ON THE PRACTICAL IRRATIONAL OF IMMORALITY

MICHAEL NELSON

University of California at Riverside
Department of Philosophy
Riverside
California
U.S.A.
mnelson@ucr.edu

Article info
CDD: 128.33
Received: 10.08.2018; Accepted: 21.08.2018
DOI: http://dx.doi.org/10.1590/0100-6045.2018.V41N4.MN

Keywords:
Morality
Autonomy
Practical Reason

Abstract: I argue that the Formula of Humanity, the principle that we should always treat the humanity of a person as an end in itself and never as a mere means, is a principle of pure practical reason. Insofar as that principle is also the fundamental grounds of morality, it follows, then, that all autonomous rational agents are committed to morality.

ON THE PRACTICAL IRRATIONAL OF IMMORALITY

According to Kant's Formula of Humanity, the second formulation of his Categorical Imperative, we should always treat the humanity of persons as an end in itself and never a mere means. This principle is put forward as a version of the "supreme principle of morality," providing the grounds of all moral requirements. A choice is morally

impermissible, then, because it involves treating humanity as a mere means. Treating humanity as a mere means is the mark of the immoral, casting explanatory light on the immorality of any given immoral action. And, on Kant's view, moral requirements are always overriding, in the sense that, if one has most reason to do *x*, then one should, all things considered and *simpliciter*, do *x*. So, insofar as the FH is the fundamental principle of morality, then, it follows that it must be itself a principle of pure practical reason (or at least implied by something that is a principle of pure practical reason).

The FH is claimed to have a peculiar set of features: Unlike specific requirements like not killing the innocent or deceiving another for one's own amusement, the FH is sufficiently general to be the foundation of all of morality; unlike purely formal principles of reason, such as the principle to always do what one has most reason to do and even the instrumental principle to take the necessary means to one's intended ends, the FH is substantive; and unlike, for example, principles of etiquette, the FH is constitutive of rational agency as such.

In this essay I develop an account of treating humanity as a mere means. I argue that treating the humanity of a person as a mere means involves judging the relative worth of one's contingent end to be more valuable than the worth of that person's humanity. Call this the Evaluativist understanding of the FH. I contrast this understanding with other competing understandings and interpretations of the FH, in particular the Consent Conception of the FH. I am less interested in interpreting Kant's writings on the FH, although I think that the Evaluativist conception fits the whole of the text better than the Consent conception. I am more interesting in arguing the FH, understood as the Evaluativist does, is true and accepting it helps to make progress defending the moral rationalism articulated in the opening paragraphs above. To act immorally, then, is, on

the view to be developed, to treat the humanity of a person as a mere means, which is to act in a way that, were one fully articulate and coherent, involves judging one's adopted end to be more valuable than the humanity of a person. This is practically irrational, and so a course of action that one should not, full stop and *simpliciter*, do, because every autonomous practical reasoner is rationally committed to judging humanity to be a grounding value, which is a judgment inconsistent with the judgment that one's adopted ends are more valuable than the humanity of a person. I argue this by developing a realist, cognitivist version of Kant's regressive argument for the value of humanity. On this view, the value of humanity is not the source of value but instead the rational grounds for one's placing the importance one does on some value, a value which, for all the argument is concerned, may have grounds independent of choice, as the realist claims. While Kant may well have been a constructivist about value, that is not, I think, what his regressive argument for the value of humanity requires or shows.

Moral rationalism is the thesis that all moral demands are, at their ground, principles of practical reason as such, so that in acting immorally, one acts contrary to what one should do. Sometimes the dictates of a domain, say the dictates of being a strong advocate for one's students, require doing something, say, overstating a student's abilities, that one should not, all things considered, do. That is, the dictates of that domain conflict with the dictates of pure practical reason, as those dictates are outweighed by competing considerations. It is intuitive to maintain that mismatches between the dictates of morality and the dictates of practical reason do not exist. The 'should' of practical reason and the 'should' of morality speak with a single tongue. That is the thesis of moral rationalism.

Moral rationalism promises a robust defense of morality's authority. If moral demands just are demands of

practical reason, the ammoralist's challenge *Why be moral*? is equivalent to the challenge *Why should I do what I have most reason to do*?, which presupposes the very authority it questions. Given moral rationalism, a person unconcerned about others is not merely unattractive, unappealing, distasteful, and a jerk; such an agent is also practically irrational, which is a criticism internal to the agent's own point of view, whatever her contingent aims and concerns happen to be. The immoral agent violates requirements that are constitutive of the bare activity of autonomous choosing. Her action is self-defeating in the sense that the commitments she undertakes in choosing to act as she does are guaranteed to be inconsistent and so it is impossible for her to satisfy them all.

Many have thought that the promise of moral rationalism rests on a sleight of hand, as the rationalist employs one set of principles when arguing that the fundamental principles of morality are principles of practical reason and subtly different principles when deriving particular substantive moral demands. The worry is that there is no single set of principles that are both genuine principles of pure practical reason, principles that every rational agent, whatever her contingent concerns, cares, and projects, is committed to, and that deliver substantive moral requirements. I am sympathetic with to this concern and want in this paper to begin to address it.

Here I take only the first steps in a defense of moral rationalism, as I argue that the FH, interpreted as the Evaluativist does, is a principle of pure practical reason. But this is an important step. What remains to defend the moral rationalist's position is that this same principle is a fundamental principle of morality. It strikes an intuitive chord that the mark of the immoral is treating persons as mere means, using them as mere things that are to be put to one's own service in any manner that one wills. What is less obvious is that the FH is a principle of pure practical

reasoning as such, which is what I will be arguing for, supporting the claim that every autonomous agent, whatever her circumstances, projects, concerns, aims, and cares, is rationally committed to the FH just in virtue of her rational autonomy.

I don't claim that it is trivial to show that the FH is a fundamental principle of morality, especially as the content I ascribe to the FH in arguing for its status as a rational principle may seem far removed from the intuitive moral content of treating person as mere means. While others have argued that cases of deception and coercion fit the bill when the FH is understood in terms of consent, a conception I shall discuss below, I do not know of a comprehensive discussion that shows that a wide range of cases of immorality can be explained in a similar fashion, and I shall give reasons to think below that the consent conception of the FH only works for a narrow range of cases. And I know of no discussions showing of the Evaluativist interpretation of the FH. However, it is beyond the scope of this essay to argue that the FH, under the Evaluativist interpretation, is a principle that covers all immorality and so I do not pretend that I am here establishing moral rationalism. I shall instead focus on showing that the Evaluativist conception of the FH is a principle of pure practical reason.

The FH is the principle that one should always treat humanity, whether in one's own person or in the person of another, always at the same time as an end in itself and never as a mere means. So, treating humanity as a mere means is forbidden. But what does that amount to? It is widely agreed that humanity is to be identified with the capacity of choice. What exactly that capacity amounts to, of course, is wildly controversial, and we also need to worry whether the sense of choice is something that nonhuman animals like chimpanzees, cats and dogs, cows, and perhaps even snakes and spiders share or, instead, whether the

operative sense of choice is *human*, autonomous choice. Furthermore, a topic live in Kant interpretation, it is debatable whether humanity, conceived of that which is good in itself, is choice *done well*, i.e., moral choice, or the underlying indifferent capacity itself, whether used well or poorly. We do not need to take a stand on these questions here and we do not, I hope, rely on anything but noncontroversial components of an adequate account of choice.

To treat the humanity of a person as a means involves treating the capacity of choice as a means to achieving one's end. Means are tools to one's ends and they have an inherited value, being meaningful to the chooser in virtue of its relation to her ends. However, one can use something as a means without treating it as a *mere* means. Similarly, one can use a person to achieve one's ends without acting immorally. For example, Susan uses the restaurant staff to get her dinner but, we can suppose, the whole affair is perfectly moral. She does not use them as a *mere* means. What, then, is added to using someone as a *mere* means?

