An Asymmetrical Approach to Kant's Theory of Freedom

forthcoming in *The Idea of Freedom: New Essays on the Interpretation and Significance of Kant's Theory of Freedom*, Oxford University Press, ed. Dai Heide and Evan Tiffany.

Benjamin Vilhauer

Introduction

Asymmetry theories about free will and moral responsibility are a recent development in the long history of the free will debate. To my knowledge, Kant commentators have not yet explored the possibility of an asymmetrical reconstruction of Kant's theory of freedom, and that will be my goal here. By "free will", I mean the sort of control we would need to be morally responsible for our actions. Kant's term for it is "transcendental freedom", and he refers to the attribution of moral responsibility as "imputation". By "Kant's theory of freedom", I mean not only his theory of transcendental freedom and imputation, but also the various ways in which he draws on these ideas in his moral theory.

The key commitment of asymmetry theorists is that the standards that must be met to count agents as free and morally responsible are different in the context of the positive reactive attitudes and their attendant practices, such as praise and reward (what I will call "positive imputation" in Kant's context) than they are in the context of the negative reactive attitudes and practices including blame and punishment ("negative imputation" in Kant's context). The most-discussed asymmetry theory, developed by Susan Wolf and Dana Nelkin, posits an ontological asymmetry: people can be blameworthy only if they had alternative possibilities, but can be

praiseworthy even if they did not have alternative possibilities.[1] I have argued that even if we do not posit such an ontological asymmetry, we should acknowledge an epistemic and justificatory asymmetry—even if the ontological requirement agents must satisfy to be blameworthy is the same as the one they must satisfy to be praiseworthy, we must have better reasons for believing that the ontological requirement is satisfied to legitimately treat agents as morally responsible in the context of the negative reactive attitudes than we must have in the context of the positive reactive attitudes. This is because it is intuitive to think that people deserve the benefit of the doubt, and that there is a hazard of injustice in getting things wrong in connection with blame which does not exist in connection with praise, or at least does not exist in the same way or to the same degree.[2]

I will not propose a reconstruction of Kant's theory of freedom that posits an ontological asymmetry. I do not think this would be would be very useful, given Kant's consistency about the ontology of transcendental freedom. But given the dramatic shifts in Kant's epistemology of transcendental freedom and the inconsistencies they inflict upon commentators, I think a reconstruction which posits a justificatory asymmetry should be of interest. The reconstruction I want to propose is meant to be revisionary: I think that while Kant got a great deal right about the building blocks of his theory of freedom, he never fits them together in a stable way in his

Acknowledgements: Thanks to the organizers, participants, and audience at the 2014 *Kantian Freedom* conference at Simon Fraser University for helpful comments, especially Evan Tiffany .

[1] Susan Wolf first proposes this ontological asymmetry theory, in (1980) "Asymmetrical Freedom", *Journal of Philosophy* 77/3: 151-166, and Dana Nelkin develops it further in independent publications, including (2011) *Making Sense of Freedom and Responsibility*, Oxford University Press.

[2] Benjamin Vilhauer (2015) "Free will and the asymmetrical justifiability of holding morally responsible", *Philosophical Quarterly* 65: 772-789.

own texts. So this reconstruction does not describe the theory of freedom I think he actually held himself, but rather the one I think he ought to have held, and one which can be defended in contemporary discussions about free will.

Kant's understanding of the ontology of transcendental freedom seems consistent throughout the critical philosophy. It is a species of agent-causation which we might gloss as "noumenal libertarianism": it affords alternative possibilities of action despite phenomenal determinism, which Kant thinks we must posit in the contexts of both positive and negative imputation.[3] While there may in principle be conceptual space to explore a reconstruction on which transcendental freedom is only necessary in the context of negative imputation, I will not explore that possibility here, both because it would alter a consistent feature of Kant's metaphysics, and also because I think that Kant is largely correct in his doctrine of transcendental freedom. I think that we would need a power much like this to satisfy the control requirement for imputation whether the empirical world is deterministic, as Kant holds, or is

[3] I argue that Kant's rejection of Hume's inductivism allows Kant to endorse the possibility of single-instance deterministic laws, that is, causal laws which are instantiated just once, or which cover just one succession of events. This in turn accommodates the possibility of types of events which occur just once, which may plausibly be found among the events of empirical psychology (MNS 471). Determination of once-instantiated laws would allow agents *qua* noumena to control their own phenomenal actions in a way that affords them alternative possibilities, without entailing control of events outside what we typically understand as the scope of our causal responsibility, such as events prior to our births. Since causal laws establish the objective order of time, agents' determination of causal laws cannot be something that happens at a point in time, and is therefore 'timeless' and compatible with Kant's commitment to in-principle predictability of all actions. See e.g. Benjamin Vilhauer, "Can We Interpret Kant as a Compatibilist about Determinism and Moral Responsibility?" *British Journal for the History of Philosophy* 12, no. 4 (2004), 719-30, and "The Scope of Responsibility in Kant's Theory of Free Will," *British Journal for the History of Philosophy* 18, no. 1 (2010): 45-71. For an interpretation on which transcendental freedom does not essentially involve the ability to do otherwise, see Derk Pereboom (2006) Kant on Transcendental Freedom, *Philosophy and Phenomenological Research* 73 (3):537-567, p. 542-4.

indeterministic, since chance threatens imputation just as much as the deterministic subsumption of actions into a single series of nature.[4]

