# Kant's Mature Theory of Punishment, and a First *Critique* Ideal Abolitionist Alternative

forthcoming in *Palgrave Kant Handbook*, ed. Matthew C. Altman

Benjamin Vilhauer, City College of New York

## Introduction

This chapter has two goals. First, I will present an interpretation of Kant's mature account of punishment, which includes a strong commitment to retributivism. Second, I will sketch a non-retributive, "ideal abolitionist" alternative, which appeals to a version of original position deliberation in which we choose the principles of punishment on the assumption that we are as likely to end up among the punished as we are to end up among those protected by the institution of punishment. This is radical relative to Kant's mature theory of punishment, but arguably it conforms better to the spirit of Kant's first *Critique* remarks on imputation and punishment than his mature theory does.

## Overview

Much of the interest and frustration of philosophy arises from the discovery that concepts that we use with easy consensus in everyday life become puzzling upon reflection. Concepts whose everyday use is fraught with controversy are even more confusing. This is common in ethics, but the concept of *just punishment* is an especially difficult case, since even a bit of reflection can raise doubts about its coherence. Though no two schools of moral philosophy express this point in quite the same way, morality's primary orientation seems to be the provision of reasons for us to avoid harming each other, and to help each other at least some of the time. Arguing for the possibility of just punishment seems, at least at first blush, to demand an inversion of that orientation, since it involves marshalling moral reasons in favor of harming people. The oddity

of this prompts some philosophers to deny that there can be just punishment, and to advocate its abolition.

Among philosophers who do think punishment can be justified, two basic strategies predominate, both historically and in the contemporary literature: retributivism and consequentialism. Retributivism is the view that punishment is justified by the fact that it inflicts on criminals the suffering they deserve for their actions. Retributivists typically acknowledge that punishing criminals can have valuable consequences, but they hold that these consequences are merely fortunate side effects of punishment, not part of what justifies it. Consequentialists see the justification of punishment in its valuable consequences, such as preventing criminals from reoffending and deterring potential criminals. They regard suffering as intrinsically bad, and as justified only if it is "outweighed" by its beneficial consequences. There are also various "mixed" justifications, according to which both retributivist and consequentialist justifications play roles in justifying punishment.

Kant makes remarks at various points that seem to support abolitionism, retributivism, consequentialism, and mixed justifications:

> <u>Abolitionism</u>
>
> A constitution providing for the **greatest human freedom** according to laws that permit **the freedom of each to exist together with that of others** [is a] necessary idea. … The more legislation and government agree with this idea, the less frequent punishment will become, and hence it is quite rational to assert (as Plato does) that in perfect institutional arrangements nothing of the sort would be necessary at all. … Even though this may never come to pass, the idea of this maximum is nevertheless wholly correct when it is set forth as an archetype, in

order to bring the legislative constitution of human beings ever nearer to a possible greatest perfection. (A316-17/B373-374)

Retributivism

Even if a civil society were to be dissolved by the consent of all its members (e.g., if a people inhabiting an island decided to separate and disperse throughout the world), the last murderer remaining in prison would first have to be executed, so that each has done to him what his deeds deserve … for otherwise the people can be regarded as collaborators in this public violation of justice. (MM 6:333)

Consequentialism

All punishments by authority are deterrent, either to deter the transgressor himself, or to warn others by his example. But the punishments of a being who chastises actions in accordance with morality are retributive.

All punishments belong either to the justice or the prudence of the lawgiver. The first are moral, the second pragmatic punishments. … All the punishments of princes and governments are pragmatic, the purpose being either to correct or to present an example to others. Authority punishes, not because a crime has been committed, but so that it shall not be committed. (LE 27:286)

Mixed Theory

There can be no *penal law* that would assign the death penalty to someone in a shipwreck who, in order to save his own life, shoves another, whose life is equally in danger, off a plank on which he had saved himself. … A penal law of this sort could not have the effect intended, since a threat of an ill that is still *uncertain* (death by a judicial verdict) cannot outweigh the fear of an ill that is *certain*

> (drowning). Hence the deed of saving one's life by violence is not to be judged *inculpable* … but only *unpunishable*. (MM 6:235-36)

It is hard to see how all these remarks can fit together in one coherent theory, and I will not claim to be able to do so here. Kant's abolitionist remark can of course be made consistent with the others, since any theory of punishment can be combined with the view that it would be good for there to be no crime to punish and thus no punishment. But it is worth noting that Kant's endorsement of abolitionism appears at the beginning of the critical period and is not a theme of his mature account of punishment, a point to which I will return. The other remarks are not so easy to reconcile. The point of MM 6:333 seems to be that murderers must get their just deserts even if it is impossible that their execution will deter anyone, and this is inconsistent with the last two remarks. The third remark suggests that state authorities should be focused only on deterrence in punishing. The fourth remark suggests a mixed theory because the claim that saving one's life by killing another is culpable but "unpunishable" seems to suggest that just punishment must be deserved but also deterring to be just.