One important account from the literature is the Consent Conception, suggested by Kant's own discussion of the second of his four cases of immorality discussed in the second section of his *Groundwork*, when he writes: "one who has it in mind to make a lying promise to others will see at once that he wants to make use of another human being *merely as means*, who does not at the same time contain in himself the end. For the one I want to use for my purposes by such a promise cannot possibly agree to my way of proceeding with him and thus himself contain the end of this action" (4:429). In this passage, Kant seems to be saying that when A makes a false promise to B in order to achieve her end of, let's say, paying her rent, A uses B as a mere means because B cannot possibly agree with A's course of action. (Kant also suggests that B thus cannot "contain in himself" A's end, which some take to be a

competing account, the so-called *end sharing* conception, of what it is to treat another a mere mean. I shall not here discuss this conception.) Both Onora O'Neill (1985) and Christine Korsgaard (1986) are proponents of the Consent Conception of treating persons as mere means. Very crudely, on this conception, one uses a person as a mere means when one involves that person in one's action in ways to which she does not (or could not possibly) consent.

Suppose that A and B chose to compete in a fair competition. Each wants to win and neither are forced nor coerced in any way to participate. While any fair competition will serve our purposes, combat sports, like a boxing match, provide particularly vivid illustrations of the issue. When A hits B with a solid combination and B falls to the canvas, B surely didn't want to be hit and even more surely didn't want to be knocked down; she was trying to win, after all. B did not "share A's end" of knocking B out and did not consent to being knocked down. Still, insofar as we set aside any thought that sport fighting is intrinsically wrong, A's action is not immoral.

We cannot account for the morality of A's treatment of B by citing the fact that B would do the same to A given the chance. That is also true of two gangsters trying to rob and murder one another, but the triumphant gangster nonetheless treats her adversary as a mere means and so acts immorally. Things are otherwise with our boxing match.

While B did not consent to being hit by the combination A in fact intentionally hit her with, B did consent to A's doing her best to hit B. In fact, insofar as B's interest was in having a true competition, it was essential to B's aim that A do her best to win and so her best to hit B. When one consents to another's doing her best to do *x*, knowing full well that there is at least a live possibility that that person will succeed, one has derivatively consented to that person *x*-ing, should it come

to pass. So, I doubt that cases of competition are genuine counterexamples to the Consent Conception of treating persons as mere means.

I think that we can get a deeper understanding of the Consent Conception, as well more clearly expose the shortcomings of that conception, by considering the discussions of the Formula of Humanity and the notion of treating persons as mere means is in chapters 8-10 of Derek Parfit's *On What Matters*. In chapter 8, Parfit discusses what he calls the Consent Principle, the principle that it is wrong to treat a person in any way to which she could not rationally consent (p. 181). Parfit does not say that we should understand the notion of treating a person as a mere means contained in Kant's Formula of Humanity in terms of treating that person in ways to which she could not rationally consent, as O'Neill and Korsgaard do, and chapter 9 is focused primarily on the notion of treating persons as a mere means, where Parfit goes beyond consent. Parfit thinks of the Consent Principle and the Mere Means Principle as separate, independent principles that together constitute the FH. At the end of chapter 8, Parfit writes: "The Consent Principle cannot, however, be what Kant was trying to find: the supreme principle of morality. Some acts are wrong even though everyone could rationally consent to them. The Consent Principle states one of the ideas that are expressed in Kant's Formula of Humanity. Since we need at least one other principle, we can now turn to another part this formula" (p. 211). However, I am going to first consider the thesis that the Consent Principle captures the full force of the FH and thus, as Parfit says, is the "supreme principle of morality," capturing what it is for each and every wrong act to be wrong, arguing that this thesis is false. My discussion will focus on Parfit, even though he explicitly rejects the view I am criticizing, because his is a clear and precise account of the Consent Principle. I then return to Parfit's thesis that

the Consent Principle is an essential but not exhaustive component of the ideas that are expressed in Kant's Formula of Humanity, arguing against even that weaker thesis. On my view, the Consent Principle is superfluous. While the impossibility of rational consent is sufficient for immorality, and so the Consent Principle, considered as providing a sufficient condition for immorality, is extensionally adequate, it does not reveal the essence of immorality, in the sense that an act's immorality is not, at the deepest level, explained in terms of the fact that there is someone who cannot rationally consent to the behavior. The fact that there is someone who cannot rationally consent to some behavior is always itself grounded in some more basic set of facts that ultimately ground the immorality of the behavior. The lack of rational consent, then, is a symptom and not the underlying ground of immorality. In explaining immorality, then, we should seek those deeper grounds and throw rational consent to the side.

The bulk of chapter 8 of *On What Matters* involves testing the Consent Principle against versions of lifeboat cases in which needed aid and so harms must be distributed across people. Parfit begins with variants of what he calls "Earthquake," in which two people, White and Grey, are trapped in collapsing wreckage, with White's life and Grey's leg under threat. A rescuer can save one but not both. Suppose that all of the parties involved are strangers to one another and, whatever exactly this amounts to, do not "differ in any other morally relevant way" (p. 186). Parfit claims, plausibly, that the rescuer should save White's life. Parfit argues that the Consent Principle delivers this result by, first, claiming that whether or not a person can rational consent to some treatment is a function of the reasons she has for and against accepting that treatment and, second, Parfit's preferred *wide value-based objective* theory, which, crudely, is the thesis that the reasons an agent has are

determined by the values, the goods and bads, of the outcomes of those treatments. Grey has sufficient reasons to choose that the rescuer save her leg, as having working legs is good for her, and sufficient reasons to choose that the rescuer save White's life instead, on the grounds that the harm White would have to endure given the alternative is much worse than the harm that she would have to endure without the use of her legs. So, she has compelling reasons either way, but neither set is decisive. White, by contrast, could not rationally choose that Grey's leg be saved at the cost of her life, Parfit argues, as the value of White's life is much greater than the value of Grey's use of her legs. So, White could not rationally consent to Grey's leg being saved in these circumstances, while all parties involved could rationally consent to White's life being saved in these circumstances. So, the Consent Principle dictates that, given the stipulated circumstances, the rescuer should save White's life, allowing Grey's legs to be crushed.

Parfit describes a first variant of Earthquake, what he calls "Means," in which White is again caught in collapsing wreckage with his life at risk, but this time the rescuer can save White's life only by using Grey's body as a shield, resulting in the destruction of Grey's legs; if the rescuer were to do nothing, Grey would be unharmed and White would die. The crucial difference between Earthquake and Means, of course, is the manner in which White's life is saved. As its name suggests, in Means, Grey body is used as a means to save White's life. Suppose now that Grey does not consent to that use, which, as was argued above, is perfectly rational, claims Parfit, in light of the value of the use of her legs. In that case, Parfit argues, White could not rationally consent to being saved by Grey's body being used as a shield, as, in that case, White has a decisive reason to not be saved by using Grey's body as a shield: Namely, that it would be morally wrong to so use Grey in that way without her actual consent. In Earthquake, by contrast,

White isn't used and so her actual consent is not necessary. The fact that she could rationally consent to the treatment makes it permissible, on Parfit's view, to allow her leg to be destroyed in Earthquake, regardless of whether or not she actually consents.

Parfit claims that, if we set aside the fact that it is morally wrong to use Grey in this manner without her consent, then Earthquake and Means are on a par, as "In both [cases], either White will die or Grey will lose her leg. These cases would differ only in how the saving of White's life would be causally related to the loss of Grey's leg. Grey would have no strong reason to prefer to lose her leg in one of these ways" (p. 202). So, in Parfit's view, then, if we help ourselves to the claim that it is morally wrong to use Grey to save White's life, then White has a sufficient reason to consent to his not being saved and so the Consent Principle delivers the result that it is morally permissible to not save White. However, if we do not help ourselves to the moral wrongness of using Grey to save White's life, then the Consent Principle delivers no such result.