Kant's consistency about the ontology of free will stands in sharp contrast to his inconsistency about its epistemology. In the first *Critique* and the *Groundwork*, Kant holds that it is possible that we are transcendentally free, in a very thin sense of "possible"—that "nature at least does not conflict with causality through freedom"(A558/B586)—but that we cannot know that we are transcendentally free, and that for all we know, we are not transcendentally free. I will refer to this view as "possibilism". Possibilism differs from the view Kant advocates in the second *Critique* and afterwards, where he appeals to the "ought implies can" principle to argue from the claim that we know we ought to act in certain ways to the claim that know we can act in those ways, which he claims to imply that we know we are transcendentally free (5:30). Like many other commentators, I doubt the force of the second *Critique* argument in the context of the main themes of Kantian epistemology, built as they are around ignorance of noumena. Kant himself argues, in the second *Critique* and afterwards, that there is a practical epistemology which gives us the knowledge not available through theoretical epistemology that we are transcendentally free. I lack space to address this argument here, other than to mention that my skepticism about it is bound up with the concerns about the injustice of inadequately justified negative imputation explained in more detail below. That is, I think that the burden of proof that must be met to justly impose the serious sorts of retributive harm Kant favors, such as executing murderers and enslaving thieves (MM 6:333), is simply too heavy to be met by his practical

[4] Kant recognized that indeterminism was a threat to freedom as well as determinism: he says that excepting a "being whose existence is determined in time" "from the law of natural necessity" would be "tantamount to handing it over to blind chance" (2C5:95). This point is also noted in Pereboom (2006) p. 541.

epistemology. So, in what follows, I will for the most part assume that possibilism is the strongest view we can hold on the epistemology of transcendental freedom, and explore its implications.

The *Groundwork* and the second *Critique* set out contradictory positions on the implications of possibilism. In the *Groundwork*, Kant holds that possibilism is the only theoretical license we need to be justified in regarding and treating ourselves and others as free and morally responsible. In the second *Critique*, on the other hand, Kant's implicit view is that we must know we are transcendentally free to have the needed justification. To my knowledge, interpreters who wish to draw a consistent view out of Kant's texts have in the past chosen one of these views and rejected the other, or sought to interpret both in ways that eliminate the contradiction. My view is that these positions are indeed contradictory as Kant sets them out, but that the best reconstruction preserves both by distinguishing the aspects of our moral ideas and practices for which each is true. I think Kant is rightly sensitive to different demands of justice when he frames these different positions, but that he never combines them in a single account that is sensitive to all the demands of justice, and I will try to sketch how such an account might go here.

I will call this the "justificatory asymmetry reconstruction". It holds that in some but not all aspects of moral reasoning, we can justifiably appeal to claims about transcendental freedom even if we know only that it is possible that we are transcendentally free. Examples include the provision of "cans" to support "oughts", imputing merit, and drawing on the imputation of merit and the hope that we are transcendentally free to promote the development of rational nature in human beings. But in other instances of moral reasoning, such as retributively justifying serious

harm, we cannot appeal to claims about transcendentally free unless we know that we are transcendentally free.[5] This calls for a non-retributive reconstruction of Kant's ethics.

Kantian Ethics without "Oughts" ?

Before I explore the justificatory asymmetry reconstruction, I first want to consider what possibilism would imply if the standard Kant implicitly sets out in the second *Critique* is correct. What if not knowing that we are transcendentally free implies that we are not justified in making any "ought"-claims at all? If there were no "ought" claims there would be no categorical imperative, clearly, but there would also be no hypothetical imperatives. That is, by Kant's lights, we would be deprived of the building blocks of not just moral reasoning, but all practical reasoning.

We can look to a version of Hume's theory of agency for a fallback position for Kantians which may preserve worthwhile elements of Kantian moral rationalism. Hume can be read as offering a model of practical reasoning that dispenses with "oughts". It is widely agreed that Hume thinks reason does not tell us that we ought to pursue any ends, and that this on its own rules out categorical imperatives. But Hume may also hold that even instrumental rationality does not involve ought-claims, when properly understood, and in Kant's terms, this would mean that there are no hypothetical imperatives either.[6]

---

[5] My position is closely related to Derk Pereboom's position (Pereboom 2006) in that Pereboom also endorses possibilism and argues that it does not suffice for negative imputation. I take myself to differ with him in my view that possibilism is adequate to provide a role for transcendental freedom in grounding "oughts" and positive imputation.

[6] For recent discussions about whether Hume thought there were "oughts" of instrumental reason, see Christine Korsgaard, "Skepticism about Practical Reason", *The Journal of*

On such an ought-eliminativist view, there are no "oughts" at all in practical reasoning. Instrumental reasoning is just a combination of theoretical reasoning about what is the case, and desires. Successful instrumental reasoners are simply constituted in such a way that when they desire an end more than any competing end, and they theoretically reason that some means to that end is the most efficacious means over which they have power, then they acquire a desire for that means. There is no role for a further claim that they ought to take the means to the end.

However, ought-eliminativism does not imply Hume's broader view that reason has nothing to say about which ends are worth pursuing. It would only imply this if it were the case that the only way we can reason about which ends are worth pursuing is in the form of "ought"-claims. A number of philosophers have argued that this is not the case.[7] We can find materials for an ought-eliminativist account of moral reasons in Kant's own texts. That is, Kant claims that we can know what a perfectly good will would be like, and that there are no imperatives for a perfectly good will, since it necessarily wills morally. This point is relevant for understanding not just the will of God, but also rational wills more generally: the " 'ought' [of the categorical imperative] is strictly speaking a 'will' that holds for every rational being under the condition that reason in him is practical without hindrance" (G 4:449). For a perfectly good will, the moral law is a description of how it actually wills, not a claim about how it ought to will. So Kantian moral reasons do not necessarily come in the form of "oughts". It seems straightforward for ought-eliminativists to accommodate the idea of the perfectly good will, and to go on to give a

---

Philosophy 83, no.1 (1986): 5-25, and Elijah Millgram, "Was Hume a Humean?", *Hume Studies* 21 (1995): 75-93.