We should probably accept that some of the difficulty in assigning Kant's position on punishment to one of the common categories used in the philosophy of punishment arises from genuine inconsistencies resulting from inadequate attention to detail on Kant's part. With a bit of cleanup work, though, we can extract a consistent and interesting theory which is grounded in his broader ethical and metaphysical views. He offers a complex mixed theory according to which we are obligated to punish retributively whether or not it deters, so long as the punishments do not violate a respect for persons principle to be described below, but which nonetheless appeals to consequentialist considerations to explain why punishment is a concern of the state. Before

exploring the details, however, it may be worth pausing to ask why Kant should have anything to do with consequentialism or retributivism in the first place.

As has been widely observed, the categorical imperative rules out consequentialist reasons as foundational moral reasons. In order to understand the basic directive to respect human beings as ends in themselves, "the end must here be thought not as an end to be effected but as an *independently existing* end" (G 4:437).[1] The rational nature essential to humanity is the only thing of absolute value, but that value does not give us reasons to, say, maximize the number of humans. It is an end to be respected rather than effected. The point is not that there is no reasoning about outcomes in Kantian ethics – interacting with others in ways that treat them as ends sometimes involves planning, for example. But the forward-looking considerations derive from foundational relations to the wills of others that are not fundamentally forward-looking. For these reasons, charitable interpreters should assume that whatever commitment to consequentialism is evident in Kant's remarks about punishment is one that derives from the underlying principle of respect for persons.

Is Kant's ethics retributive in a fundamental way? This is less clear. Attributing moral responsibility, or "imputation" in Kant's terms, is a precondition for retribution. If we take the foundation of Kant's ethics to be respect for persons, then it is important to note that in a number of remarks Kant describes imputability as a fundamental characteristic of personhood. In the *Doctrine of Right*, 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" (MM 6:223). But Kant's remarks on the fundamental characteristics of personhood are diverse, complex, and scattered throughout his texts, and do not always refer to imputation. For example, in the *Groundwork*, he tells us that

> every rational being, as an end in itself, must be able to regard himself as also giving universal laws with respect to any law whatsoever to which he may be subject; … this dignity … brings with it that he must always take his maxims from the point of view of himself, and likewise every other rational being, as lawgiving beings (who for this reason are also called persons). (G 4:438)

Does the idea of seeing ourselves as ends, and giving universal law along with all other rational beings, imply imputability? Clearly Kant's view is that it does, but his argument for this is not as clear as one might wish. Further, even if we accept that imputability is required for personhood, would we have to accept that *all* actions are imputable, or would it be sufficient to impute just some? Could there be special reasons of justice to avoid imputing blameworthy actions when serious retributive harm is at issue? I will argue below that Kant's first *Critique* view implies that there are such reasons, and I will show how this view disappears in his mature theory of punishment.

After offering an interpretation of Kant's mature account of punishment, I will propose a first-*Critique*-inspired reconstruction. Kant's mature view is that punishing in the absence of a retributive justification treats criminals as mere means, but if we emphasize another aspect of treating persons as ends, that is, respect for rational consent, then I think we can build what we might call an "ideal abolitionist" approach to punishment which avoids treating criminals as mere means and gives them an equal voice in the social contract by choosing principles of punishment in an original position in which we assume we are as likely to end up among the punished as we are among those protected by the institution of punishment.

**Kant's mature account of punishment**

With bit of charitable cleanup work, we can see that Kant's mature account of punishment involves three independent justificatory principles: a consequentialist principle that says the authorities have a duty to hinder hindrances to rights-violations with permissible means; a retributive principle that says we should punish all and only those who deserve it, with a kind and quantity of punishment equal to the crime; and a respect for humanity principle that sets limits to the kind and quantity of punishment even in cases where violation of those limits would yield more perfect retributive equality.

**The consequentialist principle**

The consequentialist principle in Kant's account of punishment is drawn out of his account of the rights of persons and the social contract. On Kant's view, our only "innate right" is "*freedom* (independence from being constrained by another's choice), insofar as it can coexist with the freedom of every other in accordance with a universal law"; this includes "innate *equality*, that is, independence from being bound by others to more than one can in turn bind them" (MM 6:237). Others are obligated to respect our equality in the state of nature, and we can hope that they do, but we are also justified in enforcing it, which we may do by coercively establishing what Kant calls a juridical state of affairs, or a condition of right:

> the state of nature … would still be a state *devoid of justice* (*status iustitia vacuus*), in which when rights are *in dispute* (*ius controversum*), there would be no judge competent to render a verdict having rightful force. Hence each may impel the other by force to leave this state and enter into a rightful condition. (MM 6:312)[2]

A rightful condition is structured by what Kant calls the "Universal Principle of Right": "'Any action is *right* if it can coexist with everyone's freedom in accordance with a universal law, or if on its maxim the freedom of choice of each can coexist with everyone's freedom in accordance with a universal law'" (MM 6:230). If my action is right, then "whoever hinders me in it does me *wrong*; for this hindrance (resistance) cannot coexist with freedom in accordance with a universal law" (MM 6:230-31). Coercion within a condition of right is justified when it is a rightful response to wrongful hindrance: "if a certain use of freedom is itself a hindrance to freedom in accordance with universal laws (i.e., wrong), coercion that is opposed to this (as a *hindering of a hindrance to freedom*)" is right (MM 6:231). What Kant presents in this account of the justification of coercion is not an argument from premises about right to a conclusion about coercion but rather an analysis of what he takes to be implicit in the concept of right (MM 6:232). In a civil state, we cede our individual freedom to coercively protect equality to the authorities, and the authorities take on a duty to do this. This yields a justification of the authorities' duty to coerce compliance with law which is consequentialist but markedly different from utilitarian consequentialism, since it aims not at happiness but instead at protecting freedom.[3]

