The People Problem

in *Exploring the Illusion of Free Will and Moral Responsibility*, ed. Gregg Caruso,

New York: Lexington Books, 2013, pp. 141-60

Benjamin Vilhauer

Philosophers who are skeptical about free will have to contend with an issue that I will call "the people problem". Do we have to treat human beings as if they have free will to treat them as *persons*, in the sense that matters for morality and the meaning of life? A lot is at stake in this question. If the answer is "yes", then a world in which human beings no longer treated each other as having free will would be impoverished in some important ways, and it would be natural for this to motivate some to argue against the view that we lack free will. If the answer is "no", on the other hand, then we have one less reason to resist this view. This question has received a fair bit of discussion in the literature, but a lot more effort has been invested in arguing for "yes" than for "no". My goal here is to contribute to the "no" side (with a caveat that I will mention shortly). I will argue that a primary concern of philosophers in the recent literature who argue for "yes" is the worry that if we stop treating people as if they have free will, then we are driven to consequentialism, and we end up instrumentalizing human beings and treating them as mere means to social ends. I will argue that this concern is misplaced, since there is a broadly Kantian ethics available to free will deniers which holds that we must always treat people as ends, and never merely as means.

Let me explain a few terms before proceeding. "Free will" refers to whatever satisfies the control condition of moral responsibility. I will sometimes use "treating as free" as shorthand for "treating as having free will". "Moral responsibility" refers to the relationship to our actions

which is necessary for us to be appropriately praised or blamed for our actions. "Free will believers" claim that we have free will. "Free will deniers" claim that we lack free will. "Free will skeptics" claim that we don't know whether we have free will or not. I will occasionally refer to both deniers and skeptics as "philosophers who take skeptical approaches to free will theory".

I will not try to establish that a particular sense of "treating as persons" is *the* one that matters for morality and the meaning of life. There are different senses that matter for different aspects of morality and life-meaning, and I do not want to deny that there is indeed a sense of "treating as persons" that does imply "treating as free", which really does matter for some aspects. My own position is free will skepticism rather than denial. I agree with Kant that it is possible that we have free will, but we do not know that we do. I also agree with Kant that the mere possibility of free will can be enough to justify us in treating each other as free.[1] But while Kant thinks the mere possibility of free will justifies us in treating each other as free across the board, I think it justifies us in treating each other as free only in some contexts, for example, when we seek to justify some kinds of praise.[2] The mere possibility of free will is not sufficient to justify retributive punishment, at least not when significant harm such as imprisonment is at issue.[3]

However, despite my own preference for skepticism over denial, I think that the free-will-implying sense of "persons" is necessary for much less of morality and life-meaning than many philosophers seem to believe, and that we can do quite well without much of the stuff for which it is necessary. I do not wish to claim that we lose nothing worth having. But I think we lose less than many think. My main effort in this paper will be to focus on a particular sense of "treating as persons" which does not imply free will in any way, and which is therefore readily

available to free will deniers—a sense which has been mentioned in the literature but does not seem to have received detailed discussion. I have in mind the Kantian notion that to treat human beings as persons is to treat them as ends in themselves, rather than as mere means to ends.[4]

I will not presuppose any background in Kant's philosophy. However, philosophers with even a passing acquaintance with Kant will probably know that Kant's own account of personhood and treating others as ends is not compatible in all respects with a philosophy that denies free will.[5] For example, Kant is a retributivist who holds that criminals can only be treated as ends when we punish them if we hold them morally responsible for their actions. He is also a strident advocate of the death penalty. I will discuss these points later, and argue that Kant would be better off without his retributivism. However, in light of Kant's actual position, it may seem confused to look to Kant's philosophy for a notion of "treating as persons" that does not imply "treating as free". But Kant's work is noteworthy for how much effort he expends in providing multiple explanations of his basic ideas, and how little effort he devotes to showing how those different explanations fit together. Commentators still argue about whether Kant's different explanations of treating others as ends cohere. To my mind, Kant's most helpful explanation of treating others as ends is in terms of treating them as they would rationally consent to be treated. Even if Kant thought that treating others as they would rationally consent to be treated implies treating them as having free will, I think it can be demonstrated that it does not. This thread in Kant's thought is helpfully drawn out in Rawls' notion of original position deliberation. Properly understood, original position deliberation serves to "filter out", as morally irrelevant, desert claims that involve attributions of moral responsibility (what I call our "action-based" desert claims) and focus our attention instead on what we deserve just by virtue of being persons (what I call our "personhood-based" desert claims). Choosing the basic principles for

society in the original position allows us to identify appropriate norms and practices for a society in which we can treat each other as persons in a robust Kantian sense without treating each other as having free will. But before we can see how this works, it will be helpful to briefly review a few points in the recent history of the people problem, which is cast in its distinctive form for the contemporary literature in P.F. Strawson's "Freedom and Resentment" (1997).[6]

1. Some Highlights in the Recent History of the People Problem

The idea that there is a special connection between treating human beings as persons and treating them as having free will been a part of philosophical reflections on free will for a very long time. But P.F. Strawson's "Freedom and Resentment" prompted a renewed interest in this issue. Most readers will know his basic story, so I will treat it in summary fashion.