[7] Derk Pereboom argues this in *Living Without Free Will*, Cambridge University Press, 2001, p. 143-148. I make this case from a different perspective in "Hard Determinism, Humeanism, and Virtue Ethics" *Southern Journal of Philosophy*, Vol. 46, No. 1, 2008, pp. 121-144. .

general account of successful moral reasoning on its basis which has the same sort of structure as the ought-eliminativist account of instrumental reasoning. Ought-eliminativists who want to preserve the Kantian rational will could say that successful moral reasoners are simply constituted in such a way that, as a matter of fact, they will to act on universalizable maxims and to treat others as ends. They could thereby avoid claiming that successful moral reasoners will as they ought to will. It seems plausible that, despite our imperfections, we are all successful instrumental and moral reasoners some of the time, according to these definitions.

But even though an ought-eliminativist account can achieve some of Kant's goals in building a rationalist ethics, it would frustrate many others. Kant sees the claim that reason provides imperatives for our wills as essential for explaining how reason can direct imperfect wills like ours. "Ought"-eliminativism can accommodate the idea that a perfectly rational will necessarily wills morally, but it has no way to offer direction to an imperfectly rational agent who, as it happens, does *not* will in a perfectly rational way, and does not want to do so. Kant thinks imperatives play (at least) two roles in directing the will: they are *commands* that we give to ourselves as practical reasoners (4:413-4), and these self-commands play a role in generating *motivation* to conform to them.[8] If I know I cannot do something, then it is practically irrational to command myself to do it, or to be motivated to do it. A reconstruction that preserves imperatives and their roles in directing the will must find "cans" to support "oughts".

Possibilism and "Ought Implies Can"

[8] Kant may also hold that it is essential to imperative that if X ought to do A, then X's failure to do A is imputable to X. I do not think this follows, and if this is Kant's view, it is not an aspect of his view that I wish to preserve in this reconstruction. I think *X is blameworthy for failing to A* entails *X ought to have done A*, but I do not think the entailment holds in the other direction.

I think that possibilism provides strong enough "cans" to offer significant support for "oughts" in our practices of moral reasoning and action. This claim is likely to provoke suspicion from two different directions—first, from commentators who agree with Kant's second *Critique* view that we must have knowledge of transcendental freedom to have "cans" sufficient to support "oughts", and that a view on which "cans" depend in any way on our ignorance about how things really stand with transcendental freedom must be too weak to help in our practices—and second, from commentators who prefer reconstructions of Kant's theory of freedom even more revisionary than the one I am proposing, which seek to eschew the difficult metaphysics of transcendental freedom altogether and to make do with deflationary, compatibilist-style "oughts" and "cans" that would available even if we knew both that determinism was true, and also that transcendental freedom was false. I won't try to argue that possibilism can support "cans" and "oughts" as strongly as the knowledge that we are transcendentally free would support them—my more modest goal is just to show that the support it provides is practically significant. I will explain and argue for this claim by contrasting possibilism with one such compatibilist, deflationary strategy, which I will call the "merely epistemic view". The "cans" and "oughts" provided by the merely epistemic view also depend on ignorance, but in a different way: on this view, what makes it consistent for us to believe that we can act in more than one way at any given time in the future is merely the fact that we typically cannot predict how we will actually act at that time.[9] I will argue that possibilism provides better support for "oughts" and motivation

[9] Hilary Bok emphasizes this merely epistemic sense of possibility in the broadly Kantian account of agency she presents in *Freedom and Responsibility*, Princeton University Press, 1998, pp. 109-122. I also discuss it in "Taking Free Will Skepticism Seriously", *The Philosophical Quarterly*, October 2012, Vol. 62, No. 249, pp. 833–852.

in our practices than the merely epistemic view provides, because possibilism blocks retrospective falsification of "ought"-claims which the merely epistemic view cannot block.

There is a broad consensus in favor of the view that deliberating about what to do at some point in the future is impossible without the belief that there are multiple things that I can do at that point. This is part of what Kant has in mind in his *Groundwork* claim that "to every rational being having a will we must necessarily lend the idea of freedom also, under which alone he acts"(G:448). But it is not obvious that the beliefs about "cans" necessary for deliberation commit us to any beliefs about the actuality, or even the possibility, of ontological alternatives of the sort that would be underwritten by transcendental freedom. What is necessary for deliberation seems to be that I can act in multiple ways at a given future time *to the best of my knowledge,* and this merely requires my ignorance about what I will actually do—an ignorance which would almost always beset me even if I knew determinism was true and transcendental freedom was false, given my limited ability to predict my actions. In other words, if I interpret "can" in this merely epistemic way, then even if I know that determinism is true and know (*contra* possibilism) that I am not transcendentally free, it is true that I can act in more than one way at any given time *t* in the future, so long as I do not know what I will actually do at *t.* Knowing the truth of determinism and the falsity of transcendental freedom implies that I have no ontological alternatives in the actual future, or said differently, that there is only one way in which it is ontologically possible for me to act in at any one point in the actual future.[10] But the

[10] Hume's compatibilism holds that we have ontological alternative possibilities even if determinism is true and we lack transcendental freedom—he argues in Section 8, part 1 of *An Enquiry Concerning Human Understanding* that even if determinism is true, "if we choose to remain at rest, we may; if we choose to move, we also may." But Hume implicitly acknowledges that such alternatives are dependent upon something having been different about the past or the laws of nature (which on his view are not under our control) that would have caused us to act differently. They are in this sense not alternatives in the actual future, and the

probability seems vanishingly small that we will ever be able to predict our actions in any detail, given the countless variables that would need measuring for prediction even given perfect knowledge of the laws of nature. However, so long as I do not know what that one particular way will be, there are various merely epistemically possible ways in which I can act.

This merely epistemic sense of "can" is sufficient to give us the logical space we need to accommodate the "commands of reason" essential to "ought"-claims. Even an agent who believed herself to be deterministic and transcendentally unfree could contemplate merely epistemically open paths X and Y, deliberate about which one she ought to take, come to a conclusion, and command herself to conform to it. If she is a psychologically normal agent, this command will have some motivational efficacy.

