Practical Reason in Kant's Moral Theory*, 24, 32; and Christine Korsgaard, 'The Normativity of Instrumental Reason' (Essay 8 in this volume).

agents over time consists in a kind of continuous deliberative control of intentional states and actions. Deliberative control exists when intentional states—such as beliefs, desires, and intentions—are formed, maintained, and modified as the result of deliberation and when actions are regulated by prior deliberations. What makes lines of deliberative control distinct—even when the intentional states, deliberative processes, and actions in each line are qualitatively similar—is lack of functional integration. Intentional states, deliberations, and actions can be ascribed to the same agent just in case they are part of the same psychic economy; intentional states must be able to interact with each other so as to modify each other and produce action. For example, A's pain will directly tend to produce B's avoidance behaviour just in case A and B are the same agent. The same is true with A's intention to vote and B's plan about how to get to the polling booth, A's belief that it is raining and B's desire to get an umbrella from the closet, etc. On this view, intentional states and actions are correctly ascribed to a single agent just in case they are parts and products of a functionally integrated deliberative system. If so, what makes someone a rational agent is that he is capable of deliberating about his desires, in light of his other intentional states, and of taking actions that reflect those deliberations. For Kant, self-consciousness requires an ability to distinguish oneself from particular impulses and desires (*KrV* B132–5); so he must think that agency requires a capacity for self-consciousness precisely because agency requires a conception of oneself and what one should do that is distinct from the various particular impulses one has and what they incline one to do.[47] It is the functional integration of this deliberative control over time that makes someone a numerically distinct and temporally extended agent. For Kant, this requires a unified consciousness, one that is sufficiently unified to support self-consciousness (*KrV* A97, A107–8, A110, A117, B132–4, A212/B258–9, A352).

Given that there are a plurality of purely rational agents, there must be requirements concerned with my own agency that apply to me just in so far as I am a particular rational agent, independently of my contingent interests and desires, just as Kant believes there are requirements of impartial concern that apply to me simply in so far as I am a rational agent. We might call the former requirements of *categorical prudence*.[48] This is not simply the claim that I have reason to be concerned about my own rational agency, as well as that of others. For this kind of self-concern would

[47] Cf. Green, *Prolegomena to Ethics*, bk. ii, esp. sects. 85–8, 100, 120–9 and Irwin, 'Morality and Personality: Kant and Green'.

[48] So the demands of a certain kind of prudence constitute categorical or external reasons. Cf. Terence Irwin, 'Kant's Criticisms of Eudaimonism', in Engstrom and Whiting (eds.), *Aristotle, Kant, and the Stoics*, 63–101.

already be included in an impartial concern for all rational agents, as in F2. Rather, the idea is that I ought to have concern for my own rational agency that is grounded in my being a particular rational agent and not simply one rational agent among others. Categorical prudence is no more included in categorical impartiality than the claims of ethical egoism are included within the claims of utilitarianism. Whereas being a responsible agent capable of having reasons for action depends upon one having the same deliberative capacities that would make anyone else a responsible agent, responsibility is ascribed to particular rational agents on the basis of the way they exercise their deliberative capacities. If so, it seems I ought to possess reasons for action in virtue of facts about my own agency as well as in virtue of rational agency as such. But then I will have reason to promote my own rational agency, as well as to promote agency impartially.

It's worth distinguishing the imperatives of categorical prudence from the assertoric imperatives of conventional prudence. Kant understands the imperatives of prudence to be imperatives to pursue one's own happiness; he understands happiness to consist in the satisfaction of one's (empirical) desires (*G* 399; *KpV* 22, 34), and he conceives desires to be aimed at pleasure (*KpV* 21). It follows that the principle of prudence, according to Kant, is 'empirical and can furnish no practical laws' (*KpV* 21). Thus, even if everyone desires her own happiness, prudential imperatives are (at most) hypothetical and prudential motivation is heteronomous (*G* 389, 415–16, 433, 442–4; *KpV* 21, 24–6, 35–6, 64–5, 109).[49] But the imperatives of categorical prudence are categorical and not merely assertoric. They are imperatives to agents to promote their own rational agency; they apply to each agent in so far as she is a particular rational agent, not in virtue of contingent aims or feelings that are extraneous to her agency. If so, the imperatives of categorical prudence express categorical norms and generate categorical reasons.

But then an argument parallel to Kant's own argument for the claim that impartial moral requirements generate categorical reasons demonstrates that self-regarding requirements of categorical prudence also generate categorical reasons. In so far as the argument is parallel, I do not see how Kant can argue that the reasons of categorical prudence are inferior to those of morality. More generally, I see no reason to suppose that the imperatives of categorical impartiality will systematically override those of categorical

[49] This is how we should understand Kant's claim that such imperatives have only 'natural necessity' (*G* 415; *KpV* 25). Imperatives of conventional prudence do not apply to agents in virtue of their rational agency, because happiness or even the capacity for happiness, as Kant understands it, is not essential to my being an agent or even a particular agent. Unfortunately, he goes on to say that such assertoric imperatives apply to me in virtue of purposes that 'belong to my essence' (*G* 416). This cannot be his considered view; if it were, he would be committed to the claim that imperatives of conventional prudence stand on a footing with moral requirements and so are categorical imperatives.