Strawson frames his thoughts in terms of a debate between "optimists" and "pessimists". Optimists think that the concept of moral responsibility still works just fine if determinism is true, because praise encourages people to continue doing more of what we want them to do, and blame encourages people to behave differently. Pessimists not only disagree, but find the optimists' vision of the world deeply disturbing. They see the optimist vision as depersonalizing, and they hold that we can only legitimately hold people morally responsible if we have libertarian free will. Strawson thinks both sides in this quarrel have a point. He thinks optimists are right to insist that we must find a way of going on with our practices of holding morally responsible which does not presuppose libertarianism, because libertarianism is merely a fantasy. But pessimists are right to see the optimist vision as depersonalizing. Optimists recommend that we regard each other with what Strawson calls the "objective attitude", rather than the "reactive attitudes" which are essential components of "ordinary inter-personal attitudes". The reactive

attitudes are the attitudes involved in our holding each other morally responsible, including praise and blame, as well as any other attitudes that have praise or blame as essential elements. Strawson thinks we avoid the depersonalizing effects of the objective attitude by maintaining the reactive attitudes. But on his view, the pessimists are wrong to suppose that we need libertarian free will to justify the reactive attitudes. In some sense that is a bit mysterious, Strawson sees the reactive attitudes as self-justifying—the reactive attitudes themselves "fill the gap" in the optimists' vision. For purposes of studying the people problem, the key point here is the claim that we can avoid a depersonalizing objectivity of attitude if we maintain the reactive attitudes.

But why, exactly, would it depersonalize us if we relinquished the reactive attitudes and took up a permanent objectivity of attitude? Strawson's remarks here are not as clear as one might wish. At first pass, we can discern (at least) two sub-problems. One is about a profound sense of isolation Strawson thinks would ensue. The other is about consequentialism. Ultimately they may best be viewed as the same problem, the former being a consequence of the latter, but the literature seems not always to have connected them. The isolation issue has already received a fair bit of discussion in the literature, so I will discuss it briefly, and then move on to discuss the consequentialism problem which is my main concern.

Strawson claims that a world where the objective attitude was universal would be a tragic world of "human isolation" (Strawson, 129). He thinks this because he thinks that the reactive attitudes are essential elements of a very wide range of human emotions and relationships, including friendship and love (124). John Martin Fischer endorses this view, and claims that such a world would be "colorless and cold", and would be missing one of the "key normative ingredients in the notion of personhood" (Fischer 1995, 2-3).

But it is not clear that love and friendship are essentially reactive, i.e. that they are intrinsically bound up with holding people morally responsible. Consider love. Suppose Anisha successfully acts in ways designed to cultivate virtues such as fidelity and courage, and Bud loves Anisha because of these virtues. If Bud supposes that Anisha is morally responsible for these actions, then he can take his love to be a response to her that she can claim to deserve. But suppose that he instead believes that she has done nothing to cultivate her character. She has always been this way—it is simply her nature. Would this give him a reason to love her any less? As Derk Pereboom puts the point,

> Explaining love is a complex enterprise. Besides moral character and action, factors such as one's relation to the other…and her affinities with persons or events in one's history all might have a part….But even if there is an important aspect of love that is essentially a deserved response to moral character and action, it is unlikely that one's love would be undermined if one were to believe that these moral qualities do not come about through free and responsible choice. For moral character and action are loveable whether or not they merit praise. Love of another involves, most fundamentally, wishing well for the other, taking on many of the aims and desires of the other as one's own, and a desire to be together with the other. (2009, 20)

To take the next step in reflection, now suppose that Bud comes to see his love for Anisha as simply a consequence of his own nature, rather than something he has in any way willed or cultivated. It is not clear that any of this should make him see this love as shallower, or less valuable. As I discuss in more detail elsewhere, the same sort of argument can be made regarding many other attitudes that Strawsonians sometimes claim to be essentially reactive, such as friendship, and respect. (We can respect people for their fidelity and courage without liking or loving them for it, even if we do not suppose that it involves free will.) There are no doubt some attitudes that are essentially reactive, such as the attitudes involved in praise and blame, and some species of gratitude, anger, resentment, and vengefulness. A world devoid of these attitudes would be a different world, without question. But if the worry is that it would be

so devoid of emotional engagement that the human beings populating it could no longer be counted as persons, then the recognition that it could still contain love, friendship, and respect should suffice to allay this worry.[7]

Now I want to turn to the problem about consequentialism. Consequentialists hold that the rightness of actions is determined entirely by the consequences they produce. Non-consequentialists deny this. They hold (among other things) that there are some actions we cannot take, no matter what the consequences may be. Utilitarianism is probably the most widely discussed form of consequentialism. Utilitarians hold that the morally correct end is maximum overall happiness, and that whatever means maximizes overall happiness is the means we should take.[8] Strawson associates optimism and the objective attitude with utilitarianism. He explains that the optimist finds "an adequate basis for certain social practices in calculated consequences", an approach which he calls "a one-eyed utilitarianism"(140). When the optimist "undertakes to show that the truth of determinism would not shake the foundations of the concept of moral responsibility and of the practices of moral condemnation and punishment, he typically refers, in a more or less elaborated way, to the efficacy of these practices in regulating behavior in socially desirable ways. These practices are represented solely as instruments of policy, as methods of individual treatment and social control"(137). The pessimist "is apt to say…that the humanity of the offender himself is offended by *this* picture of his condemnation and punishment"(137). Strawson acknowledges that we should have "some belief in the utility of practices of condemnation and punishment"(139). But he goes on to say that "the social utility of these practices, on which the optimist lays such exclusive stress, is not what is now in question. What is in question is the pessimist's justified sense that to speak in terms of social utility alone is to leave out something vital in our conception of these practices". He thinks that

the "vital thing can be restored by attending to the complicated web of [reactive] attitudes and feelings which form an essential part of the moral life as we know it, and which are quite opposed to objectivity of attitude"(139). As Susan Wolf puts it, in commenting on Strawson,

> To justify praise and blame in the way the optimist suggests is to…leave out of account any question of whether it is an individual's fault that he has done something wrong or whether it is to the individual's credit that he has done something right. In short, to justify the praise and blame of persons in the way the optimist suggests is…to justify these practices only as a means of manipulation or training. The pessimist's fear may now be expressed as the fear that if determinism is true, this consequentialist justification of praise and blame is the only kind of justification that would be available to us. (390)

If I may gloss Wolf's claim in the final sentence of this excerpt in slightly broader terms, the worry seems to be that attributions of free will and moral responsibility are not just *a* way for us to avoid a slide into consequentialism, but are *necessary* if we are to avoid a slide into consequentialism. It is probably fair to assume that this is Strawson's view too, even if he never quite states it explicitly. But philosophers who take skeptical approaches to free will can resist this view.