Both possibilism and the merely epistemic view hold that "oughts" and "cans" are conditioned by our ignorance of how things actually are, but in different ways. Both possibilism and the merely epistemic view hold that "oughts" and "cans" are compatible with (1) the truth of determinism, (2) the falsity of transcendental freedom, and (3) knowledge of the truth of determinism. But only the merely epistemic view is compatible with (4) knowledge of the falsity of transcendental freedom. On the merely epistemic view, it is the mere fact of my ignorance about how I will actually act (due to the limits of prediction) which creates the alternatives supporting "oughts". On possibilism, my ignorance derives not only from the limits of prediction (which obtains in any plausible theory of freedom) but also from noumenal ignorance: since it is possible that I am transcendentally free, it is possible that I have ontological

---

problem for advocates of this compatibilist "conditional analysis" of "can" is to explain how such alternatives are accessible to us in a way that would justify imputation. For a reconstruction of Kant's theory of freedom along these lines, see Hud Hudson, *Kant 's Compatibilism*, Cornell University Press, Ithaca: 1994, pp. 92-98.

alternatives. In this way, possibilism offers a richer sense of an open future, and a past which is such that it is ontologically possible that I could have acted otherwise than I in fact acted.

But does the difference in the ways possibilism and the merely epistemic view are conditioned by ignorance make a difference for our practices? If the availability of the merely epistemic view means that "oughts" would be secure in our practices even if we knew we lacked transcendental freedom, then perhaps there is no reason aside from faithfulness to Kant's texts to find a role for transcendental freedom in the preservation of imperatives, and to put up with the complicated metaphysics it brings with it.

I think there are some cases in which the merely epistemic view is adequate in practice. Suppose Larry knows that determinism is true, and that if he turns a key, he frees an unjustly accused prisoner who is good by Kant's (or any other reasonable ethicist's) standards and is about to be painfully executed. Nearly all Larry's inclinations dispose him favorably toward turning the key, with the exception of a nagging desire to keep reading the book he will have to set down if he turns the key, which makes him hesitate. He reflects for a moment, realizes that turning the key would be the just thing to do, and concludes that he ought to do it, and this gives him the additional nudge of motivation that he needs to put down the book and turn the key. Is there any practical significance in a case like this whether we take the "ought" and its "can" to be merely epistemic, on the one hand, or possibilist, on the other? Assuming that Larry had no way to predict that he would not turn the key, he had the logical space necessary to issue a command of reason to himself even if he believed himself to be transcendentally unfree, and his self-command gave him the extra bit of motivation he needed to act. So it seems to me that the merely epistemic view arguably provides us with everything we need for practical purposes in such a case.

In other cases, however, possibilism offers resources that the merely epistemic account of "can" lacks. An example can be found in famous second *Critique* case in which Kant imagines asking someone

> whether, if his prince demanded, on pain of . . . immediate execution, that he give false testimony against an honorable man whom the prince would like to destroy . . . he would consider it possible to overcome his love of life . . . He would perhaps not venture to assert whether he would do it or not, but he must admit without hesitation that it would be possible for him. He judges, therefore, that he can do something because he is aware that he ought to do it and cognizes freedom within him . . . (2C 5:30)

This is similar to the key case, in that Kant wants it to be clear that the normative question of whether one ought to resist giving false testimony is to be answered in the affirmative just as clearly as the question about whether to turn the key was. But this agent's motivational struggle is more serious than the trivial one that made Larry hesitate. Now, Kant's own point about this case is just that if I find myself in such a situation, then despite the motivational struggle, it is nonetheless clear that I ought to refuse to give false testimony, and this implies that I can refuse (at least according to his practical epistemology of transcendental freedom in the second *Critique*). I want to adapt and expand on this case to argue that it points out a practically significant role for possibilism in supporting our motivation to conform to commands of reason, because of the way it blocks retrospective falsification of "ought"-claims.

Kant's general account of motivational struggle involves a sort of force with which respect for the moral law suppresses ("humiliates") the inclinations of self-conceit (see e.g. 5:72-6). It seems reasonable that the more confident I am in my belief that I ought to act in some way, the more effective this feeling will be in suppressing my inclinations of self-conceit. Part of that confidence derives from confidence in my understanding of my obligations: do I really owe it to the person against whom I am asked to give false witness to refuse to act in this way? Kant

thinks it clear that we should answer "yes". Another part derives from the "ought implies can" principle: can I do the thing that I am commanding myself to do? This question highlights a limitation in the merely epistemic account of "can". If it is merely my ignorance of what will actually happen at time t that underwrites the alternatives at time t necessary to sustain deliberation and "can"-claims, then the corresponding "ought"-claims will often be falsified in retrospect, after I know what I actually do at time t. Suppose that the prince has come to me seeking a false witness because I have always agreed to be a false witness before, but never without a motivational struggle between my belief that I ought to refuse and my inclinations. On the merely epistemic account of "can", my previous failures to refuse imply that I was wrong, in those past cases, to believe that I ought to refuse. That is, if what supports the "can" paired with "ought" in "I ought not bear false witness" is only my ignorance about whether I will bear false witness, then once I learn that I *do* in fact bear false witness, the "can" disappears, and with it, the "ought". My past failure to be sufficiently motivated to act as I believed I ought to act retrospectively falsifies my past claims that I ought to act in that way. This is unattractive both semantically and morally. The fact of failing to act as I believe I ought to act does not seem like the right sort of fact to falsify the claim that I ought to act in that way. It is also motivationally problematic—that is, human nature being what it is, if I know I was always wrong in the past to believe that I ought not give false witness, then it is natural for me to be less confident in the belief that I ought not give false witness now, and if that reduced confidence reduces the force of respect for law on my self-conceit, my motivation to act morally will be eroded.