prudence. If so, Kantian moral psychology must recognize a *dualism of practical reason* that threatens the supremacy of impartial moral requirements.[50]

It's not clear whether this dualism represents a conflict within morality or a conflict between morality and some practical perspective external to morality. The answer depends upon whether the demands of categorical prudence are themselves moral demands or are extra-moral demands. Because Kant does not recognize the demands of categorical prudence, it's hard to know what he would think. On the one hand, he thinks that moral requirements are expressions of the perspective of a rational agent as such (*G* 408, 412, 423, 425–7; *KpV* 20–1, 29–30; *M* 216). We can understand these requirements as applying in virtue of properties common to all rational agents. We can model the way in which these requirements are generated in terms of a choice behind a veil of ignorance that abstracts from various identifying features of the chooser (*G* 427; *KpV* 21). This reasoning naturally leads to F2's impartial concern for rational agents. In so far as Kant argues this way, he seems committed to regarding agent-centred requirements of categorical prudence as extra-moral demands. On this view, recognition of categorical prudence threatens the supreme authority of morality.

On the other hand, Kant appears sometimes to equate categorical imperatives and requirements of morality (*G* 416, 420). Because the demands of categorical prudence are categorical norms that generate categorical reasons, this may give us reason to regard them as moral demands. Moreover, Kant thinks of moral requirements as depending upon features common to all rational agents as such. Whereas each has the capacities for deliberative self-government that all the others have and that would make anyone a moral agent, each also is a particular rational agent. So particularity is also a feature common to all rational agents as such. But if what is true of rational agents as such grounds moral requirements, then the requirements of categorical prudence are moral requirements. On this view, recognition of categorical prudence threatens morality's impartiality.

But this obscures what the two interpretations of the dualism of practical reason have in common; both challenge the supremacy of impartial moral requirements. One accepts morality's impartiality and challenges its supremacy; the other accepts its supremacy and challenges its impartiality.

[50] This dualism of practical reason might be profitably compared with Sidgwick's dualism between egoism and utilitarianism; see *The Methods of Ethics*, esp. 496–509. However, the comparison is imperfect. Categorical prudence is not the same as hedonistic egoism; categorical prudence takes the rational agency, rather than the pleasure, of the agent to be the thing to be promoted. Similarly, categorical impartiality is not the same as hedonistic utilitarianism. It too is concerned with rational agency, rather than pleasure. Moreover, it is not clear that categorical impartiality should be understood in consequentialist terms; however, unlike many, I do not think it clear that it should not be understood in such terms (cf. Sect. 8 above).

Either way, the most serious challenge to Kantian rationalism is not the move from the inescapability of impartial moral demands to their authority but the one from their authority to their supremacy.

This leaves two responses to our puzzle about the rational authority of morality. The first response would be to abandon the supremacy of impartial morality. This would provide a weak form of rationalism without Kant's stronger rationalist aspirations. On this view, though we reject the agent-centred assumptions about practical reason in premiss (3) and maintain that necessarily there is reason to act on impartial moral requirements such that failure to do so is *pro tanto* irrational, we must also reject the strong rationalist premiss (2) that failure to act on impartial moral requirements is necessarily on balance irrational. Impartial moral requirements would necessarily enjoy authority, but would not necessarily enjoy supremacy. The other response would be to seek a *practical* resolution of the dualism by showing that the interests of distinct rational agents, when properly understood, are interdependent in such a way that acting on an impartial concern for rational agents is a counterfactually reliable way of promoting the agent's own rational agency (and vice versa). On this view, despite a dualism between agent-centred and impartial practical reason, we can try to maintain the strong rationalist thesis (2) by rejecting the independence assumption in premiss (4). To develop this response, however, we would need to look outside Kantian ethics to the eudaemonist tradition in Greek ethics, which Kant rejected, or to the ethics of self-realization found later in British idealism.[51]

[51] I have discussed this alternative in 'Self-Love and Altruism'. Of course, even if we can reduce the conflict between impartial and agent-centred concern with rational agency, we may not be able to eliminate it completely. If so, rejecting premisses (3) and (4) in the puzzle about the authority of morality may not be sufficient to deliver the strong rationalist commitment to the supreme authority of impartial morality, expressed in (2). See 'Self-Love and Altruism', sect. 12.

I would like to thank Henry Allison, Richard Arneson, Anne Margaret Baxley, Richard Boyd, Joshua Cohen, Garrett Cullity, Stephen Engstrom, Berys Gaut, Michael Hardimon, Paul Hoffinan, Brad Hooker, Terence Irwin, Patricia Kitcher, Christine Korsgaard, Wayne Martin, Paul Pietroski, Geoffrey Sayre-McCord, Alan Sidelle, John Skorupski, Michael Smith, Jennifer Whiting, a UCSD graduate seminar, participants in the Ethics and Practical Reason Conference at the University of St Andrews, and an audience at the University of California, Davis for helpful discussion.