Wolf goes on to ask us to

> Imagine for a moment what a world would be like in which we all regarded each other solely with the objective attitude. We would still imprison murderers and thieves, presumably, and we would still sing praises for acts of courage and charity…But these actions and words would have a different, shallower meaning than they have for us now…they would be bits of positive and negative reinforcement meted out in the hopes of altering the character of others in ways best suited to our needs. (1981, 391)

Tamler Sommers (2007) argues that Strawson and Wolf are too grim about the implications of rejecting the reactive attitudes. He considers a world of universal objectivity of attitude in which a character named Sally loses a wallet and a helpful stranger finds it and returns it to her:

> Sally should thank the woman, but not only because it may reinforce the behaviour. She should also thank the woman because she deeply appreciates the gesture. And while it is true that the woman is not ultimately deserving of praise

> for her actions (ultimately, Sally believes, it is a matter of luck that she became the kind of person who performs them), there is no reason for Sally to be coldhearted to her. She can warmly appreciate the gesture and the person who performed it without attributing desert-entailing responsibility to her. Sally can exult in the gesture, if she wants to; she can think 'What a nice world it is that produces clumsy, absentminded people like me who drop money-stuffed wallets, and sweet unselfish women like her who return them'. And the greater the heroes – the Danes, for example, who protected Jews during the Holocaust – the more profound one's feelings of appreciation will be. (2007, 229-330)

I take part of Sommers' point here to be that even if we endorse a consequentialist, behavior-reinforcement justification of thanking people who do good turns for us, we are not required to think of this as a calculating, cynical manipulation of them. We can have a sincere, emphatic, emotionally rich experience of this appreciation. This is what a utilitarian would encourage, since we will probably contribute more to the maximization of overall happiness in this way. But this point does not go far enough to resolve the consequentialism problem (though in fairness to Sommers, he does not claim that it does—he does not appear to take up this broader worry about consequentialism). I see Sommers' point here as adding another reply to the isolation issue. That is, one might think that the root of the isolation issue is the fear that a thoroughgoing consequentialist justification of all our interactions would mean that we should always be calculatingly putting on a show when interacting with others, trying to reason out what sort of emotional experiences we should be pretending to have, in order to cause the people around us to behave as we desire. If we looked out across our social world and saw that all our potential interlocutors were manipulative, dissembling play-actors, then this would be isolating indeed. I take Sommers' point to be that if that is not what we want our lives to be like, then the mere fact of relinquishing the reactive attitudes gives us no reason to try to make them that way. This is an important point, but the problem about consequentialism extends further. Sommers goes on to say that

> Kantian proclivities may begin to rebel at this picture, but the rebellion can be suppressed, at least for now. True, we are not attributing to these heroes a dignity and respect as autonomous agents. But this does not prevent us from admiring and applauding their characters and the actions that arise from their characters. We are grateful to the world for having such people in it, and we appreciate the heroes themselves for being what they are (even if they are not morally responsible for it). This is a deep, warm, *un*bleak, *un*barren, *un*ironic appreciation, and it is entirely consistent with denying free will and taking the objective attitude.
>
> (2007, 330)

I think Sommers is correct that Kantian proclivities need not rebel when what is at issue is whether we can legitimately have sincere and heartfelt emotional experiences in contexts where they raise no moral qualms, barring misplaced concerns about the objective attitude. But what about when confronted with the issues that more traditionally divide consequentialists and Kantians—issues about human rights, for example? To treat each other as persons is surely not only to have sincere, engaged emotions in our interactions, but also to respect each others' rights, and to develop social institutions together that safeguard our rights. What should free will deniers say about a situation in which, by depriving just a few agitators of their right to vote, or their right to free speech, we can dramatically improve social harmony, thereby maximizing overall happiness? Or, what if, by punishing criminals with torture, we will get such vigorous general deterrence that the overall crime rate will be dramatically lower, and overall happiness dramatically higher? Many philosophers seem to think that free will deniers have to accept consequentialism, and then choose between accepting disturbing answers to these questions, on the one hand, or on the other, taking on the burden of complex consequentialist attempts to offer less disturbing answers. But I think free will deniers should develop a non-consequentialist theory which allows them to attribute to human beings what Sommers calls "a dignity and respect as autonomous agents". Free will deniers can make such dignity and respect for people a central moral principle if they respect people as *rational* agents rather than as *free* agents, and if

they regard agents as autonomous not with respect to the laws of nature, but instead with respect to the undue influence of other agents. There is of course a long and rich consequentialist tradition of developing sophisticated strategies for accommodating rights—rule consequentialism, indirect consequentialism, and so on. But it is widely held that it is harder for consequentialists to explain rights in a satisfying way than it is for non-consequentialists. So it seems worthwhile to explore a non-consequentialist alternative.