I understand Parfit to claim that, regardless of whether or not Grey actually consents to the treatment, it is, as far as the Consent Principle by itself is concerned and setting to the side, as he says, the independent and prior fact that it is morally wrong to use Grey to save White without her consent, Grey has no strong reason to prefer to lose her leg as she would in Earthquake as opposed to how she loses her leg in Means. I disagree. While both result in the same harm resulting from the loss of her leg and the physical pain of having her leg crushed, the second is far worse, as on top of those "evils" Grey has also been treated as a mere means to saving White's life. And White should be sensitive to this fact and so refuse to be saved in such a manner, were the choice up to him; that is, asked whether or not to be saved in those circumstances, White should say that the matter is up to Grey, even if he would prefer to

continue living. Grey's actual consent is essential to the permissibility of the rescuer's treatment in Means. If Grey refuses or is simply never consulted, it is intolerable, and Grey has very strong reason to protest, being pushed in the wreckage and having her leg crushed in order to save White. If she refuses or simply is never asked, then Grey is seriously wronged by being used to save White in Means. But this is not true in Earthquake. In Earthquake, Grey does not need to be consulted to see if she should be allowed to have her leg crushed so that White's life may be saved instead. It is no doubt regrettable in Earthquake that Grey's leg was crushed, but she has not been wronged, as she did not have a claim on anyone to save her leg, although she does have a claim, and a claim on everyone, we can suppose, to not cause her leg to be crushed without her actual consent, even if it was to save another's life. This difference, that Grey is wronged in Means but not in Earthquake, makes the harms Grey suffers very different.

It might be claimed that, in arguing as I have above, I am not really "setting aside" the fact that it is morally impermissible to save White by causing Grey's leg to be crushed as is the case in Means. However, I don't think that that fact is really independent of the FH. Grey has authority to determine whether or not her leg is crushed to save White's life in Means but not Earthquake because saving White's life in the manner described in Means, absent Grey's actual consent, is to treat Grey as a mere means and so is impermissible, while saving White's life in the manner described in Earthquake, whether absent Grey's consent or in the face of her protest, is not to treat her as a mere means.

Parfit's methodology is to appeal to facts about moral wrongs to determine whether or not an agent has sufficient reason to assent or dissent from some treatment whose morality is under question. (He writes: "Since acts can be wrong in other ways, or for other reasons, what the

Consent Principle implies may in part depend on which acts would be wrong for such other reasons. So when we apply this principle, we must sometimes appeal to our beliefs about which acts are wrong." (201).) For example, he claims that, insofar as we accept a morally relevant distinction between doing and allowing, one might claim that it is morally wrong to save White's life in Means but not Earthquake. Those facts then in turn ground whether or not it is rational to consent to the treatment, which in turn grounds the moral permissibility and impermissibility of the action, given the Consent Principle. In that case, then, those facts about moral wrong must be prior to and more basic than the Consent Principle, as they are fed into the Consent Principle as input before the verdicts of moral permissibility and impermissibility are delivered by the Consent Principle as outputs. This strategy is not available to those who, like O'Neill and Korsgaard, conceive of the Consent Principle as explicating what it is to treat a person as a mere means and so providing the content of the FH. That is because the FH is to be conceived of as the fundamental principle of morality, in which case there are no moral facts prior to its application and so no such facts to ground the claim that the agent has sufficient reason to assent or dissent.

If we help ourselves to a rich set of facts about moral permissibility, rational consent perfectly tracks moral permissibility, in the sense that an act A is morally permissible exactly when it is possible to rationally consent to it. But if what we want is a principle that *explains* the underlying grounds of impermissibility, that makes intelligible, in a satisfying and full way, *why* an impermissible act is impermissible, the Consent principle is unacceptable. There are two further reasons, independent of issues concerning the distinction between merely allowing a person to suffer a harm and causing that harm and so independent of distinguishing between Earthquake and

Means, to support this contention.

The first concerns duties and wrongs to the self. Kant famously thought that there are duties to the self and that one can perform an action that wrongs no one but oneself. His cases are infamous; in the *Groundwork*, he present the case of committing suicide in order to avoid a future filled with more suffering than goods and the case of failing to develop one's talents. I won't try here to defend Kant's cases, although I think that both are pretty convincing when narrowly understood. Instead, let's just grant that it is possible to wrong oneself. Such cases cannot adequately be understood in terms of consent. Consent necessarily involves two parties. The idea that I consent to my own treatment of myself, at the time of acting, fails to get any purchase, as any purchase it receives is empty, as I would seem to default consent to the treatment just by having chosen it. My treatment of myself my well be irrational. (I think, in fact, that all cases of wronging oneself, which is a nonempty class, are cases of acting irrationally.) But the irrationality has nothing to do with my consenting or failing to consent to that treatment. An adequate understanding of the FH, at least if we are going to accept with Kant the idea that that principle is the fundamental principle of morality and that that principle is a principle of practical reason itself, should explain the irrationality of my immoral treatment of myself, which the Consent Principle is unable to do in these cases.

A second, complimentary and I think more compelling case is that of Armin Meiwes and Bernd-Jurgen Brandes, who, in March 2001, engaged in mutually consenting acts of sexual cannibalism. Because Meiwes is the member of the pair that survived the experience (he was found guilty first of manslaughter and then, after a second trial, of murder and sentenced to life in prison), we know more about his history and state of mind. As a pre-teen, Meiwes became obsessed with cannibalism as a means of forever

being with another. He fantasized about acts of cannibalism, created elaborate scenes enacting acts of cannibalism on mannequins he built, which he filmed and posted online, and he met with people to play-act cannibalistic acts for the sexual gratification of each. All of this led up to his posting an advertisement on an online forum *The Cannibal Café* reading: "Looking for a well-built 20 to 30 year old to be slaughtered and then consumed." The ad received multiple replies. Meiwes met with at least two people, bringing each back to his house, where he had constructed a butchery, discussing his plans for slaughter. When each backed out, Meiwes spent the remaining evening with him watching movies and then each went his separate way. Meiwes wanted a *willing* victim, someone who wanted to be killed and eaten, which is what he found in Brandes. Brandes drove to Meiwes home, where he took sleeping pills, drank half a bottle of schnapps, and then encouraged Meiwes to bite off his penis, which Meiwes tried unsuccessfully to do, ultimately resorting to a knife. Brandes and Meiwes then tried unsuccessfully to eat the dismembered appendage raw, as they planned during their online chats leading up to the visit, to eat Brandes's body together. Brandes moved to a warm bathtub to bleed to death and Meiwes sautéed Brandes's penis in some of Brandes's fat, which he burnt and so fed to his dog, and then read a book, checking frequently on Brandes in the tub, waiting for him to bleed to death. Meiwes ultimately moved an unconscious Brandes from the tub to his slaughter room, killed him by stabbing him in the throat, hung his lifeless body on a meat hook, butchered his body, and stored it in his freezer, eating over 40 pounds of Brandes's flesh over the course of the next ten months. Meiwes recorded the four-hour affair, which was used in his second murder trial.

It seems clear to me that, despite Brandes's consent and willing participation, the whole affair was deeply immoral.