Kant explicitly makes our right to hinder hindrances a right to constrain actions, not to constrain goodness of will or maxims of actions. This is the basis for his distinction between the distinct legislations of reason in ethical lawgiving, the topic of the *Doctrine of Virtue*, and juridical lawgiving, the topic of the *Doctrine of Right*. Ethical lawgiving requires an incentive which is itself ethical. But whether we follow the law because it is right is not relevant for juridical lawgiving. Juridical lawgiving couples laws with incentives of "aversion" which coerce

compliance so that all members of society are motivated to comply irrespective of the quality of their wills (MM 6:218).

When Kant speaks of coercively hindering hindrances through incentives of aversion, he seems to have fear of punishment chiefly (perhaps exclusively) in mind. But he does not argue that the authorities' duty to coerce through aversion contributes to the justification of punishment, and I think his considered view is that it does not. This is sensible, because there is no necessary connection between coercing through aversion and punishing to deter. There are many ways of coercing, and many kinds of aversions: for example, clever engineers may someday offer a roving corps of ticklebots always on the scene to make would-be criminals collapse in unbearable giggles before completing criminal acts.

The notion of a duty to coercively hinder hindrances provides no substantive guide to the means we may permissibly take to this end. When we focus on particular wrongful acts in progress, we may seem to get clear and intuitively satisfactory guidance – we are entitled to use whatever means are necessary to prevent that wrong from occurring. Now, Kant's theory of right as it stands may be enough to explain why it is right to apply just enough gentle, gradual pressure to the fingers of a potential apple thief to loosen his grip on my apple if this suffices to prevent the theft. But it is not helpful when we look at the wider range of coercive means. What if our would-be apple-snatcher is a dedicated but nonviolent thief who is bruising only produce in his thievish tenacity, but the only means at my disposal sufficient to prevent the theft is to seriously injure or kill him? If this is a rightful way of stopping the theft, then the fact that it is necessary to hinder a hindrance to my freedom does not provide an intuitively satisfactory explanation of why it is.

The questions become even more perplexing when we turn to coercive mechanisms that may be imposed on a criminal after the crime, such as preventative imprisonment or punishments that deter others. These practices are not aimed at any particular actual crime but instead at a field of possible crimes in the future. We might aim to prevent or deter all crimes, but even if we imposed a public, maximally painful execution on an apple-thief, we could not be sure of deterring all future apple-thieves. Even if we could be sure to deter them all, the "hindering hindrance" idea cannot plausibly be thought sufficient to allay moral concerns about severe punishments.

The consequentialist principle in Kant's account of punishment is also of little help in explaining *whom* we should punish. What if we could not capture the actual criminal? Could we frame and punish a scapegoat in his place to generate deterrence? Kant recognizes this as a challenge to purely consequentalist justifications:

> justice does not follow, if it is inflicted to improve the criminal, or as an example to others. This would simply have to do with its usefulness, and then it would be merely a means to that intent, for example, if somebody is flogged, whether guilty or not, in order to frighten people by his outcry, and create an impression. (LE 27:553)

To rule this out, we must regard punishment as "an immediately necessary consequence of the morally bad act" (LE 27:552-53), which Kant sees as an element of the retributive justification to be discussed below.

Kant himself may be tempted by the thought that guidance about means flows from the idea of reciprocal coercion. He says that the "law of a reciprocal coercion" can be constructed "by analogy with … bodies moving freely under the law of the *equality of action and reaction*"

(MM 6:232), which may be meant to suggest a necessary connection with the retributive equality in *ius talionis* (to be discussed below). But it is far from clear how an analogy between people interacting in conformity with duties of right and bodies moving in conformity with natural laws could provide a morally substantive link between the concept of reciprocal coercion and the idea of retributive equality. Presumably Kant is not making the crude claim that the law according to which, when rock A hits rock B, rock B hits A back just as hard, helps justify *ius talionis.* However, even the more abstract notion of a construction by analogy endows this natural law with more moral significance than seems sensible.[4]

Another point at which Kant may seem to suggest that guidance about legitimate punitive means flows from the scheme of reciprocal coercion is when he claims that "the mere idea of a civil constitution among *human beings* carries with it the concept of punitive justice belonging to the supreme authority." However, he goes on immediately to point out the question of "whether it is a matter of indifference to the legislator what kinds of punishment are adopted, as long as they are effective measures for eradicating crime" (MM 6:362), implicitly acknowledging that the answer to this question does *not* follow from the idea of a civil constitution. He then goes on to refer to both the retributivist principle (*ius talionis*) and the respect for persons principle ("crime against *humanity* as such") discussed below as necessary parts of the answer.