2. A Kantian Deontology for Free Will Deniers

It may be easier to determine what free will deniers need in a non-consequentialist ethics if we a have better understanding of why so many accept the view that eschewing attributions of moral responsibility requires us to be consequentialists.[9] I think this view derives from an impoverished picture of what we might call "directions" of justificatory grounding in ethics. The impoverished picture is that, when justifying claims about how people ought to be treated, we have two options: we can either offer "forward-looking" justifications, or "backward-looking" justifications. Forward-looking justifications are consequentialist justifications: they justify treating people in particular ways based on the future consequences of that treatment. Backward-looking justifications justify treating people in particular ways based on events prior to the proposed treatment. Backward-looking justifications can in principle refer to any sort of event prior to the proposed treatment. In principle, they can refer to the event of being born under a bad sign. But the only sort of reference event that mainstream ethicists typically accept in backward-looking justifications is action—more specifically, actions that they take agents to be morally responsible for performing (if they suppose that there are such things). According to backward-looking justifications of this form, we deserve to be treated in particular ways because

of how we have acted. I will refer to such justifications as *action-based desert claims*. Since we can only deserve to be treated in particular ways based on our actions if we are morally responsible, action-based desert claims imply that the agent at issue acted with free will. Those who accept this "two directions" picture, and who think that action-based desert claims are the only legitimate backward-looking justifications, see action-based desert claims as the only way to constrain consequentialist justifications.

With a little reflection, however, this "two directions" picture looks like an oversimplification, because there certainly seem to be non-consequentialist justifications that do not look backward. Care ethicists think we can justify kinds of treatment by referring to the ways in which we care about, or are attached to, other people. Virtue ethicists refer to the virtuousness of treating people in particular ways. Deontologists refer to our duties to treat people in particular ways. The things these views refer to, in grounding their justifications, are not facts about the past or future in any obvious way. So it seems reasonable to speculate that we might develop a variety of any of these theories that avoids a depersonalizing consequentialism and still eschews attributions of moral responsibility.[10]

Some may object that the very idea of ethics without free will is absurd, since there is nothing we *ought* to do if there is no free will, because "ought" implies "can", and this "can" implies free will. But there is an epistemic interpretation of this "can" which does not involve free will: I can do x so long as it is possible *to the best of my knowledge* for me to do x. Even if determinism and incompatibilism are true, there typically will be a variety of mutually exclusive actions that I can take, in the epistemic sense of "can", at any given point in the future, given the limits of our predictive abilities.[11] Objectors may go on to claim that, in a non-consequentialist ethics, there cannot be any right or wrong way to treat people if we're not praiseworthy or

blameworthy for how we treat people. But this is simply to assert that no sense can be made of a non-consequentialist ethics that dispenses with action-based desert claims, and that is precisely what is in question here.

Objectors may now lodge a more specific objection: the only way for human beings to have the moral status necessary for a non-consequentialist justification to be plausible is for human beings to be morally responsible. The objection I have in mind states that when we justify something on the basis of it being required by, say, duty or virtue, we aren't necessarily offering an action-based justification, but for creatures to have the right sort of moral status for us to have duties to them, or for virtue to require us to treat them in particular ways, they must be creatures with the capacity for morally responsible action. Let me frame this objection in different terms. I think that everyone should acknowledge the pull of consequentialist justifications. Even staunch non-consequentialists typically accept the legitimacy of consequentialist justifications when it comes to the lower animals. If saving most of the inhabitants of an overpopulated earthworm farm means killing a few, most non-consequentialists will not be too burdened by non-consequentialist qualms, because they do not regard earthworms as having the right sort of moral status to impose non-consequentialist obligations on us. What makes human beings different? A common explanation in the philosophical tradition is that human beings have free will, which allows them to (for example) set their own ends, and consequentialist justifications go wrong in (for example) setting their ends for them.

I think the best way to reply to this objection is to work out a counterargument in detail, with respect to one of these non-consequentialist theories, and in this paper I am focusing on Kantian deontology. Kantian deontologists are non-consequentialists who hold that we distinguish permissible and impermissible means to our ends by applying the principle that

persons must always be treated as ends, and never as mere means. It may be objected that personhood itself implies free will, so that human beings cannot have the moral status of persons unless they are free. At points, Kant himself advocates this view. In the *Metaphysics of Morals*, he states that "A person is a subject whose actions can be imputed to him", while "A thing is that to which nothing can be imputed. Any object…which itself lacks freedom is therefore called a thing" (1996, 6:223).[12] For Kant, to impute something to someone is just to hold them morally responsible for it, and by "freedom" here he means free will. But Kant does not need to define persons in this way. He can define persons as beings who autonomously set their own ends, and who must therefore be treated only as they would rationally consent to be treated. Free will skeptics can follow Kant in making respect for our right to independently set our own ends a central moral principle. So long as we understand the influence that we are independent from, in autonomously setting our own ends, as the undue influence of others, rather than the influence of the causal nexus constituting the world, this view is entirely compatible with free will skepticism. We might think this was a specious distinction if we thought that doing things for reasons implies doing things with free will. But I have never seen an argument for this view which does not include controversial compatibilistic or libertarian premises, and it seems to be a straightforward matter for free will skeptics to draw a line between the kind of control involved in doing things for reasons and the kind of control necessary for moral responsibility.

We can have a robust Kantian deontology without looking backward to action-based desert. We need only look to the people with whom we are interacting to find a basis for desert-claims that constrain consequentialist justifications. Personhood can provide a basis for desert-claims which is irreducibly different from action, and which does not depend upon free will in

the way action-based desert does. Persons deserve to be treated only as they would rationally consent to be treated, just because they are persons.