If, however, we adopt possibilism, this restrospective falsification is avoided. That is, if the sense of "can" at work incorporates the possibility that I am transcendentally free, and that I therefore may have been able to do otherwise at time t, then my knowledge that I did not do x at t

need not falsify the claim that I ought to have done x at t, because I can appeal to the possibility of a metaphysical mechanism that would have enabled me to act differently. So failure to act as I believe I ought to act would not erode motivation in the same way.[11]

It would of course be possible to preserve the merely epistemic account of "can" and avoid the retrospective falsification problem if we carefully time-indexed our ought-claims. But I do not think that such time-indexing would do much to help with the motivational erosion problem for agents who were self-conscious about the semantic device they were employing, since time-indexing cannot really do anything to change the fact that, on the epistemic account, claims about what we ought to do are grounded only on our ignorance about what we actually do. I do not think there is any way to avoid motivational erosion once we attend to that fact. This shows that possibilism is superior to the merely epistemic view in the support it offers for "oughts" in our practices.

Some will no doubt object that what we really need to avoid motivational erosion is knowledge that we are transcendentally free. There is no way to rule out the possibility that agents who take themselves to know that they have transcendental freedom will have stronger moral motivation than agents who merely believe that transcendental freedom is possible. This is in significant part an empirical question, and I cannot address it here. But it seems clear that possibilism offers practically significant advantages over reconstructions that dispense with transcendental freedom altogether, and that is all that I hope to establish here.

Possibilism and Imputing Blame

[11] I discuss "oughts" and motivational erosion in more detail in Vilhauer 2008.

What epistemic standard would we have to meet to legitimately appeal to claims about transcendental freedom to retributively justify executing someone accused of murder? Here, there is the obvious issue of assessing the accusation—did the accused actually commit the murder?—and there is a strong moral intuition that we must apply a very high justificatory standard in assessing such claims when serious punishment is at issue, which is manifested in the "beyond reasonable doubt" standard in criminal trials. Derk Pereboom has argued that the same high justificatory standard should be applied to claims about free will because of the way they are applied in this context.[12] Kant is also alert to this problem, since he claims in the first *Critique* that the unknowability of the intelligible character means that we can never impute actions with "complete justice" (A551/B579). Further, in the Vigilantius *Lectures on Ethics* notes, we find Kant emphasizing a standard for the imputation of blame ("*imputatio demeriti*") which may be even higher than absence of reasonable doubt. Imputation "presupposes that an action can be considered as a *factum*", which requires it to be "the effect of a *causa libera qua talis*", that is, "chosen with free will"(V27:561). A judgement "whether somebody is declared to be *auctor facti*, must always, if it involves an *imputatio demeriti*, be based on certainty. It is otherwise *invalidum*, the accused suffers injustice and is injured" (V27:564). Later he describes this standard in more detail as the "utmost moral and logical certainty" (V27:566).

By the time of the Vigilantius notes Kant has long since claimed in the second *Critique* to have discovered an argument that proves the reality of transcendental freedom in the "absolute

---

[12] Pereboom introduces the idea that the reasonable doubt standard is relevant for the metaphysics of moral responsibility into the contemporary free will literature. See *Living Without Free Will*. Cambridge University Press, 2001, p. 161. I also discuss this idea in "Free Will and Reasonable Doubt," *American Philosophical Quarterly* 46, no. 2 (2009): 131-40, and the 2015 paper cited above.

sense in which speculative reason needed it"(CPrR 3), so he would by then have claimed there to be no metaphysical obstacle to meeting this standard. In fact, already by the time of the *Groundwork*, when Kant continues to maintain possibilism, he appears to be moving away from his first *Critique* worry about justice in imputation. In Kant's *Groundwork* exposition of the moral implications of possibilism he claims that since every rational being must act under the idea of freedom, "all laws that are inseparably bound up with freedom hold for him just as if his will had been validly pronounced free also in itself and in theoretical philosophy"(G 4:448). Since he states no exception for the laws connected with imputation, it is natural to assume that he means to include them. But it seems clear that this does not follow—even if some beliefs about my own freedom or control are necessary when I deliberate about imputation, it does not follow that I must have those same beliefs with regard to the object of my deliberation. Whether I am deliberating about imputing actions to myself or to another, I represent the agent as the object of my deliberation, and I have the conceptual space to refrain from regarding the agent as transcendentally free, and "refer" the action "only to the empirical character", that is, only to the deterministic phenomenon, which Kant himself claims to be required by justice in the A551/B579 passage just mentioned. So the practical perspective on the question of transcendental freedom that Kant advocates in the *Groundwork* can easily accommodate skepticism about imputing blame when suitably reconstructed.[13] Even if there is something correct about Kant's *Doctrine of Right* claim that "A person is a subject whose actions can be imputed to him" (MM 6:223), it is not clear why we would have to impute *every* action to a person – if there are reasons of justice to refrain from imputing blame in some cases, then the

[13] Evan Tiffany also makes this point in (2013) "Choosing Freedom: Basic Desert and the Standpoint of Blame", *Philosophical Explorations* 16 (2):1-17.

practical perspective allows us the conceptual space to refrain from these imputations, and instead focus on the imputation of meritorious actions.

If the only way to meet the standard for serious retributive harm is to know that we have transcendental freedom, then the standard Kant sets out in the Second *Critique* is the right one in this context. Assuming the truth of possibilism, it cannot be met. This means that Kant's preferred responses to criminals, such as the execution of murderers and the enslavement of thieves (MM 6:333), cannot be justified, and that imprisonment under torturous conditions like those that prevail in contemporary prisons cannot be justified.