For these reasons, the consequentialist dimension of Kant's account plausibly generates the principle that *the authorities are obligated to use morally permissible means to the end of coercing compliance with laws that protect right*, but is properly seen as silent about what the morally permissible means are.[5]

**The retributive principle**

That Kant is committed to a retributive principle in the justification of punishment is clear from at least the Collins lecture notes (dating from 1784) onward, for example, at LE 27:286 (cited above). In that passage Kant is unclear about the relationship between deterrence and retributivism in the justification of punishment of criminals at the hands of the authorities. On the one hand, he states that "all punishments by authority are deterrent. … But the punishments of a being who chastises actions in accordance with morality are retributive," seeming to claim that deterrence has a role in state punishment that retribution lacks. But on the other hand, he states that "all punishments belong either to the justice or the prudence of the lawgiver," clearly implying that justice is about retribution and prudence is about deterrence, and thus that retribution has a role in lawgiving too.

By the time of the *Metaphysics of Morals* in 1797, Kant makes it clear that consequentialist considerations are morally and legally irrelevant unless a retributive justification has been established:

> *Punishment by a court* (*poena forensis*) … can never be inflicted merely as a means to promote some other good for the criminal himself or for civil society. It must always be inflicted upon him only *because he has committed a crime*. … He must previously have been found *punishable* before any thought can be given to drawing from his punishment something of use for himself or his fellow citizens. (MM 6:331)

It is important to emphasize that, on Kant's account, retributivism implies not only that punishment is *permitted* when we legitimately impute a crime, but also that it is *required*. This is what Kant means in describing punishment as "necessary" (LE 27:552, quoted above), in his

stern admonition that "the law of punishment is a categorical imperative, and woe to him who crawls through the windings of eudaimonism in order to discover something that releases the criminal from punishment" (MM 6:331), and in his command that the last murderer be put to death even as society is about to dissolve (MM 6:333, quoted above). This means that the retributive principle is doing a lot more heavy lifting in Kant's account of punishment than the consequentialist principle is: the only contribution of the consequentialist principle is to explain why it is appropriate to turn the retributively justified practice of punishment to the end of coercing compliance with law.

Earlier we saw that Kant's consequentialist principle is rooted in his account of right. Is his retributive principle similarly rooted elsewhere in his broader moral theory? Kant expends little effort in arguing that it is, and commentators have understandably taken him to task for this.[6] In the Vigilantius lecture notes, Kant remarks that the necessary link between "moral badness" and punishment implied in retributivism "cannot be discerned through reason, nor proved either, and yet it is contained in the concept of punishment," apparently suggesting that the spade turns here and no further explanation can be given (LE 27:552).

However, retributivism has deep roots in Kant's doctrine of transcendental freedom, a kind of free will which would make it the case that we have alternative possibilities of action despite the determinism of the phenomenal world. The indiscernibility of retributivism is part and parcel of the indiscernibility of the transcendental freedom it presupposes. In the first *Critique,* Kant holds that we cannot know we are transcendentally free: he claims to prove only that determinism "**does not conflict with** causality through freedom" (A558/B586), and he acknowledges that the idea of transcendental freedom may be "merely invented" (A545/B573). We can only have knowledge of things that happen in experience as events in a deterministic

series, because the empirical world must be constructed deterministically for there to be an objective order of time. In the second *Critique*, there is a shift in Kant's epistemology and he argues that "the moral law assures us" of transcendental freedom, but it remains "inexplicable" (CPrR 5:99).

At A448/B476, Kant sets out a view that he holds throughout the critical philosophy, that the "transcendental idea of freedom" constitutes the idea of "the absolute spontaneity of an action, as the real ground of its imputability" (A448/B476). In a "judgment of imputation" about "blame," "blame is grounded on the law of reason, which regards reason as a cause that, regardless of all the empirical conditions … could have and ought to have determined the conduct of the person to be other than it is" (A555/B583). In other words, the imputation of blame is grounded on the claim that person blamed could have acted differently than he in fact acted in doing the wrongful deed. In more contemporary terms, transcendental freedom serves as the basis of an alternative possibilities account of moral responsibility and desert, both of which are presupposed by retributivism. The claim that someone could have done otherwise is of course not sufficient to justify the claim that he deserves to suffer – it is also necessary, for example, that he understood what he was doing and that it was wrong. But Kant is certainly not alone in thinking the alternative possibilities condition is a necessary condition which, in combination with whatever other necessary conditions there may be, is sufficient to justify retribution.

To explain what kind and quantity of punishment we should impose, Kant turns to what he calls the "*law of retribution* (*ius talionis*)":

> what kind and what amount of punishment is it that public justice makes its principle and measure? None other than the principle of equality. … Accordingly,

> whatever undeserved evil you inflict upon another within the people, that you inflict upon yourself. If you insult him, you insult yourself; if you steal from him, you steal from yourself; if you strike him, you strike yourself; if you kill him, you kill yourself. (MM 6:332)

There are of course many reasons to object to this principle. One objection that Kant thinks he can meet concerns the notion of "equality" at work here: how do we take an eye for an eye if the person to be punished has lost one eye already? Kant appeals to a notion of similarity when a strict notion of equality is inapposite:

> A fine … imposed for a verbal injury has no relation to the offense, for someone wealthy might indeed allow himself to indulge in a verbal insult on some occasion; yet the outrage he has done to someone's love of honor can still be quite similar to the hurt done to his pride if he is constrained … not only to apologize publicly to the one he has insulted but also to kiss his hand. (MM 6:332)