Let me explain the distinction between personhood- and action-based desert in more detail. Action-based desert claims are very diverse—examples include claims about praise and blame, the Lockean claim that we come to deserve property when we "mix our labor" with objects, and the retributivist claim that criminals deserve the suffering of punishment based on their criminal actions. But on reflection, it seems clear that not all desert claims are action-based.[13] Some are based on the mere fact of being a person.[14] For example, when we claim to deserve the right to free speech, to be granted due process, and to be respected as persons (the sort of respect that Darwall calls "recognition respect"[15]), these claims are not based on facts about our actions, but instead on the fact that we are persons. Since personhood-based desert does not depend on how we act, we do not need to appeal to moral responsibility to make sense of it. Therefore free will deniers can endorse personhood-based desert claims even though they must reject action-based desert claims. Our rights to due process offer an especially clear illustration of personhood-based desert. It is part and parcel of our understanding of due process that there is nothing anyone could conceivably do to deserve not to be treated as innocent until proven guilty. And consider the concept of *inalienable* rights—rights can only be inalienable if they are disconnected from action-based desert, because they are by definition rights that we cannot lose through action. I think that personhood-based rights can all be seen as ramifications of a more basic right to be treated as ends and not mere means. Since many of the desert claims that we are pretheoretically inclined to accept are action-based, we would be in for substantial revisions of everyday notions of desert if we excised action-based desert claims from our thinking. But I think that enough of the rights that make up our pretheoretical ethical

understanding are personhood-based that free will deniers need not entirely lose grip on that understanding.

I think Kant's most helpful way of explaining how to treat others as ends is as follows. We refrain from coercing, deceiving, or otherwise manipulating other people into serving as means to our ends, by causing them to do things they would not rationally consent to do, and (at least some of the time) we actively share others' ends, by taking on their ends as our own. The "others as ends" principle doesn't require us to avoid treating each other as means in all cases, because it is sometimes rational to consent to serving as a means to another's end, for example, when he is a collaborator reciprocally serving as a means to one's own end. If I have the end of teaching philosophy, then the students I teach are among the means to my end. If they have the end of learning philosophy, then I am among the means to their end. Since our ends are complementary in this way, we can rationally consent to this interaction, and treat each other as ends as well as means.

3. Rational Consent Without Free Will

To my claim that rational consent doesn't imply free will, some may object that consent is only morally significant insofar as someone *acts* to give or withhold consent, in a way that makes them morally responsible for consenting or withholding consent. But this is a mistake—one which can be avoided if we properly distinguish between actual consent (consent given in the actual world) and hypothetical consent (counterfactual consent, consent that would have been given under different circumstances). When we properly understand the ethical significance of consent, it is clear that sometimes (though by no means always) the rationality of consent matters more than actuality of it. Sometimes what matters is hypothetical rational consent, and we quite

properly disregard actual consent. Since we can only be morally responsible for what we actually do, we cannot be morally responsible for what we would hypothetically consent to do. So it follows from the moral significance of hypothetical consent that moral responsibility is not essential to the moral significance of consent.

Let me explain in more detail. There are times when we disregard actual consent precisely because it is *ir*rational—in other words, we care about what the agent would consent to *if he thought rationally about things*, not what he actually consents to, because he is not reasoning properly for one reason or another. Imagine a situation in which someone has been kidnapped by a cult and brainwashed into accepting his own enslavement. We kidnap him back and force him to undergo deprogramming therapy. After brainwashing but prior to deprogramming, he actually consents to enslavement, and actually refuses consent to deprogramming. But in this case it is clearly appropriate to tell the cult members that his actual consent to enslavement does not license them to enslave him, and to force him to undergo deprogramming therapy even though he actually refuses consent. It is appropriate because in this case his actual consent is clearly irrational, and it is his hypothetical rational consent that determines how he should be treated. He would consent to deprogramming therapy if he thought rationally about things, and it is that consideration which carries the day. Respecting autonomy sometimes means treating people as they would rationally consent to be treated, rather than as they actually consent to be treated.

When we turn from extreme cases like the brainwashed slave to subtler cases in everyday life, there will of course be more debate about when it is appropriate to override actual consent in favor of hypothetical rational consent. Too often in our society, labelling someone's viewpoint as irrational amounts to nothing more than a prejudiced rejection of his perspective. A great deal

turns on how we understand what it would be rational to consent to do. A bit later, I will offer an account of rational consent that draws on Rawlsian original position deliberation. I think this is a good way to proceed, but I will not argue that here, and I will certainly not try to argue that it is the only way, or the best way, to proceed. I will merely claim that this seems to be an approach in the Kantian tradition which is worth exploring. There may be others that are worth exploring too.

I must also emphasize that I am in no way claiming that actual consent is not morally significant. In many cases, it is clearly wrong to override actual consent based on claims about hypothetical rational consent. These include cases where we are not sure what would be rational to do, or we think it important to tolerate disagreement about what is rational. Salient examples include voting, and choosing our friends and romantic partners. In pluralistic societies like those of the contemporary West, most of us agree that it is important to tolerate quite a lot of disagreement about what is rational. Kantians should hold that when in doubt, we ought to assume that an agent's actual consent is rational, and that claims about hypothetical rational consent have to meet a very high standard to justify overriding actual consent. But it is nonetheless important to see that we all draw the distinction between actual and hypothetical rational consent, and we all acknowledge that there are cases where hypothetical rational consent is what matters. This point severs the connection that may initially seem to exist between the ethical role played by consent, on the one hand, and the concept of moral responsibility, on the other. As mentioned above, this is because we can only be morally responsible for things that we actually do—not for things we would do if we acted rationally. Since moral responsibility plays no role in explaining the moral significance of hypothetical rational consent, moral responsibility is not essential to the moral significance of rational consent. This point gives us

no reason to get more comfortable with overriding actual consent. It is a metaethical point about the modal structure of actual and hypothetical consent, not a normative claim that actual consent matters less than we might previously have thought. A Kantian deontology for free will deniers should recognize the moral importance of both hypothetical and actual rational consent.