Retribution involving bodily injury and coercion is not the only sort of seriously harmful retribution relevant for Kant's ethics and contemporary practice: we use blame and guilt to inflict emotional retribution upon others, and also self-reactively upon ourselves through the pain of conscience. I take this idea to be a part of what passes for everyday moral common sense in society, but it is also worked out in a descriptive model of conscience in Freud, and arguably as a normative model in Kant.[14] At 6:394, Kant says that to experience "*pain*...from the pangs of conscience" is to "deservedly suffer…inner reproach". He also argues that just courts apply the principle of retribution, and gives a detailed account of conscience as an inner court. Emotional retribution can amount to harm just as serious as the harm of bodily torture; a sign of this is that both can prompt the desire to commit suicide. Kant suggests at 6:485 that his moral theory does not encourage a debilitatingly painful conscience, but this remark is in tension with others. For example, at 6:439, he claims that when conscience makes a negative judgment, it "pronounces the sentence of…misery, as the moral results of the deed". To hold that we deserve to suffer misery as the moral result of evil deeds is arguably to retributively justify serious emotional

[14] See e.g. S. Freud, *Civilization and its Discontents* (New York: Norton, 1989), pp 83-96.

harm. Since Kant's own accounts of punishment and conscience cannot meet the epistemic standard we must apply here, both require repair, to which I will return below.

Possibilism and Imputing Merit

The justificatory asymmetry reconstruction holds that possibilism is adequate to support the role of transcendental freedom in positive imputation in at least some cases. When we examine Kant's own approach to positive imputation, we find that Kant remains a kind of skeptic about positive imputation throughout the critical philosophy, even after he has rejected possibilism in favor of knowledge of transcendental freedom, and that he nonetheless gives positive imputation a central role in his moral theory. For this reason, I think Kant's own view of positive imputation incorporates a low enough epistemic standard to allow the justificatory asymmetry reconstruction to leave it largely intact.

Kant's main term for what we positively impute is "merit". In the *Doctrine of Right*, Kant says that "If someone does *more* in the way of duty than he can be constrained by law to do, what he does is *meritorious*." We act meritoriously in fulfilling duties of virtue—an example is when "at considerable self-sacrifice I rescue a complete stranger from great distress"(6:228) without, say, hoping for a reward. But we can also act meritoriously in fulfilling mere duties of right:

> Although there is nothing meritorious in the conformity of one's actions with right (in being an honest human being), the conformity with right of one's maxims of such actions, as duties, that is **respect** for right, is *meritorious*... since another can indeed by his right require of me actions in accordance with the law, but not that the law be also my incentive to such actions. (6:390-1)

Kant emphasizes in the second *Critique* that we are not morally worthy if we act in accordance with the law *out of a desire to be meritorious*, but we are nonetheless meritorious if we act in accordance with the law from duty (5:85).

In the first *Critique*, Kant endorses a kind of skepticism about merit which is tightly bound up with skepticism about transcendental freedom: his concern that the unknowability of the intelligible character implies that we can never impute actions with "complete justice" extends to both "merit and guilt" (A551/B579). In the *Groundwork*, where Kant argues that "all laws inseparably bound up with freedom hold" for us (G 4:448, quoted above), we might have expected merit skepticism to dissipate, but it does not:

> It is indeed sometimes the case that with the keenest self-examination we find nothing besides the moral ground of duty that could have been powerful enough to move us to this or that good action and to so great a sacrifice; but from this it cannot be inferred with certainty that no covert impulse of self-love…was not actually the real determining cause of the will; for we like to flatter ourselves by falsely attributing to ourselves a nobler motive…(G 4:407)

In this passage merit skepticism seems to derive from the limits of our introspective insight into ourselves, rather than skepticism about transcendental freedom. Since Kant still maintains possibilism in the *Groundwork*, it is natural to wonder whether doubts about the reality of transcendental freedom play an implicit role, but this seems to be ruled out by Kant's expression of the same thought in the *Doctrine of Virtue*:

> a human being cannot see into the depths of his own heart so as to be quite certain, in even a *single* action, of the purity of his moral intention…even when he has no doubt about the legality of the action. Very often he mistakes his own weakness, which counsels him against the venture of a misdeed, for virtue. (6:392-3)

This passage appears long after his rejection of possibilism, and this makes it clear that he thinks there are limits to introspection which ground merit skepticism and are independent of skepticism about transcendental freedom. Kant clearly does not regard this skepticism as a

problem for drawing on the idea of merit in our practices. (He does not disregard it altogether, of course—for example, it is a reason to avoid overestimating our worth and denying our continuing need for self-improvement.) It seems clear that irrespective of how things stand with transcendental freedom, Kant thinks there are reasons to hold the imputation of merit to a lower epistemic standard than the imputation of blame. I think this is because of his implicit recognition that the hazard of injustice present in holding imputations of blame to too low a standard is not present in the same way with imputing merit. If people deserve the benefit of the doubt, then we do not have to be certain that they are meritorious to legitimately treat them as meritorious. Because of this, it seems reasonable to suppose that his mature account of the imputation of merit should be able to tolerate possibilism about transcendental freedom alongside the merit skepticism based on limits to introspection which he explicitly acknowledges. If Kant's account of imputing merit can tolerate possibilism, then the justificatory asymmetry reconstruction need not disturb it.

My claim is not that there is *no* hazard of injustice in the context of imputing merit. The degree to which merit can be imputed "has to be assessed by the magnitude of the obstacles that had to be overcome"(6:228), and we can never know "how many people who have lived long and guiltless lives may not be merely *fortunate* in having escaped so many temptations"(6:393). We must therefore be alert to the risk of imputing merit to some in a way that unjustly excludes others. But this can be avoided by (for example) singling out particular individuals for the imputation of merit privately, or striving to be egalitarian by imputing merit to everyone who has tried to act in a morally worthy way. (It seems plausible to assume that everyone sometimes tries to act in a morally worthy way except psychopaths who lack the capacity to care about merit, and

it is hard to see how it could be unjust to exclude such individuals from the imputation of merit.)[15]