Clearly such fine-tuning cannot be an exact science, and the inevitable asymmetries between most criminals and victims make this problem ubiquitous. But perhaps if we are as exact as we can be, this is not a fundamental limitation in *ius talionis* as a moral principle in determining the kind and quantity of punishment. There are, however, cases that *do* point out such a limitation, that is, "crimes that cannot be punished by a return for them because this would be … itself a punishable crime against *humanity* as such, for example, rape" (MM 6:363). Kant also includes torture and execution by torture in this category: "No punishment should be coupled with cruelty, i.e., it must not be so framed that humanity itself is thereby brought into contempt" (LE 27:556). He also states that, when "death is judicially carried out upon the wrongdoer," it must "be freed from any mistreatment that could make the humanity in the person suffering it into

something abominable" (MM 6:333). In cases like the verbal insult case, we ought to return like for like, an injury to honor with an injury to honor, and the limitation in arranging a relation of like injury for like injury is due to differences between the agents involved, not due to a moral prohibition on returning like for like. The question of punishments themselves punishable as crimes against humanity arises in contexts where there is a moral prohibition on returning like for like because of the respect for persons principle discussed in more detail below.

**The respect for persons principle**

It is easy to overlook Kant's reliance on an independent respect for persons principle in his account of punishment when confronted with the remarks on *forfeiture* scattered through the *Doctrine of Right*: a criminal is someone who has "forfeited his personality" (MM 6:358), the only "human being in a state [who] can be without any dignity … is someone who has lost it by his own *crime*" (MM 6:329-30), and a criminal can be "condemned to lose his civil personality" (MM 6:331). He explains this doctrine at greater length in "Theory and Practice," where he explicitly connects the innate "equality" contained in the innate right to freedom (MM 6:237-38, discussed above) with criminal forfeiture of this right:

> whoever is *subject* to laws is … subjected to coercive right equally with all the other members of the commonwealth. … No one of them can coerce any other except through public law …, through which every other also resists him in like measure; but no one can lose this authorization to coerce … except by his own crime, and he cannot give it away of his own accord, that is, by a contract, and so bring it about by a rightful action that he has no rights but only duties; for he

> would thereby deprive himself of the right to make a contract and thus the contract would nullify itself. (TP 8:291-92)

Kant tells us that "beings that have only duties but no rights" are "human beings without personality (serfs, slaves)" (MM 6:241), and prescribes enslavement for thieves: he remarks that it is "by the principle of retribution" that the convicted thief "must let [the state] have his powers for any kind of work it pleases (in convict or prison labor) and is reduced to the status of a slave for a certain time, or permanently if the state sees fit" (MM 6:333). Together these remarks suggest that Kant's doctrine of forfeiture involves bringing his retributive principle to bear upon the scheme of universal reciprocal coercion to justify the criminal's expulsion from it. That is, he looks to the retributive principle to explain why some people lose their right to equal coercive power and can instead be used as means to the end of coercion. Since they could have avoided committing crimes and therefore deserve to suffer, their expulsion from this scheme and use as deterrence-generators uses them as means but not as *mere* means. If this is right, then the retributive principle operates in a sense at two distinct levels: it explains why criminals deserve to suffer and why criminals can be coerced but lose the right to coerce in return (have duties but no rights).

Now, this idea of forfeiture is crucial in Kant's account of punishment, but by Kant's own lights what he says about it cannot be quite right. If criminals truly had no rights, then there could be no such thing as a punishment which is "itself a punishable crime against *humanity* as such" (MM 6:363) – on Kant's theory of right, there can only be such a punishment if criminals can legitimately coerce others through public law not to impose such a punishment. Kant's considered view seems to be that we cannot forfeit our personality entirely, no matter how we act. He refers to the unforfeitable core as "innate personality" (MM 6:331). It is on the basis of

this innate personality that he says, "I cannot deny all respect to even a vicious man as a human being; I cannot withdraw at least the respect that belongs to him in his quality as a human being, even though by his deeds he makes himself unworthy of it," and on this basis he rules out "disgraceful punishments that dishonor humanity itself (such as quartering a man, having him torn by dogs, cutting off his nose and ears)" (MM 6:463).[7]

The upshot of this discussion is that, while it is clearly Kant's view that by the principle of retribution we can forfeit a large proportion of what we tend to think of as key attributes of personhood in Kant's ethics, it is equally clear that there remains a much-diminished but nonetheless crucial notion of innate personhood that we cannot forfeit, and which plays a necessary role in constraining the principle of retribution, by limiting the kind and degree of punishment we can impose by the *ius talionis*.

**An objection to the retributive principle**

Critics have made various objections to Kant's retributivism over the years and I lack space to comment on them here in any detail, so I will focus on an objection derived from a point that Kant himself makes in the first *Critique.*

As mentioned above, transcendental freedom is a power which would make it the case that we have alternative possibilities of action despite the determinism of the phenomenal world, and would serve as the "real ground" of imputability (A448/B476, cited above). As also already mentioned, according to the first *Critique*, we cannot know that we have transcendental freedom. This means we cannot know that our imputations have real grounds. We apply the idea of transcendental freedom as a "regulative principle" (A554/B583) that we apply when we impute actions, but we cannot know that these imputations have the sort of ground they require to be just.