Objectors may now take a different tack—they may argue that the metaphysics of moral responsibility is still crucial to explaining the moral significance of hypothetical consent, because hypothetical consent can only be morally significant if agents *could have* rationally consented in cases where they did not actually do so, and this implies that a possible world in which they rationally consented was *accessible* to them. Objectors could point out that the metaphysics of accessible alternatives depends on the concept of free will, which is a prerequisite for moral responsibility. They might claim that this shows that the moral significance of hypothetical consent depends on free will, even if it does not directly depend on moral responsibility. This would pose just as much of a problem for free will deniers as a direct dependence on moral responsibility would pose.

The problem for this way of objecting is that if we look at how claims about hypothetical consent are typically formulated and applied, it is clear that they do not imply accessibility. That is, a claim about hypothetical consent typically states that an agent *would* have consented *if* he had thought rationally about things, but this does not imply that he *could* have thought rationally about it. And this is not simply a quirk of phrasing. Let us return to the case of the kidnapped and brainwashed cult slave. It is obviously morally significant that he would have refused consent to enslavement, and consented to deprogramming, if he had thought rationally about it, but it is just as obvious that *as things were,* in his brainwashed state, he *could not* have thought rationally about it. In other words, in his brainwashed state, his hypothetical rational consent

remains significant even though a possible world in which he thought rationally about things was not accessible to him. This demonstrates that neither moral responsibility nor free will are prerequisites for the moral significance of hypothetical consent.

Objectors may persist—they may claim that moral responsibility and free will are crucial to the significance of rational consent in the cases where actual consent matters, even if they are not in the cases where hypothetical consent matters. But given that these concepts have no role when it comes to explaining the significance of hypothetical consent, it seems unmotivated to insist on a role for them in an account of actual consent. It seems entirely adequate to say that respecting actual consent is the best way to respect autonomy in cases where we do not have extremely good grounds for favoring hypothetical consent. We can, after all, respect actual consent without supposing that actual consent is important because it is an expression of free will. Further, treating people as they would hypothetically consent to be treated, instead of as they actually consent to be treated, often involves coercion, which we must assume to be an invasion of autonomy unless we have extremely good grounds to believe otherwise.

Presumably all will agree that law enforcement is an area where coercion is necessary. Broadly speaking, Kantians should prefer the least invasive legal system consistent with the protection of human rights. But given human nature, whatever laws we do have will have to be enforced, and this seems to imply coercion. On the view I am advancing here, law enforcement sometimes requires us to disregard actual consent and instead treat people as they would rationally consent to be treated. The notion is that even when we sanction or punish law-breakers, we are obligated to treat them as they would rationally consent to be treated. On this view, it is legitimate to garnish the wages of a tax cheat, or imprison a violent criminal, so long

as it is the case that if these law-breakers had thought rationally about things, then they would have consented to this treatment.

This approach can be worked out in more detail with the help of social contract theory, which can be used to model rational consent. If it would be rational to choose to join in a social contract with a particular institutional structure, then we can view that institutional structure as one to which we would rationally consent. Kantian free will deniers should model rational consent based on personhood- but not action-based desert. The Rawlsian social contract seems especially well-suited to this approach, because original position deliberation has the effect of ‘‘filtering out’’ action-based desert claims. The original position is an idealized standpoint in which the members of a society choose the basic principles that will govern their society. It is formed by drawing what Rawls calls a "veil of ignorance" between the people who make up a society and all their particular characteristics. Original position deliberators must choose principles to govern society without knowing where they will end up within society. In the original position, one cannot know whether one is among the best or worst off, what one’s religion, ethnicity, or sex is, or what patterns of action one exhibits, for example, whether one is industrious or lazy. Rawls holds that the veil of ignorance ensures that the principles chosen in the original position will be just, since one will not be able to choose principles that make any of one's particular characteristics advantageous. Further, since people are self-interested and risk-averse in conditions of uncertainty, deliberators will worry most about ending up among the worst-off members of society, and they will therefore choose principles which make the circumstances of the worst-off members of society the best they can be.

When it comes to human rights and distributive justice, Kantian free will deniers can take on board Rawls’ view of original position deliberation in its entirety. Rawls thinks that original

positions deliberators will insist on equality of rights and basic liberties, as well as what he calls the "difference principle". This is the principle that economic inequalities are just if and only if they improve the conditions of the worst-off members of society. This implies that it is just for the industrious to derive advantages from their industry to the degree that it produces economic dynamics that raise the standard of living for the worst-off, for example, by creating incentives for hard work, and by making redistributive taxation possible. Once we raise the veil of ignorance, we can justly disregard a wealthy person's actual withholding of consent to taxation, because original position deliberation demonstrates that he would rationally consent to the taxation of the wealthy.

Rawls himself refuses to apply original position deliberation to penal justice. He allows us to assume that we will be able to follow the laws we choose in the original position. It is a matter of controversy why Rawls takes this approach.[16] But free will deniers should part ways with Rawls on this point, since applying original position deliberation to penal justice makes possible a non-retributive justification of punishment that avoids treating criminals as mere means. This means that we must assume ignorance about whether we are law-followers or law-breakers. It is more difficult to apply original position deliberation to punishment than to distributive justice, since there is no singular worst-off position here in the way there is with respect to distributive justice. That is, potential victims of crime and the people punished "compete" for the position of the worst-off, and also with one another, in the sense that if we assume that punishment deters, then modifying the penal system to improve things for one party tends to worsen things for the other party, and vice-versa. If the purpose of original position deliberation is to ensure fairness, then I think original position deliberators must assume that they have an equal chance of finding themselves among the punished, and among those protected by

the institution of punishment. Under these circumstances, it would be rational to choose to imprison violent criminals in benign prisons (prisons much less harmful than contemporary ones). But it would not be rational to choose an institution of punishment that included the death penalty, or torture, or imprisonment under harsh conditions. For people with normal social attachments, and a normal desire to be free from interference in the pursuit of their ends, prison would inevitably be harmful, even under benign conditions, and would therefore serve as a deterrent. But it would make sense to risk that harm if it significantly diminished our chances of being murdered or seriously injured by a violent criminal.