If possibilism is adequate to support the role of transcendental freedom in imputing merit, then it seems reasonable to think that the justificatory asymmetry reconstruction can achieve a variety of Kant's other goals in his theory of freedom. As mentioned earlier, Kant makes imputability an essential aspect of personhood in the *Doctrine of Virtue*, and his moral writings are full of remarks that connect regarding ourselves as transcendentally free with a special sense of dignity. More broadly, Kant thinks that regarding ourselves as transcendentally free contributes to the development of rational nature in us in a wide variety of ways that I cannot adequately catalog here, which include disclosing a sublime moral vocation that is not confined by the limits of the empirical world, and supporting the beliefs that we can resist the impulses of the senses and that we are autonomous law-givers. Cases like the key-turning case above raise questions about whether it is really necessary to regard ourselves as transcendentally free to see ourselves in these other ways. But it does not strike me as implausible to speculate that it may be at least an empirical fact about human beings that regarding ourselves as having the sort of freedom that transcendental freedom would confer can help us to see ourselves in these other ways, and presumably we can use all the help we can get. Why should it be necessary to draw on both positive and negative imputation to take advantage of this? Suppose we have answered all the empirical questions about the contexts in which, drawing solely on positive imputation, we can promote the development of rational nature in human beings by regarding ourselves as transcendentally free. Would not possibilism be a sufficient ground for regarding ourselves in this way, in these contexts?

[15] I discuss a parallel point in Vilhauer 2015.

Some may object that it would be irrational to regard ourselves as free unless we believed we were free, and the possibility that we are free does not justify the belief that we are free. This objection is relevant for the justificatory asymmetry reconstruction's account of imputing praise as well as the account of the broader benefits of regarding ourselves as transcendentally free that I have just discussed. But it oversimplifies the range of cognitive stances we can take toward representations of ourselves. I might have a reason to coach a basketball player to *imagine* herself as a bird if it helped her jump higher, and it would presumably not be necessary or helpful in this case to claim that it was possible that she actually was a bird. But the cognitive stance toward transcendental freedom that best fits the justificatory asymmetry reconstruction's account of positive imputation is the *hope* that we are transcendentally free. Rational agents can pretend or imagine that things they know to be false are true, but they cannot hope that things they know to be false are true. The possibility of transcendental freedom would seem to be an adequate basis to strive to instill the hope that we are actually transcendentally free—it allows us to impute merit to agents not in the trivial spirit of encouraging them to pretend to themselves that they are meritorious, as a sort of internal emotional play-acting, but in a way that allows them to rationally hope that they are.

Others may object that this approach to regarding-as-free is needlessly cautious. It may seem that even if I knew that transcendental freedom was impossible, I could still legitimately deceive people into believing that they are transcendentally free if I knew that it would contribute to the development of rational nature in them. A view much like this is developed in intriguing detail by Saul Smilansky.[16] But for Kantians, such deception should be recognized as

[16] Smilansky, Saul (2000). *Free Will and Illusion*. Oxford University Press.

using the people deceived as mere means, even if it would benefit them—it would be a kind of epistemic paternalism.

It may be that people naturally default to the belief in libertarian free will, so that philosophers would not need to actively deceive people to maintain broad social acceptance of the belief in free will. Can Kantians simply acquiesce in people's maintenance of the false belief that they know they have libertarian free will? Perhaps there are some cases where acquiescence would be permissible, though I am not sure. But it is unsatisfactory as a general strategy, because people often appeal to unreflective libertarianism to defend retributivism, and philosophers ought to object to bad arguments used to justify harm.

Kantian Ethics Without Retribution

Serious retributive harm of both bodily and emotional kinds demands the second *Critique* epistemic standard of knowledge of free will, and given that it cannot be met, Kant cannot justify either of these. Since Kant's own ethics depends on them, repair is required. Optimistic anarchists might hope that rational nature is strong enough in us that society can hobble toward a condition of right even without the crutches of punishment and conscience. While Kant may be too pessimistic in thinking we need to solve "the problem of organizing a nation…for a people comprised of devils"(8:366), I think it is imprudently utopian to do ethics without accounts of punishment and conscience. So I think the best repair is to seek non-retributive Kantian accounts of punishment and conscience.

Kant's texts offer resources here which have not yet been fully explored. I advocate a non-retributive Kantian approach to punishment which derives from our perfect duty to avoid

treating others as mere means, and a non-retributive account of conscience based on a kind of remorse which derives from our imperfect duty to take others' permissible ends as our own.

Kant holds that the only alternative to retributively justifying punishment is to appeal to the good consequences of punishment for society, such as deterrence, which treats criminals as mere means to the end of a better-functioning society. But Kant is mistaken here. I advocate a non-retributive "ideal abolitionist" account of punishment inspired by Kant's first *Critique* skepticism about imputing blame, and his remark that in "a perfect state no punishments whatsoever would be required", and that we must strive "to bring the legislative constitution of human beings ever nearer to [this] possible greatest perfection"(A316-7/B373-4).

The idea is to select the principles of punishment in a version of Rawls' original position, and thereby draw on the idea of rational consent to punishment rather than retributive desert.[17] Suppose that we had to choose institutions of punishment behind the veil of ignorance, assuming that we had an equal chance of finding ourselves among the punished and among the unpunished. Our first priority would be to make immediate progress with all means at our disposal toward a society that dispensed with institutions of punishment and emphasized non-coercive preventative strategies to diminish incentives to commit crime, like better access to public services, jobs, education, and voluntary therapies for those most at risk of offending. But while we work toward the ideal of abolishing punishment, it would be would be rational to maintain a scheme of reciprocal coercion which is currently our best hope for approximating a

---

[17] I discuss this approach in more detail in Benjamin Vilhauer, "Punishment, Persons, and Free Will Skepticism," *Philosophical Studies* 62, no. 2 (Jan. 2013), 143-63, and in "Kant's Mature Theory of Punishment, and a First *Critique* Ideal Abolitionist Alternative", forthcoming in the *Palgrave Kant Handbook*, ed. Matthew Altman. Sharon Dolovich advocates a similar approach, but not in the context of Kant interpretation or free will skepticism ("Legitimate Punishment in Liberal Democracy," *Buffalo Law Review* 7, no. 2, 2004, pp. 314-29.)