Kant's theory of transcendental freedom involves a distinction between agents' empirical and intelligible characters. In an agent's empirical character, "actions, as appearances, would stand through and through in connection with other appearances in accordance with constant [deterministic] natural laws," and "in combination with these other appearances, they would constitute members of a single series of the natural order." But we represent the agent as also having an "**intelligible character**, through which it is indeed the cause of those actions as appearances, but which does not stand under any conditions of sensibility," and this allows us to represent the intelligible character as exempt from deterministic laws (A539/B567). A human being's intelligible character is the nondeterministic noumenon we posit as the ground of the actions we impute according to the regulative idea of transcendental freedom. If, as Kant holds in the first *Critique*, we cannot know that we are transcendentally free, then we cannot know that we have intelligible characters that provide real grounds for imputation, and this calls into question the justice of imputation:

> The real morality of actions (their merit and guilt), even that of our own conduct, therefore remains entirely hidden from us. Our imputations can be referred only to the empirical character. How much of it is to be ascribed to mere nature and innocent defects of temperament or to its happy constitution (*merito fortunae*) this no one can discover, and hence no one can judge it with complete justice. (A551n/B579n)

The context of this quotation makes it clear that Kant's worry was not that, while we could be sure that our imputations *sometimes* had real grounds, we could not be sure when they did, or how robust they were, as he is sometimes taken to mean. The worry is rather that our imputations might *never* have real grounds – that is, grounds in the intelligible character through

transcendental freedom – at all, and therefore that the real morality of our actions cannot be judged with complete justice. If we think that imputing blame must be a matter of complete justice, then this would prohibit it. Some may worry that this concern is overblown. How much confidence do we really need about the foundations of retributive justification to legitimately justify punishment retributively in our everyday legal practices?

In the Vigilantius notes, Kant advocates a very strong standard for the imputation of blame (*imputatio demeriti*), that is, one of certainty:

> The *factum* itself, for the man who wishes to impute it, and claim that the accused is its *auctor*, must have the utmost moral and logical certainty. … For even the highest degree of probability, if it were to play the part of moral certainty, would be [accompanied by] the risk of injuring the other's rights; the imputation would be staked on such a risk. But here the slightest gamble is at all times disallowed. (LE 27:566)

It may be objected that surely Kant here means to speak of certainty only on the matter of whether the particular action was done by the particular person with a relevant motive, not on the matter of the metaphysical foundations of imputation. However, the Vigilantius notes have Kant painstakingly drawing a contrast that appears to be the same as the empirical/intelligible character contrast that he makes in the first *Critique,* and stating very clearly that *imputatio facti* lies "not simply and solely in the fact that [the criminal] is a rational being" with a "motive to the action" – it is "absolutely necessary in addition, that he act with freedom, indeed it is only when considered as a free being that he can be accountable" (LE 27:559).

It may also be objected that Kant discovers in the second *Critique* that we really can be certain that we have transcendental freedom after all, and it is this discovery that allows

imputation to meet the certainty standard propounded in the Vigilantius notes. That is, by the time we arrive at the second *Critique,* all doubt about transcendental freedom has been mooted by Kant's discovery of an argument that proves the reality of transcendental freedom in the "absolute sense in which speculative reason needed it" (CPrR 5:3): the "moral law is given, as it were, as a fact of pure reason … which is apodictically certain" and it "serves as the principle of the deduction of an inscrutable faculty …, namely the faculty of freedom, of which the moral law … proves not only the possibility but the reality in beings who cognize this law as binding upon them" (CPrR 5:47). This argument turns on the "ought implies can" principle: we judge that we "can do something" because we know we "ought to do it" and thereby "[cognize] freedom within" ourselves (CPrR 5:30). But this argument violates a basic epistemological constraint that Kant establishes in the first *Critique*: it is a transcendental argument that yields a conclusion about things in themselves rather than the form of experience. It is also a weak argument, because it is not clear why the "ought implies can" principle cannot be adequately accommodated by Kant's earlier view in the *Groundwork*, which does not rely on knowledge of transcendental freedom.

In the *Groundwork*, Kant argues that rational beings necessarily reason and deliberate "*under the idea of* [transcendental] *freedom*," and that this implies that "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). Part of the guiding thread in this argument is the idea that we cannot make a decision about what to do unless we assume that we have alternative possibilities of action, and when confronted with alternatives, the question of which one we *ought* to pursue inevitably arises, and in reflecting on this question in pure practical reason we see that the only thing we *unconditionally* ought to do is act morally. In this

argument Kant's view is that the alternatives necessarily posited in practical reasoning provide a robust enough ontological foundation for "oughts," and it is not clear why we ought to reject this view in favor of the second *Critique* view.