This justification of punishment countenances using criminals as means, since the rationale for accepting an institution of imprisonment is that it protects potential victims. However, since we would choose this approach to punishment in the original position, with the understanding that we might well turn out to be criminals when the veil of ignorance is lifted, we all satisfy the hypothetical rational consent requirement for punishment. So we can punish criminals without treating them as mere means, given their hypothetical rational consent. In other words, once we raise the veil of ignorance, we can justly disregard a violent criminal's actual withholding of consent to imprisonment, because original position deliberation demonstrates that he would rationally consent to punishment.

4. Conclusion: Reforming Kant's Own Theory

To conclude, I would like to apply the results of this inquiry to reforming Kant's own theory. As I mentioned toward the beginning, Kant himself is a retributivist. That is, he holds that punishment must be justified in terms of action-based desert. Kant bases his argument for

the legitimacy of the death penalty on the claim that murderers deserve to die. The core of Kant's argument goes like this:

> Punishment… can never be inflicted as a means to promote some other good for the criminal himself or for civil society. It must always be inflicted on him only *because he has committed a crime.* For a human being can never be treated merely as a means to the purposes of another…(1996, 6:331)

In this passage, Kant seems to accept the "two directions" picture I criticized earlier. He assumes that if we do not hold that criminals deserve punishment based on their actions, then the only other way to justify punishment is with a consequentialist appeal to the positive consequences of punishment for society. But as I have argued, there is another alternative—a personhood-based justification of punishment that turns on the claim that criminals would have rationally consented to punishment—and Kant's own theory has helped provide the resources to demonstrate this. In a nearby passage (1996, 6:335) Kant denies that, in the social contract, one has "consented to lose his life in case he murdered someone else", and he is certainly right here, at least insofar as we take a Rawlsian approach to social contract theory. Original position deliberation about punishment can justify benign imprisonment, but not seriously harmful measures. It may be that the only way to justify the death penalty is with action-based desert claims. But contemporary ethicists should see this as a point against the death penalty, rather than a point in favor of justifications that include action-based desert claims. Kant's own reasoning here may be driven by such a deep commitment to the death penalty that he cannot be content with any justification of punishment that will not support it. But this is not a part of Kant's theory that we should seek to preserve.

It seems quite clear that Kant's ethical theory as a whole becomes stronger if he gives up his retributivism and instead adopts a non-retributive justification of punishment.[17] As I mentioned at the outset, Kant holds that it is possible that we have free will, but we cannot know

that we do. The mere possibility of free will is simply not adequate to justify retribution, at least not seriously harmful retribution of the sort that is involved in imprisonment under harsh conditions, and certainly not the sort involved in the death penalty. We recognize that we must meet a very high burden of proof to justify harming people when we impose the "reasonable doubt" standard on arguments in criminal courts, and we ought to apply the same standard in assessing justifications of punishment.[18] Kant himself acknowledges a parallel standard when he argues that the priests of the inquisition could not have known that God wanted the inquisition's victims to die, and that no evidentiary standard short of knowledge could justify anyone in imposing so severe a penalty (*Religion Within the Limits of Reason Alone,* 6:185). I think it is fair to ask Kant to apply the same high standard to claims about free will when they appear in justifications of serious retributive harm.

Bibliography

Darwall, S. 1977. "Two Kinds of Respect." *Ethics* 88.1 (1977): 36-49.

Feldman, F. 1996. "Responsibility as a Condition for Desert." *Mind* 105.417 (1996): 165-68.

---. 1995. "Desert: Reconsideration of Some Received Wisdom." *Mind* 104.413 (1995): 63-77.

Fischer, J.M. 1995. *The Metaphysics of Free Will* (Aristotelian Society Series v. 14). Malden, MA: Blackwell Press.

Kant, I. 1996. *Metaphysics of Morals,* trans. and ed. Mary Gregor. New York: Cambridge, 1996.

---. 1998. *Religion Within the Boundaries of Mere Reason,* trans. and ed. Allen Wood and George di Giovanni. New York: Cambridge, 1998.

Mills, E. 2004. "Scheffler on Rawls, Justice, and Desert", *Law and Philosophy* 23, 261-72.

Moriarity, J. 2003. "Against the Asymmetry of Desert", *Noûs* 37:3, 518-36.

Pereboom, D. 2001. *Living Without Free Will.* New York: Cambridge University Press.

---. "Is Our Conception of Agent-Causation Coherent?" *Philosophical Topics* 32 (2006): 275-86.

---. "Free Will, Love, and Anger." *Ideas y Valores: Revista de Colombiana de Filosofía* 141, 2009, 5-25.

Rachels, J. 1978. "What People Deserve." In *Justice and Economic Distribution*, ed. J. Arthur and W.H. Shaw. Englewood Cliffs, NJ: Prentice-Hall.

Rawls, J. 1999. *A Theory of Justice, Revised Edition.* Cambridge, MA: Harvard Belknap.

Sadurski, W. 1985. *Giving Desert Its Due: Social Justice and Legal Theory*. Dordrecht: D. Reidel.

Scheffler, Samuel. (2000) "Justice and Desert in Liberal Theory", *California Law Review* 88, 991-1000.