Brandes's was used as a mere means for sexual gratification. While I think that Brandes was equally guilty, let's here focus on Meiwes. Despite Brandes's consent, Meiwes did not respect Brandes in killing and eating him. A proponent of the Consent Principle may agree, insisting that Brandes´s actual consent is irrelevant, as it was not rational consent and it is not possible to rationally consent to being slaughtered, butchered, and eaten for sexual gratification. (Cannibalistic acts due to starvation, from the Starving Time of the winter of 1609 in Jamestown to the famines in the Soviet Union in the 1920s and again in the 1932 and 1933 during the Holodomor, are arguably different. Meiwes did not eat Brandes's flesh out of hunger and a lack of alternate nutrition.) But why is Brandes's consent irrational? For the reasons given above, a proponent of the view that the Consent Principle is the fundamental principle of morality cannot say that it is because the act is immoral. Feeding that fact into the Consent Principle gets matters upside down. I think that any account of why the consent was irrational will itself be the ultimate grounds of the immorality of the act upon which the impossibility of rational consent will supervene, which is inconsistent with the claim that the Consent Principle is a fundamental principle of morality.

I have argued that we should not see the impossibility of rational consent as the explanatory grounds of immorality and so that we should reject the consent conception of the FH. I already noted that Parfit would not disagree. In chapter 9 of *On What Matters*, Parfit discusses the Mere Means Principle, according to which it is wrong to treat a person merely as a means. He conceives of this principle as distinct from the Consent Principle, and so he does not claim that what it is to treat a person merely as a means is to treat her in ways to which she could not rationally consent. I shall briefly discuss Parfit's account of the FH, according to which both the Consent Principle and the

Mere Means Principle together constitute the content of the FH.

Parfit thinks that someone who uses *x* as a means to her ends does not *merely* use *x* as a means when her use of *x* is "restricted by her concern for their well-being" (213). He thinks that this is brought out be comparing the attitudes of two scientists to their laboratory animals, the one who treats her animals in whatever way is most effective to her experiments regardless of the pain endured by those animals and the second who avoids any treatment that would cause her animals to suffer, even if the painless methods are less effective than the alternatives. Parfit claims that the first but not the second treats her laboratory animals merely as a means. More fully, Parfit claims that the following, Principle (B), comes close to explicating the notion of treating a person as a mere means: "We do not treat someone merely as a means, nor are we even close to doing that, if either: (1) our treatment of this person is governed or guided in sufficiently important ways by some relevant moral belief or guided in sufficiently important ways by some relevant moral belief or concern, or (2) we do or would relevantly choose to bear some great burden for this person's sake" (214).

I think that principle (B) in no way captures an adequate conception of what it is to treat a person merely as a means. First, it is, on Kant's view, the *humanity of* a person that is treated as a mere means. A person is treated as a mere means only indirectly, by having her capacity for choice not properly respected when, for example, matters that are properly up to that person are settled by another. The Consent Principle at least gestures in the direction of choice and autonomy. Parfit's explication of treating persons as mere means in chapter 9 seems to omit even the gesture. Second, I do not share Parfit's intuitions. Both scientists, it seems to me, treat their lab animals as mere means. The second at least has the decency to show concern for the

animal's suffering, but both are used as mere means. (I am not taking a stand on whether or not the treatment is justified. If the lab animals lack the humanity that the FH demands we respect, then, as far as the FH is concerned, the treatment is not immoral.) Furthermore, Parfit claims (on page 215) that bandits who robbed his mother as she travelled on a boat, taking only half of the property of ordinary people, did not treat her as a mere means (or even come "close to doing that") by letting her choose whether they would take her engagement ring or her wedding ring. Again, I simply do not share his intuition. It is better that they showed some consideration toward his mother, and much better than bandits who take everything and leave their victims with nothing.

Parfit argues that (B) needs to be modified, as it is possible to satisfy those conditions but not act immorally and not treat persons as mere means. To illustrate, he considers the case of a gangster buying a cup who "would steal from the coffee seller if that was worth the trouble, just as he would smash the machine. But though this gangster treats the coffee seller merely as a means [in virtue of his attitudes towards the coffee seller], what is wrong is only his attitude to this person. In buying his cup of coffee, he does not act wrongly" (p. 217). Parfit uses this case to motivate adding the to (B).

The condition that the person be harmed by the treatment. As the coffee seller is not harmed by the gangster's treatment, even though he has vicious attitudes toward the coffee seller, the gangster's act is not wrong.

In *The Metaphysics of Morals*, Kant distinguished between the doctrine of right and the doctrine of virtue. In the first instance, the distinction concerns whether or not the duties can be externally regulated, where duties of virtue cannot and duties of right can. The state, for example, can provide incentives for people to pay for their coffee instead of taking it by force, which makes Parfit's coffee gangster

behave legally, but it cannot legislate and provide incentives, either in the form of promised rewards or promised punishments for failure, for virtuous behavior, as that is a function not just of the externally individuated actions like *pay for the coffee you consume* and *smash the coffee seller and take your coffee without paying*, but of the quality of the agent's will, the nature of her intentions, in acting as she does. We can impose costs on people who behave poorly in an effort to prevent them from so acting – to make them act right, in the juridical sense – but we cannot make a person act ethically in this manner. We cannot incentivize the gangster to have respectful and reverent attitudes toward the coffee seller; if he is responding to threats or promises of reward, then the quality of his will is still not virtuous and he does not have a respectful attitude, even if he behaves correctly.

Parfit's case runs roughshod over this distinction. Parfit is correct that the gangster does not act wrongly in the sense that he conforms to all of the duties of right; that is, he acts legally. His act is, nonetheless immoral, open to moral criticism, as it is not virtuous, and not merely in the sense that he acts with an absence of the quality of will that makes his choice and resulting action of moral worth, something true of all of us, but with attitudes that are positively criticizable. I don't pretend that Kant's distinction is perfectly clear and I also don't pretend that it is obvious how to relate duties of right and duties of virtue. But all we need to note is that it is possible to act in accordance with the duties of right even though one acts unethically by treating the humanity of a person as a mere means. Parfit's intuition that his gangster coffee connoisseur does not act immorally even though she acts with attitudes that do not express proper respect for the humanity of others, then, can be adequately accommodated. There is, then, no need to add to our conception of treating persons as mere means that the

person is harmed. And, furthermore, it is a mistake to add that extra condition, as it is possible to treat a person as a mere means even though the treatment does not harm her. One example is that of well-disguised paternalistic treatment. When Jill paternalistically doesn't tell John about the job offer that just came in for him because she rightly thinks that he is likely to take it and that the job is not right for him and he will be miserable, she treats John as a mere means, as though his will does not need to be involved in determining matters that are in fact rightly in its domain, even though she does not harm him and even helps him, especially if he never finds out about the omission and so is never upset by the poor treatment.

I end my discussion of Parfit by returning to variations of the Earthquake and Means cases from above. In chapter 9, Parfit describes more traditional versions of these cases in which the rescuer must chose between the lives of one and the lives of many. He describes three versions: In *Lifeboat*, one person is stranded on a rock in the rising tide and five people on another. It is impossible for the rescuer to get to both rocks in time and so can only save either the one or the five. In *Tunnel*, a driverless train is heading into a tunnel with five people on the track. The rescuer can divert the train onto another track and save the five, but the train will then strike and kill another, single person on that track. Finally, in *Bridge*, a train is again headed for five people and the only way to save them is to push a person onto the track so that the train's automatic braking system will be triggered when it strikes and kills the one.

The relevant differences between the three cases concerns the manner in which the death of the one is causally related to the saving of the five. In Lifeboat, the one is allowed to die; aid that would save her life is instead directed to save five. In Tunnel, the death of the one is a foreseen side-effect of the saving. While the presence of the single person on the track the train is diverted to is not part

of the means by which the five are saved in Tunnel, in Bridge, it is. So, in Bridge, the death of the one is the means by which the five are saved.

The standard view is that, while it is at least morally permissible to save the five in Lifeboat (and perhaps even that one should save the five), it is impermissible to save the five in Tunnel and impermissible to save the five in Bridge. Parfit argues, however, that "the Kantian principles," i.e., the Consent and Mere Means principles, entail that it is morally permissible to save the five in all three cases. Tunnel and Bridge are common counterexamples to consequentialist accounts, motivating either side-constraint conditions or nonconsequentialist accounts altogether. It would therefore be very surprising, and frankly disappointing, if one of the core Kantian ethical principles cannot deliver the intuitively correct verdicts on these familiar cases from the literature. I shall argue, however, that Parfit's argument does not convince.