Now, in the *Groundwork* discussion, Kant makes no reference to his first *Critique* worry that our lack of knowledge of transcendental freedom leaves us with a concern about justice. That is, if *all* laws bound up with freedom hold for us, then presumably those include the principles governing imputation and retribution which become central in the *Doctrine of Right*. But there would seem to be plenty of conceptual space for views that combine the *Groundwork*'s acting-under-idea view with the first *Critique* concern. We could hold, for example, that beings who must act under the idea of transcendental freedom ought to act under all the laws bound up with freedom except in cases where our lack of knowledge raises concerns of justice. Further, Kant's undeveloped first *Critique* notion of imputation to the empirical character might be a useful tool in mapping this conceptual space: it seems quite intuitive to think that while it may be necessary to regard myself as having alternative possibilities of action when I deliberate about how to respond to a crime, I can at the same time regard the crime as caused by the history of the criminal's empirical character, and stop short of the imputation to the intelligible character which would provide a real ground for imputation.

Kant's shift to his second *Critique* epistemology of transcendental freedom may in part be due to the need to overcome his sense of injustice about the imputation of blame on uncertain grounds. And overcome it he does, in dramatic style: there he remarks on "children" who are taken to be "born villains" because they show "such early wickedness and progress in it so continuously." Since the "moral law assures us" that this "has as its basis a free causality" of "evil and unchangeable principles freely adopted," this makes their behavior "only more

culpable and deserving of punishment" (CPrR 5:99-100). (This must be one of the saddest things Kant ever wrote. From the vantage point of contemporary child development theory, it is hard not to imagine a single series of the natural order in which wicked children punished by parents on Kant's grounds grow up to punish their own wicked children on Kant's grounds with each new generation.) Astonishingly, Kant shares with us that, in this way, "appraisals can be justified which, though made in all conscientiousness, yet seem at first glance quite contrary to all equity" (CPrR 5:99) As we have seen, the appearance of contrariety to equity is not just at first glance, but upon thorough consideration in the first *Critique*.

For these reasons, it seems fair to conclude that the second *Critique* shift in the epistemology of imputation fails to meet the "utmost moral and logical certainty" standard set out in the Vigilantius notes. I think we should also resist the idea that we should accept a compatibilist or priority-of-the-practical reconstruction of Kant's theory of freedom *in order to* resolve problems in the epistemology of imputation, because I think that would be to philosophize in bad faith. (This is not to argue that there are no other reasons to prefer these reconstructions. I am not convinced that there are, but I cannot argue that here.)

**An "ideal abolitionist" original position reconstruction**

How might a reconstruction that did not incorporate transcendental freedom look? We would have to reject retributivism and *ius talionis*, but also the underlying notion that criminals *could have done otherwise* when they committed their crimes – what I will call the "avoidability of crime assumption."

At many points in Kant's texts, his strategy for treating people as ends does not focus on desert, but instead on treating people as they would rationally consent to be treated. In his

philosophy of right, for example, the justice of the social contract derives from it being one to which would receive the hypothetical rational consent of all parties to it (TP 8:297). But Kant rejects Beccaria's social-contract-based critique of capital punishment (no one could rationally consent to execution and thus capital punishment cannot be included in the social contract) and suggests that Beccaria is making a conceptual error in posing the question of whether criminals can rationally will punishment, since "it is no punishment if what is done to someone is what he wills." However, Kant endorses the view that "saying that I will to be punished if I murder someone is saying nothing more than that I subject myself together with everyone else to the laws, which will naturally also be penal laws if there are any criminals among the people" (MM 6:335). So Kant does think that the social contract involves my rational consent to a law that prescribes my execution *if* I become a murderer. Clearly execution is a very undesirable outcome, and it is rational for us to avoid undesirable outcomes when evaluating potential laws for the social contract, and to give or withhold our consent accordingly. Kant thinks that undesirable outcomes for commoners of laws granting "hereditary privilege of *ruling rank*" are reasons to reject such laws (TP 8:297).[8] So why is the undesirable outcome of harsh punishments for criminals not a reason to reject laws that prescribe them? The idea seems to be that it is not under my control whether I am a commoner or a hereditary noble, but it is under my control whether I become a criminal or not, and I must therefore consent to the social contract under the assumption that I will avoid committing crimes (an assumption of what Rawls calls "ideal theory"[9]). If it is rationally incumbent upon us to regard our consent to the social contract in this way, then, as Kant puts it, criminals "cannot possibly have a voice in [this] legislation" (MM 6:335). But if skepticism about transcendental freedom means that we must reject the

avoidability of crime assumption, then it cannot be rationally incumbent upon us to make the avoidability of crime assumption when we consent to the social contract.

What would a reconstructed Kantian social contract look like that rejected the avoidability of crime assumption? One plausible approach is by way of a version of original position deliberation in which we assume that we are just as likely to end up among the punished as we are among the unpunished.[10] If we assume this, we would have a strong preference for a society that did not punish at all – we would necessarily make abolition our goal. Our first priority would be to pour resources into noncoercive preventative strategies, such as free voluntary therapy for people at risk of crime, drug treatment, publicly funded education and job programs in areas where unemployment may be an incentive to crime, and much greater funding for public services.