condition of right, in order to avoid falling back into the state of nature (MM 6:221). For example, it would be rational to choose to imprison violent criminals even knowing we might be among them, under the right conditions of imprisonment. I think that to pass the rational consent test, these would have to be conditions that offered criminals more choiceworthy lives than the state of nature. It may sound absurd to suppose that any conditions of imprisonment could pass this test, but from the Kantian perspective worthwhile freedom is not the "wild, lawless freedom" of the state of nature, but instead lawful freedom (MM 6:315). From the perspective of right, worthwhile freedom is the freedom of acting without violating the limits of others' rightful freedom. Placing too much weight on this point would be totalitarian. However, applied with humane caution, I think it is helpful. It seems plausible that a *radically* reformed institution of imprisonment which provided meaningful opportunities for social interaction, work, education, voluntary therapy, and continual parole review could afford violent criminals more choiceworthy lives than they would have in the state of nature. This approach to punishment can by no means justify the harsh measures which Kant himself prefers, but I think it hews closer to the dominant impulses of Kantian ethics. That is, it follows Rawls in drawing on what I think is the core conception of the moral person in Kant's ethics, as the rationally autonomous legislator in the kingdom of ends.

The non-retributive Kantian account of remorse I advocate is loosely based on a rationalist analog of Hume's sympathy-based moral psychology. The main idea is that there is a kind of remorse in which the wrongdoer suffers in sympathy with the pain he has caused the victim of his wrongdoing. It is to be understood as having family resemblances to other kinds of sympathetic suffering, for example, the sympathetic suffering we feel for our friends and loved ones when they suffer. The key point is that this kind of remorse is valuable not because it is

deserved, but because it shares in whatever makes these other kinds of sympathetic suffering valuable. It is clear that the value of sympathetic suffering does not derive from its being deserved. It would be absurd to think that by befriending or loving someone, I have gotten myself into a position such that I deserve to suffer when she does. Its value has instead to do with the way it partially constitutes a valuable relationship. It is not good merely because of its good consequences. I don't deny that it is likely to have good consequences, since sympathetic suffering disposes us to relieve the suffering of the people with whom we sympathize. It is natural to think that sympathetic remorse would not only prompt wrongdoers to try to restore their victims' well-being, but would also sensitize wrongdoers to the effects of their wrongs and prompt them to act better in the future. However, for Kantians and non-consequentialists more broadly, it is important to be able to understand sympathetic suffering as intrinsically valuable. Its intrinsic value can be explained in terms of care ethics, or, for Kantians, in terms of the emotional conditions for the possibility of taking others' permissible ends as our own.[18]

Kant's ethics can at first blush seem thoroughly hostile to any moral role for sympathetic engagement with others. But further reflection shows something different.[19] For Kant, having a permissible end implies having a particular determination of feelings which orients one conatively and emotionally toward that end. So, to fulfill our imperfect duty to take others' permissible ends as our own, we must cultivate determinations of our feelings that correspond to determinations of others' feelings. In other words, we cannot take others' permissible ends as our

[18] I develop this account of remorse in the context of care ethics in "Hard Determinism, Remorse, and Virtue Ethics" Southern Journal of Philosophy, Vol. 42, No. 4, 2004, pp. 547-564.

[19] Two recent discussions of Kant on sympathy which I take to lend support to this approach are Melissa Seymour Fahmy, "Active Sympathetic Participation: Reconsidering Kant's Duty of Sympathy", *Kantian Review* 14 (1):31-52 (2009), and Allen Wood, *Kantian Ethics*, Cambridge: New York, 2008, pp. 24-43.

own without sharing their feelings in important ways. In this way sympathy can take on a role in Kantian ethics which parallels the role sympathy plays in Hume's ethics, though in Kantian philosophy the emotions involved are not metaethically foundational in the way they are in Hume's ethics—instead they are the emotional manifestations of rational relations among finite rational beings. I read Kant as beginning to appreciate this in his later ethics, for example, in his discussion of active sympathy in MM 6:456, "Sympathetic Feeling is Generally a Duty". I take this line of thought to show that we have moral reasons in some cases to be emotionally pained by the things that emotionally pain the people whose permissible ends we take as our own. One implication of our finitude is that we lack the capacity to take *all* others' permissible ends as our own. Partly for this reason, Kant accepts that "one human being is closer to me than another" (MM 6:451), for example, my parents, children, and friends. How this point fits with universalizability is a complex question which I cannot address here, but it is clear that we have special reasons to take as our own the permissible ends of people close to us, and consequently to be pained by what pains them. Kantians can hold that people we have wronged are similarly morally close to us—we ought to share in the suffering of people we have wronged in a way that parallels the way we ought to suffer in sympathy with our loved ones when they suffer, because it is the determination of feeling which expresses the rational relation of taking their ends as our own. In this way, I think Kantians can have an appropriately rationalistic, non-retributive account of conscience.

To conclude, I would like to have a try at a metaphor that both Derk Pereboom and Allen Wood have employed to illuminating effect.[20] Wood thinks of Kant's effort to justify the

assumption that we are transcendentally free in doing ethics, despite our theoretical ignorance, as like the effort of a defense attorney to demonstrate that his client deserves the benefit of the doubt if we are not certain of his guilt. Pereboom points out that Kant's willingness to appeal to transcendental freedom in his justification of the death penalty, despite our theoretical ignorance, is more akin to the attitude of a prosecuting attorney who would unjustly deny the accused the benefit of the doubt. I think there is something apt about both these employments of the trial metaphor. On my view, the standards that transcendental freedom must meet to play its roles in our moral notions and practices are asymmetrical, and they can be met in some cases but not in others. So I think that transcendental freedom is the defendant in a number of cases before the bar, and that we should defend it in some of these cases, but prosecute it in others.

[20] See Pereboom 2006, p. 564, and Allen W. Wood, 'Kant's Compatibilism', in *Self and Nature in Kant's Philosophy*, edited by Allen W. Wood (Ithaca: Cornell University Press, 1984), p 73-101.