Parfit argues that the lone person in Lifeboat could rationally consent to the five being saved, even though there is also, of course, sufficient reasons for her to chose that she is saved instead. So, as far as the Consent Principle by itself is concerned, saving the five is morally permissible. He then argues that the same is true in both Tunnel and Bridge, as the lone person again has sufficient reasons to save either herself or the five and so she could rationally consent to the five being saved. Parfit argues as follows: "It would make no relevant difference that [the lone person] would here [in the Tunnel case] be saving the five by redirecting the train so that it would kill [her] instead. This way of dying, we can suppose, would be no worse…. [The lone person] would have sufficient reasons to jump in front of the train, so that it would kill [her] rather than the five. And compared to killing [her]self as a side-effect of saving the five, in *Tunnel*, it would be no worse for [her], in *Bridge*, if [she] killed [her]self as a means of saving the five" (p.

220).

I think that this argument is unconvincing, as it too quickly aligns the harms in the three cases. While the outcome to the lone person in the three cases is, at a coarse level, the same, as, in all three cases, the lone person dies, there are still important differences between merely being allowed to die, in Lifeboat, the death being foreseen as a side-effect of a positive act, in Tunnel, and the death being a means, in Bridge. However, I am going to set this aside to focus on what I think is a more important problem with Parfit's primary argument, from pages 221-222. So, I will grant Parfit his contention that, because the lone person in Lifeboat can rationally consent to the five being saved, so too can the lone person in Bridge rationally consent to the five being saved. As the Consent principle entails that an action is wrong when someone could not rationally consent to it, it follows, Parfit claims, that, as far as that principle is concerned, the lone person is not being wronged when the five are saved. Parfit concludes that "Even if [she] would not in fact consent, the Consent Principle would not condemn this act" (p. 221). And, because a person is not treated merely as a means when the treatment of that person is governed by the Consent Principle, as, in that case, the treatment is guided by a moral concern for the person, it follows that the Mere Means Principle does not condemn killing the one to save the five in Bridge. Parfit endorses this argument, claiming that "It might be wrong for [the rescuer] to kill [the lone person], without [her] consent, as a means of saving the five. But this is not implied by these Kantian principles" (p. 221).

There are several mistakes in this argument. First, as I argued above, Parfit is wrong to claim that there are no reasons to find the death of lone person worse in Bridge than it is in Lifeboat, precisely because it is worse that one's death is a means to others being saved than it is to be allowed to die in order to save others. Second, it does not

follow from the fact that I could rationally consent to *x* that, even if I would not in fact consent to *x*, the Consent Principle would not condemn *x*. Let's grant that it is possible for A to rationally consent to having sex with B. It does not follow that the Consent Principle would not condemn B's having sex with A even if A did not in fact consent. It is one thing to have consensual sex with someone and quite another to have nonconsensual sex with that person or even to have sex unconditioned by that person's consent. Similarly, it is one thing to kill oneself by jumping of one's own free will in front of a train to save five and quite another to be pushed in front of the train to save five without one's consent.

In the Lifeboat, there is sufficient reason for the lone person to choose that the five be saved instead of her, if we imagine the rescue to be entirely her choice. From this, it follows that she could rationally consent to the five being saved. And it is also true that it remains permissible to save the five without the lone person's consent and even in the face of her protest. But contrary to Parfit's suggestions, this is not a consequence of the first claim. Instead, the irrelevance of the lone person's actual consent to the morality of saving the five stems from the fact that the lone person does not have a right to the life saving aid (under the conditions Parfit describes, where no one was at fault for the lone person being in the state she found herself and no one promised, say, or was hired to save her). So, suppose for the sake of argument that Parfit is right that the lone person in Tunnel and Bridge also have sufficient reason to choose that the five be saved. From that, it would follow, assuming the Consent Principle, that it is permissible for the lone person to save the five and it may even follow that it is permissible for the rescuer to save the five *given the lone person actually consented*. But it does not follow that it is permissible for the rescuer to save the five "even if [the lone person] would not consent," particularly

in Bridge. There is a morally relevant difference between one's demands for the aid of others and one's demands against actions that will lead to one's own bodily harm and death, even if a precise carving of that distinction is hard to come by. So, Parfit's transition from the claim that the lone person in all three cases has a sufficient reason to save the five to the claim that saving the five is morally permissible even in the absence of the lone agent's actual consent and even if she would not consent is unjustified.

There are two reasons for why this transition is unjustified. One concerns the individuation of "the treatment" involved. When Parfit argues that the lone agent in each of the three cases has a sufficient reason to save the five, the "treatment" being agreed to is, in Tunnel and Bridge, voluntarily sacrificing oneself or agreeing to be sacrificed. But that of course isn't the "treatment" that, by the conclusion, Parfit argues is morally permissible, as the lone person's agreement or consent has been dropped. The second reason why the transition highlighted in the previous paragraph is dubious concerns the domain of a person's authority. I claimed that the lone person in Bridge has authority over the integrity of her body that is morally different from the authority the lone person in Lifeboat has over the aid of another, even if both are conditions necessary, given the circumstance, for her continued existence. That is, it is extremely plausible that the right to life that everyone has covers not being killed to save five others but does not ensure that needed lifesaving aid will be provided. I agree that it is a very difficult matter to carve the contents of these rights out fully and precisely and it is an even more difficult matter to say what grounds those rights. But that work is unnecessary here, as my claim is that Parfit rides roughshod over these matters. What is clear, I claim, is that there is some truth to the claim that the rights of the lone person in Lifeboat are not violated while the rights of the lone person in Bridge are when the

five are saved without the lone party's consent.

I conclude that Parfit is unjustified in his claim that the Kantian principles fail to entail that it is wrong to kill the lone party, without her consent, as a means of saving the five. Such behavior is a paradigm of using a person as a mere means and Parfit has not shown that the FH, even interpreted in the manner Parfit interprets it, does not imply as much.

I have said a lot about what I think the FH is not. It is now time to say something more positive. My aim is to support what I call the Evaluativist conception of the FH and then argue that that principle is a principle of pure practical reasoning. In particular, I argue that conforming to the FH characterizes an ideal form of action, autonomous action, that we necessarily aspire to, given the kind of beings we are.

Practical reason is animated by an aim to act in accordance with reason, where being motivated to act in accordance with reason is to be motivated to make sense of one's behavior. (See Velleman 1989 and the essays collected in Velleman 2000.) When I move my body around the world, I have in mind a goal the movements are thought to achieve. The behavior is nonarbitrary, in at least the minimal sense that I, perhaps only implicitly, have available as I act some explanation of why I am so behaving in light of those ends and a conceptual fit between my ends, so described, and my behavior, so described. Without considerations that would explain why I am so behaving involved in the production of my behavior, answers, if you like, to the question *Why?*, I am a mystery to myself and my behavior cannot be seen as stemming from me. But things don't stop there. Unlike my cats, I don't have my ends simply given by instinct and bodily needs. I can reflect on and choose my ends; for better and sometimes worse, reason is involved in the adoption of ends. Complete self-governance requires a *complete* explanation of what one is

doing and why; an end to the questions *What are you doing and why are you doing it?*. Such a complete explanation, then, must be grounded in an *end in itself*, something that is unconditionally and absolutely worth doing and thus something to which the question *Why are you pursuing that?* has the same sense as the question *What are you doing what you have most reason to do?*; that is, a question that can have a nontrivial question only if the 'why' is not taken to be one of asking for a reason. And regressing on this chain of explanatory relations bottoms out in the value of choice and so the FH. That is the line of argument I shall develop in the remainder of this essay. The FH, then, is a principle of pure practical reason because it is necessary to have an adequate and complete explanation of what one is doing and why and practical reason is constitutively animated by a motivate to make sense of oneself.