Sommers, T. "The Objective Attitude", *The Philosophical Quarterly*, 57, 228 (2007), 321-41.

Smilansky, S. 1996. "Responsibility and Desert: Defending the Connection." *Mind* 105.417 (1996): 157-63.

---. 2000. *Free Will and Illusion.* New York: Oxford University Press.

---. 2005. "Free Will and Respect for Persons." *Midwest Studies in Philosophy* 29 (2005): 248-61.

---. 2006. "Control, Desert, and the Difference between Distributive and Retributive Justice." *Philosophical Studies* 131 (2006): 511-24.

Strawson, P.F. 1997. "Freedom and Resentment." In *Free Will*, ed. Pereboom, D. Hackett, 1997. 119-42.

Vilhauer, B. 2004a. "Can We Interpret Kant as a Compatibilist about Determinism and Moral Responsibility?" *The British Journal for the History of Philosophy*, Vol. 12, No. 4, 2004, 719-30.

---. 2004b. "Hard Determinism, Remorse, and Virtue Ethics', *Southern Journal of Philosophy*, 42, 4 (2004), 547-64

---. 2008. "Hard Determinism, Humeanism, and Virtue Ethics", *Southern Journal of Philosophy,* 46, 1 (2008), 121-44.

---. 2009a. "Free Will Skepticism and Personhood as a Desert Base", *Canadian Journal of Philosophy*, 2009, Vol. 39, no. 3, 489-511.

---. 2009b. "Free Will and Reasonable Doubt", *American Philosophical Quarterly*, 2009, Vol. 46., No. 2, 131-40.

---. 2010a. "The Scope of Responsibility in Kant's Theory of Free Will", *The British Journal for the History of Philosophy*, 2010, Vol. 18. No. 1, 45-71.

---. 2010b. "Persons, Punishment, and Free Will Skepticism", *Philosophical Studies* 'Online First', DOI 10.1007/s11098-011-9752-z, forthcoming in print.

---. 2012. "Taking Free Will Skepticism Seriously", *The Philosophical Quarterly*, doi: 10.1111/j.1467-9213.2012.00077.x, forthcoming in print.

Wolf, S. "The Importance of Free Will". *Mind*, New Series, Vol. 90, No. 359. (Jul., 1981), pp. 386-405.

Wood, A. 2010. "Punishment, Retribution, and the Coercive Enforcement of Right". In Lara Denis (ed.), *Kant's Metaphysics of Morals: A Critical Guide*. New York: Cambridge University Press. 111-29.

Sadurski, W. 1985. *Giving Desert Its Due: Social Justice and Legal Theory*. Dordrecht: D. Reidel.

Notes

[1] Kant claims that we cannot have "theoretical knowledge" that we have free will, but that we can nonetheless have "practical knowledge" that we have free will. How to interpret this claim about practical knowledge is a matter of longstanding controversy. It seems to have something to do with the idea that it is necessary to think of ourselves as having alternative possibilities of action when we deliberate about how to act. But without an account of the sort of necessity Kant has in mind, it is not clear what moral implications it has. If it is a merely psychological necessity, due (for example) to contingencies of evolution, then it would be hard to see how it could have any moral implications at all. Kant clearly means something stronger than mere psychological necessity, but he also seems to mean a sort of necessity that is consistent with acknowledging that we may not *really* be free after all. In my view, such practical necessity is of little help if we seek to justify holding each other morally responsible. So I think that Kant's approach must ultimately rest on the notion that the mere possibility that we have free will is enough to justify the assumption that we have free will whenever we reason about our practices.

2. Influential views in the contemporary literature which bear interesting relationships to Kant's view include those of Pereboom and Smilansky. I take Pereboom to hold that it is possible that we have free will, but that it is impossible to provide evidence that we do, and that the mere possibility that we have free will does not justify us in treating anyone as free in any context.

Smilansky holds that we can know we lack free will, but that free will is so important in ethics and practical reasoning that we are nonetheless justified in treating people as free in many contexts. See e.g. Pereboom 2001 and Smilansky 2000.

3. I discuss this in more detail in Vilhauer 2009b and 2012.

4. See Pereboom (2001, 151) for some helpful points on Kant and free will denial.

5. I discuss Kant's free will theory in Vilhauer 2004a and 2010a.

6. All references to Strawson are "Freedom and Resentment" (1997).

7. I make this argument in more detail in “Free Will Skepticism and Personhood as a Desert Base”.

8. A more detailed account of utilitarianism would have to distinguish between act- and rule-utilitarianism, but Strawson does not mention this distinction.

9. Parts of this section and the following section are adapted from Vilhauer 2009a and 2010b. I re-work ideas from those papers here in order to make them more general, explain their connection to Kant, and show their application to the people problem.

10. I discuss virtue ethics and attachment approaches for free will deniers in Vilhauer 2004b and 2008.

11. I discuss this further in Vilhauer 2012. Also see Pereboom 2001, 137.

12. All Kant citations are by Akademie pagination.

13. Philosophers who hold that all desert is action-based include Rachels (1978, p. 157) and Sadurski (1985, p. 131). Smilansky holds a related position, i.e., that giving up the belief that human beings are morally responsible for their actions implies giving up all our morally significant beliefs about desert (1996, 157-63).

14. Fred Feldman discusses this point, but not in the context of free will skepticism. (Feldman 1995a).

15. Darwall 1977, 38.

16. See Saul Smilansky (2006), Eugene Mills (2004), Jeffrey Moriarity (2003), and Samuel Scheffler (2000) for discussion of this issue.

17. Also see Wood 2010 for discussion of related issues.

18. Pereboom proposes applying the "reasonable doubt" standard in the free will debate (2001, 161). I also discuss this issue in Vilhauer 2009b.