A reason for acting is a consideration in light of which an agent can make sense of what she is doing. Insofar as the agent acts autonomously in acting on a consideration, the consideration's effectiveness in producing behavior must be compatible with the agent being the nonarbitrary source of that behavior. In that case, the consideration must be seen as speaking for the agent. So, an agent's acting autonomously constrains what considerations she acts on: The considerations must shed explanatory light on what she is doing and why while being consistent with her viewing her behavior as flowing from herself. And not just any consideration can simultaneously satisfy both of these conditions, as we can see by surveying cases of nonautonomous action, trying to isolate what they lack. I will argue that what is lacking in all of these defective cases is a set of motivating reasons with which the agent is identified from which she can explain what she is doing and why in a way consistent with seeing herself as the ultimate source of the action. I will assimilate immorality to these cases. Immorality is compatible with (and in fact requires)

negative freedom, in the sense of acting of one's own free will, not being compelled and coerced to behave as one does, but it excludes positive freedom, acting autonomously, which requires the availability of a full explanation of what one is doing and why that is compatible with the agent herself being the source of the action. This condition of full identification requires conformity with the FH that is incompatible with acting immorally.

I begin by constructing a menu of nonideal examples of agency, familiar from the action literature. Let's begin with cases of acting from an unconscious motive. I find myself in front of the local ice cream store while out for a walk through my town. Later reflection tells me that I have been, from the beginning of my walk, guided by the aim of securing a frozen treat, although I wasn't aware of that fact as I was walking. I unconsciously decided to get a treat when I left my house and set out to realize that intention unaware of its presence and force, but for all of that quite deliberately and intentionally. Because its operations were unconscious, I cannot view my behavior as both directed by that intention and stemming from myself. Ascribing that motivate to my actions makes sense of what I did and why, but at the cost of extracting myself from its production. My behavior is intentional but not autonomous. In order for an intention's functioning in deliberation and action to count as my controlling my deliberation and action, I must be, at some level and however implicitly, aware of its operation at the time of action.

Unconscious motives provide one class of intentional but nonautonomous action. But there are other cases where the motive is perfectly conscious but its role in the production of behavior does not amount to the agent's guidance and direction of her behavior. Consider first Harry Frankfurt's unwilling addict (Frankfurt 1971). The unwilling addict chooses to perform complex intentional

actions, guided by the aim of getting high. Her behavior is done for reasons and she is aware of the existence and operation of the motives that animate her drug-taking behavior. But, for all that, she is not really in control of herself. Because of her alienation from those motives, explaining her behavior in light of them excludes her from also being the source of her behavior. When she acts on those motives, she acts, in some deep sense, despite herself. Acting autonomously requires that the behavior in question flow from the agent's sense of what to do while the unwilling addict's behavior flows from alien, because rejected and disvalued at the time of action, motives. Such cases show that there is a wedge between minimal notions of acting for reasons, acting intentionally, and acting intelligibly, on the one hand, and a more robust notion of acting autonomously.

Consider next cases of perversity. Perversity comes in degrees. Milder forms are cases of clear-eyed weakness, where one acts of one's own free will on considerations that one judges, at the time of deliberation and action, to be outweighed, in the circumstance, by competing considerations. I grab another cookie of my own free will. I am not overrun by passion; I do not act compulsively or in spite of myself, as is arguably the case in the unwilling addict. To echo Austin, I calmly and carefully eat down the whole plate. But I am not identified with the motives that move me. Even while I reach out and start chewing, fully aware that there is some real and genuine sense in which I am able to stop myself and resist the temptation, I really wish I would refrain from eating more and despise my weakness. Yet still I eat. Such cases demonstrate two points. The first concerns Frankfurt's claim of the connection between identification and acting of one's own free will. (See (Frankfurt 1971, 1977, 1987).) The second is, I think, more germane to my primary aims in this paper. I discuss each in turn.

Frankfurt employed the notion of identification in an account of acting of one's own free will. While Frankfurt didn't fully spell out what he takes to be involved in addiction, it is plausible that addiction undermines freedom of the will, or so it is at least not implausible to insist. This contrasts with cases of ordinary weakness. The phenomenology is that I simply allow myself to be determined by my desire for sweets—I am not overrun by rebellious motives—and so I act of my own free will. In both cases, the agent acts independently of her evaluative determination and acts on motives with which she is not identified, but only in cases of addiction does the agent fail to act of her own free will.

The distinction between the unwilling addict and the person who willingly acts contrary to her evaluative judgments is related to Watson's distinction between compulsion and weakness (Watson 1977). The distinction proves problematic for Frankfurt's claim that acting on a motive with which one is identified is necessary for acting of one's own free will, as the merely weak agent acts of her own free will but on a motive with which she is not identified. Furthermore, weakness is compatible with moral responsibility; the weak, but arguably not the compelled, agent is morally responsible for her behavior. (This also proved problematic for Watson's own evaluativist conception of identification, as he did not clearly and explicitly draw the distinction between negative and positive freedom. The weak willed agent, unlike the compelled agent, acts of her own free will, but neither act autonomously.)

Identification is not necessary for acting of one's own free will and being morally responsible. But acting autonomously does require being identified with the motives that move one. Frankfurt did not distinguish between acting of one's own free will and acting autonomously, which hampered his ability to do justice to

the distinction between the unwilling addict and the merely weak agent. That failure also kept him from seeing the real role identification plays in the proper functioning of human agency.

Because I judge it better to not eat the cookies, I judge myself to have not done what I should. I am normatively deficient. But I am also a bit of a mystery to myself. The problem I have isn't just a normative problem; it isn't just that I am unable to fully justify myself when challenged. My problem is also an explanatory problem; indeed, on my view, it is a normative problem because it is an explanatory problem. "Why am I doing this?" I ask, without being able to adequately answer. Sure, the behavior has something going for it; it isn't like sticking tacks under my fingernails, for which there is really nothing that speaks in its favor. The wonderful taste and fabulous sensations of eating the cookies in some sense renders my behavior intelligible, as after all, that's something I enjoy and even crave. The behavior isn't aimless; it is directed at some end that has something to be said for it. Furthermore, given my past similar action, it is hardly surprising that I broke down; I always seem to when faced with a plate of cookies and a little free time. Anyone who knows me, myself included, would predict exactly what in fact happened. But there is still a mystery. What remains unexplained is why I am allowing myself to be directed by those considerations. Citing the pleasant sensations and good flavors fails to make sense of my behavior in light of the aim of getting some confectionary pleasure. They provide a mere psychological explanation that is not compatible, in light of my negative assessment of the weight of those considerations, with my furthermore being the source of my actions. Those considerations are not connected up with my other views to properly enable me to make sense of myself as in control of my behavior while acting on those considerations. They cannot, then, "speak for me"

and their involvement in the production of my behavior does not constitute or amount to my directing myself. Even worse for playing this role of "speaking for me" are the considerations that lead me and all of those who know me to predict that I will eat the plate of cookies. While noting my weak nature and recalling all the past times I similarly failed to control myself may make my behavior predictable and reveals something about me, in the end it only makes my failure all the more mystifying, as it still sheds no light on why I allowed myself to be ruled by something I deem unfit to lead given that it is not overpowering me and, as a rational agent, can hardly entice me except through considerations that would lead me to endorse. Being defeated by an alien force I could not resist, as the compulsive agent is, is one thing, while willingly giving into a consideration I deem to be outweighed is quite another. While mere weakness is far more common that compulsion, it is, for all that, far more mystifying, at the same time.

All of the cases of defective agency share the following feature: The agent acts on motives that she does not, at the time of action, consider as authoritative. This feature makes it so that the agent lacks, at the time of action, an account of what she is doing and why that is consistent with her viewing her behavior as also being self-determined. Absent endorsement of a motive, then, one does not act autonomously in acting on that consideration. When one voluntarily acts contrary to one's evaluative outlook, we cannot simultaneously see that behavior as being based on the deemed lesser considerations—the considerations that the cookies taste good, in the example above—and as stemming from the agent herself.

Acting autonomously requires acting from one's evaluative perspective. But then there is rational pressure to act from one's evaluative perspective in virtue of one's being an autonomous agent. This is because autonomy is

the default position of practical deliberation. While we can fall short of our autonomy by acting nonautonomously, whenever we are making choices, we are subject to the demand to act autonomously and we cannot set out with the aim of being nonautonomous or even the presumption that we won't be autonomous whenever we deliberate about what to do and make decisions. Going into a situation with a fatalistic, as it were, attitude is already to give up deliberation and deciding what to do; going into a deliberative situation with such an attitude is an instance of taking up a passive attitude of waiting to see what happens, rather like when one watches a movie. Insofar as practical reasoning is productive thought, that is, thought directed at realizing its own truth, then, practical reasoning must presuppose, for an agent capable of having the thought (i.e., a normal functioning, self-aware and self-conscious human adult), that one is autonomous and self-directing.

To summarize: I act intentionally in walking with the unconscious motive of getting a treat, but not of my own free will and not autonomously. I act of my own free will in eating the cookies but I do not act autonomously. I don't act autonomously because I act on a motive with which I am not identified. I am not identified with my effective motive because I judge eating to not be among the best things to do given my circumstances. There is a constitutive connection between acting autonomously and acting from one's evaluative perspective, as only then can one have a complete explanation of what one is doing and why that is consistent with viewing that behavior as being self-governed. Evaluative endorsement and self-governance are conceptually related because evaluative judgments constitute the agent's principled stand on what to do and so are capable of being the states from which the agent directs her own behavior. So, I proposed a restricted version of evaluativism: While it is perfectly possible and all too actual to be motivated by, act intentionally and even of one's own

free will in ways that are independent of one's evaluative outlook at the time of acting, there is a conceptual connection between acting autonomously and acting "under the guise of the good." Acting autonomously requires acting from one's evaluative judgments, acting only on motives with which one is identified. So, in virtue of our autonomy, we are subject to the demand of acting from our evaluative judgments. This demand stems from the need for an adequate and complete explanation of what one is doing and why and is thus grounded in an explanatory conception of acting for reasons.

What does this have to do with my main claims about practical reasons, practical rationality, and practical reasoning as such, on the one hand, and morality and the FH? Because practical reasoning requires a presupposition of autonomy, and in particular that the resulting behavior will be autonomous, self-governed action, and autonomous action is action under the guise of the good, self-reflective and articulate practical reasoning involves evaluative judgments concerning the good of one's actions and the ends for which one acts, the FH plays a crucial explanatory role necessary for fully autonomous practical reasoning. The role it plays is exposed by the regressive argument for the FH. On this picture, evaluative judgments get into the mix because they alone can play a certain explanatory role; only when one acts from one's values can one have an explanation of what one is doing and why that is compatible with conceiving of oneself as being the source of the action. But one's evaluative judgments themselves are subject to explanatory coherence. That is, a fully reflective and rational agent will not just explain her choices in terms of her evaluative judgments but will also have an explanation of her evaluative judgments. I explain my walking across the room in terms of the good of getting a drink of water. But the demand for explanation is not satisfied as explanatory coherence requires that I have an

explanation of why it is good to get a drink of water. The role of the FH, then, is to unify and ground all evaluative judgments. The ultimate grounds of the particular evaluative judgments invoked to explain one's particular choices, the argument goes, is the value of humanity. The FH, then, is a commitment of autonomous practical reasoning and so morality is rationally nonoptional. It is nonoptional because any action that does not conform to the FH lacks a full and adequate explanation compatible with the agent viewing herself as the author of her action.

Kant introduces the FH in paragraph 49 of section II of the *Groundwork*. He reconsiders his four examples of immoral action, that of suicide in order to "escape from a trying condition," making a false promise to secure one's end, that of failing to develop and maintain one's natural talents, and that of nonbenevolence, this time using the FH instead of the formula of universal law. Seven paragraphs later, he introduces his third formulation of the Categorical Imperative, the Kingdom of Ends formulation. Then, in paragraph 68 of section II, Kant introduces a distinction between *price* and *dignity*. I mention this crude explication of the text only to emphasis that it is far from clear that this last distinction is meant to reflect on the FH. But that is exactly how I shall take it. I think that we should read into the distinction between treating the humanity of a person as a mere means and as an end in itself the distinction between price and dignity. The only thing with dignity on Kant's view is rational nature; i.e., humanity. And Kant explicitly connects dignity, humanity, and being an end in itself in these paragraphs. That much, I think, is beyond dispute. It is quite natural, then, to employ his notion of price, and especially of a market price, in understanding his notion of treating the humanity of a person as a mere means. To treat a person as a mere means is to treat her as though she had a market value; i.e., to treat her in ways in which, were one maximally reflective and explicit, involve

judging the worth of one's contingent end to be more valuable than her humanity. Insofar, then, as we are rationally committed to valuing humanity as an end in itself in virtue of our autonomous rational agency, treating persons as mere means is practically irrational.

An entity has a *market price* when it is useful to satisfy a human inclination or need. Part of what it means to say that it has a market price is that its worth is relative; in particular, it is relative to the worth of the inclination or need it serves or furthers. Part of what this means is that the worth of something with a market price is an inherited worth, exactly the way that money is valuable only because of the useful goods it can be exchanged for. This structure is mirrored in the means-end structure, where the value of the means is derived from the value of its end. An otherwise valueless or perhaps even negative, when considered by itself, means, like getting a shot, say, or depriving oneself of some immediate pleasure, becomes valuable and worthwhile because of the end it brings about.

The ordinary ends that we pursue are typically only derivatively valuable, in that they inherit their value from their connection to some larger, more comprehensive end. For example, it is useful to get a shot because it will help prevent one's getting sick during the flu season, which is a worthwhile end because it is part of maintaining one's health, which it turn is a worthwhile end because maintaining one's health is necessary to be an effective agent. Thus, we regress on the value of our ends, seeking more comprehensive ends that explain our more particular aims that back our particular decisions and actions.

The thought that there is a single grounding value, the value of humanity or rational nature as such, at the bottom of this regress on any adopted end, is the basis of the *regressive argument for the value of humanity* that Christine Korsgaard ascribes to Kant and developed in Korsgaard (1986). Evaluative judgments are subject to demands of

explanatory coherence. One cannot reasonably think, "It is good for me and bad for you simply because it's me." Those grounds do not enter into an explanatory web with one's other beliefs and attitudes. Hence, if that's all that can be said about one's ends, one's ends are unreasonable. To be reasonable, then, one's evaluative judgments must be supported by explanatory relations to other, more general evaluative perspectives. Reason seeks generalities in support of the particular. When one acts, then, on the judgment that it would be best to eat a bowl of clams, that judgment must be supported by and connect with a more general evaluative picture that explains why that judgment is good. Reasonable evaluative judgments, then, must have supporting general grounds. Reason is not satisfied until all *Why?* questions are answered, which, in the realm of valuable ends, requires that there be an end that is valuable in itself, which is rational agency as such. The thought is that only humanity, conceived as the capacity to adopt ends for reasons, is good in itself and so can be the explanatory grounds of one's contingently adopted ends.

So, insofar as one is perfectly rational, reflective, and articulate autonomous agent, then one acts on ends that one judges to be good and to have their worth to be ultimately grounded in the value of rational agency as such. Suppose now that all immorality involves treating the humanity of a person as a mere means and that to treat the humanity of a person as a mere means is to involve the person in one's ends in such a way that, were one maximally reflective and articulate, would involve judging the person's worth to have a market price, as one would be judging that the person's capacity to set his or her own ends and to be a determiner of her own future were less valuable than one's adopted contingent end that one is acting on. Let's consider a case. I need a new roof on my house and decide that the best way forward is to falsely promise to fix my neighbor's cars in exchange for his fixing

my roof. I treat him as a mere means to my end of fixing my roof. On the view under discussion, this means that I treat him in a way that requires, assuming that I am maximally reflective and articulate, that the value of his capacity to determine his own involvement in my plans as a market price, as I am bypassing his capacity to determine his involvement by being untruthful about what my plans are, for the purpose of achieving my end of fixing my roof. In that case, then, maximal reflection and articulation of what I am doing and why involves contradictory evaluative judgments: I am judging that my all instances of humanity have a dignity (for the reasons given in the regressive argument) and yet the humanity of my neighbor has a market value. Those judgments are logically inconsistent.

The evaluative judgments involve in my making a false promise to my neighbor to get him to build me a new roof are contradictory because an entity with a dignity is a grounding value. To say, however, that an entity has a market price is to say that it is exchangeable. We can compensate a diminishment of overall value involved in acting against that entity by more values of another kind. But if *x* is the grounds for the value of *y*, then it is absolutely impossible to make up for the loss of *x* through having more *y*; *y* is valuable only *because* of the value of *x* and so taking away *x* entails that *y* is without value. It is like pursuing the goal of making more money at the expense of one's overall happiness. Money is worthwhile only because of its usefulness in finding happiness. So, it is deeply irrational to pursue money so relentlessly and at the expense of other aspects to one's life that one is unhappy, as though money were an adequate final end. Similarly, it is deeply irrational to pursue an end like fixing one's roof at the expense of disrepecting rational agency, as the grounds of the value of that end is found in the value of rational agency. One and the same entity cannot both have a market and a grounding value.

So far my argument for the irrationality of immorality is that fully reflective and articulate immorality involves inconsistent evaluative judgments, the contradictory evaluative judgments that humanity is the grounding value of some contingent end one has adopted and that some instance of humanity has a market price. But what about less than full reflective and articulate immorality, which is just to say most if not all immorality that actually occurs? While I have perhaps shown than a fully reflective and articulate agent rationally must not act immorally, it is less than clear what conclusion has to do with the rest of us. So, what I have said falls short of a defense of moral rationalism, as it leaves open the claim that a less than fully reflective and articulate agent's immorality is perfectly rational.

My idea is that normally functioning adult human is subject to the rational demands not only of consistency but also of explanatory coherence. What is true, then, of the fully reflective and articulate rational agent – namely, that their immorality is irrational because inconsistent – will trickle down to the rest of us through the rational demand for explanatory coherence. Very crudely, the idea is that, were the unreflective agent who lies to her neighbor to get her roof replaced but without thinking through the grounds of the value of that end, perhaps not even thinking that her end is valuable, and so does not in fact, in acting as she does, judge that humanity is the grounds of all values, for example, to fully satisfy the demands of explanatory coherence, then she would have the same inconsistent judgments as the fully reflective and articulate agent who acts from the same set of attitudes. That is because the Kantian story of the value of adopted ends and the value of humanity is the objectively true story, the only rationally acceptable story, and so full satisfaction of the demands of explanatory coherence demand that that be the account she would adopt, in which case she would judge that humanity

is an end in itelf, which then conflicts with her judgment that her end of getting a new roof built is more important than getting her neighbor's willing involvement in her end.

Every rational agent is subject to the rational demand for explanatory coherence. However, full reflection and articulation has its costs and we often reasonably fall short of satisfying that demands, as working everything out is simply not worth those costs. Here is a simple example. I am sitting here looking out my window as I type watching a hummingbird in my tree. I form the belief that there is a hummingbird in my tree without forming the belief, which is a self-evident logical consequence of that belief, assuming an adequate account of the conditional, that, if Trump is Donald Trump is the 45th president of the US, then there is a hummingbird in my tree. My system of overt beliefs fails to be explanatorily coherent, is it is not closed under logical consequence, or even self-evident logical consequence, whatever exactly that comes to. I fall short of full rationality because I have explanatorily incoherent beliefs. But it seems wrong to say that I am, for that, irrational, in the way I would be if I believed that it is not the case that, if there is a hummingbird in my tree, then there is a hummingbird in my tree; that is, I am not irrational in the way I would be if I believed a contradiction.

So, suppose A adopts F as her end in deciding to do such and such, but A does not reflect on this attitude and does not try to articulate and incorporate it into a maximally coherent system of beliefs. A does not, then, believe that F is (among) the best thing to do in her circumstances, or, if she reflect to that degree, she does not continue further in forming a belief about the grounds of the worth of that end. In that case, even if A's end is to, let's say, get a new roof by getting her neighbor to install it by falsely promising that she will fix his cars in exchange, her system of beliefs is not contradictory, as it does not include, as A has simply never reflected on the matter,

anything about the grounding value of humanity. To justify the thesis of moral rationalism in its full generality, as I intend to do, I need to show why even this agent is *criticizably irrational*, even though she does not have inconsistent beliefs and even though a mere failure of a maximally coherent set of beliefs is not sufficient for being criticizably irrational, as the case of my merely believing there's a hummingbird in the tree illustrates.

The problem with A's consistent set of beliefs that she has when she decides to make a false promise to her neighbor to get a new roof for her house is that there is no way to extend that system of beliefs into a maximally coherent system of beliefs without rendering the system inconsistent. This is because, in order to have an adequate explanation of the grounds of value of my adopted end, assuming the regressive argument is sound, is to believe that humanity is an end in itself. But if that belief is added to A's system of beliefs to satisfy the demands of explanatory coherence, A's beliefs become inconsistent because A also believes that it is more important to secure a new roof than to respect her neighbor's capacity to determine for himself his use of his agency. This is in stark contrast to my explanatorily incoherent set of beliefs concerning the hummingbird in the tree. Supposing I don't have any inconsistencies in my beliefs, we can simply extend my system of beliefs by adding the other conditional beliefs and we will ultimately get to an explanatorily complete system of beliefs that is also consistent (assuming that the additions are done by testing, for every pair of contrary propositions *p* and *not*-p, which can be added to the previous system of beliefs without inconsistency and adding that member of the pair and discarding the other). So, we can say that one is criticizably irrational in having explanatorily incoherent beliefs when there is no way to extend that system of beliefs by adding further beliefs

(without subtraction) without violating the demands of consistency.

The result, then, is that all immorality is always criticizably irrational. The fully reflective and articulate immoral agent acts on inconsistent evaluative judgments. The less than fully reflective and articulate immoral agent acts on an intention that cannot be part of any consistent and fully explanatorily cohererent system of beliefs.

REFERENCES

FRANKFURT, H. "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68, 5–20. Reprinted in Frankfurt 1988, 11–25, 1971.

_____ *The Importance of What We Care About* (Cambridge University Press), 1988.

KANT, I. *Groundwork of the Metaphysics of Morals, 1785*.

KORSGAARD, C. "Kant's Formula of Humanity" *Kant-Studien* 77, 183-202, 1986.

O'NEILL, O. "Between Consenting Adults" *Philosophy and Public Affairs* 14, 252-277, 1985.

PARFIT, D. *On What Matters, Volume 1* (Oxford University Press), 2011.

VELLEMAN, J. D. *The Possibility of Practical Reason* (Oxford University Press), 2000.

WATSON, G. "Skepticism about Weakness of Will," *Philosophical Review* 86, 316–39. Reprinted in Watson 2004, 33–58, 1977.