But it would be reasonable to be concerned that abolishing all criminal coercion before we were able to make broader social changes that diminished the incentives to commit crimes would effectively throw us back into the state of nature, leaving the freedom of everyone insecure. If we could devise a coercive scheme sufficient to maintain a condition of right that places a light enough burden on those convicted of crime to give them more choiceworthy lives than they would have in the state of nature, then it would be rational to consent to it. In determining whether there could be such a scheme, a lot turns on how we understand "valuable." If we define value here in utilitarian terms, then it is easy to imagine that there could be such a scheme. A life on a carefully titrated opioid drip in a utility-maximizing prison would surely be more pleasurable than the nasty, brutish, and short life one can anticipate in the state of nature. But from the Kantian perspective, the most choiceworthy life is the freest life, and it might sound absurd to suggest that even the lightest criminal coercion scheme could offer criminals greater

freedom than the state of nature. But it must be kept in mind that the kind of freedom valuable from the Kantian perspective is not the "wild, lawless freedom" of the state of nature, but rather lawful freedom (MM 6:316). From the perspective of ethics it is the freedom of acting for the sake of the moral law, and from the perspective of right it is the freedom of acting without violating the limits of others' rightful freedom (which corresponds roughly with the freedom to act permissibly from the perspective of ethics). The wider scope of lawless freedom, including the freedom for criminals to steal or kill in the state of nature, would not make such a life worthy of choice from the Kantian perspective. (It would be Orwellian to place too much weight on this point, but it is instructive if applied cautiously.) So it is not absurd to speculate that a social contract with the right sort of coercive scheme might afford even convicted criminals more worthwhile freedom than the state of nature would.

What sorts of coercive constraints would be worth considering if we knew that we were as likely as not to be constrained by them? Preventative coercion focusing on crimes in progress would take on a new importance, since it would seem to be rational from the Kantian perspective (and other perspectives too) to prefer being prevented from completing a criminal act to being subject to coercive constraints after the fact. But the availability of choiceworthy strategies for doing this is relative to a society's development of relevant technology and social practices.[11] If we wanted to rely completely on preventative strategies, then the options readily at hand today would include combinations of blanket surveillance and militarized policing, which would protect us by making the whole world a prison. Even ticklebots might not seem as benign in the hands of a police state as they do in a thought experiment.

Are there after-the-fact coercive constraints that it would be rational to endorse while we work on more choiceworthy technological and social preventive innovations? I think that for

nonviolent thefts, it would be rational to consent to measures such as fines and ankle-monitors for the thieves we could catch, and a system of social insurance to compensate victims of theft when we could not catch the thieves. In the original position, we would not care so much about protecting our property that we would risk imprisonment for it.

Violent crimes are another matter. It would be rational to risk imprisonment in the right sort of prison in order to be protected from murder or serious injury. But to pass the test of affording greater lawful freedom than the state of nature, it would obviously have to be a very different sort of prison that what we have today, one with real opportunities for social interaction, voluntary therapy, education, meaningful work, and maintenance of relationships with friends and family in the world outside. In the original position we would not merely pay lip service to these goals as most of us do today – we would recognize the radical reform needed to achieve them. We would only consent to imprisonment while offenders were still risks for violence and would insist on a continuous parole review process leading to rapid though probably monitored release into the community.

Would it be rational to consent to after-the-fact coercive measures designed to deter, or just ones designed to prevent reoffending? We would not consent to harsh measures for the sake of deterrence, and certainly not the death penalty. But we could not forgo considerations of deterrence completely, because if after-the-fact responses to thieves or violent criminals made criminal life too pleasant, then they would become new incentives to commit crime, and we would be cast back into the state of nature once again. So we would need to maintain at least the minimum level of aversiveness necessary to prevent the system of criminal justice from becoming an incentive to commit crime. But it seems plausible to think that we could maintain this minimum in a prison and still give prisoners a life with greater scope for lawful freedom

than the state of nature affords. It is important to keep in mind that aversiveness is a relative matter – we have an aversion to a less-pleasant option when a more-pleasant option is available, even if the less-pleasant option is not intrinsically unpleasant. So the minimum level of aversiveness would be relative to conditions outside prison, and given the way the original position prioritizes noncoercive preventative measures such as education, jobs, and social services, life on the inside of prison would not have to be intrinsically unpleasant for it to be less pleasant than life on the outside.

In this way, the original position approach to punishment provides guidelines about the kind and quantity of punishment we should impose. Can it tell us whom we should punish? The retributive principle tells us to punish the deserving, but if we design an original position to rule out considerations of desert, we need a different reason to limit punishment to actual criminals in cases where punishing a scapegoat would improve deterrence. It may be objected that without an appeal to desert there can be no legitimate reason to prefer punishing the actual committers of crime in such cases. But I think that our intuition about the wrongness of scapegoating really has two roots, one having to do with desert, and another having to do with the deception of the public required for scapegoating to generate deterrence.

Our lack of knowledge of transcendental freedom pulls up the desert root, but not the deception root. While Kant does not distinguish them, his philosophy nonetheless provides reasons for rejecting scapegoating which are grounded in the deception root. Deception manipulates in a way that uses rational beings as mere means. In the context of right, Kant propounds the "*transcendental formula* of public right": "'All actions relating to the rights of others are wrong if their maxim is incompatible with publicity.' This principle is not to be regarded as *ethical* only (belonging to the doctrine of virtue) but also as *juridical* (bearing upon

the right of human beings)," and rules out maxims "that I cannot *divulge* without thereby defeating my own purpose" (PP 8:381). A publicized principle of punishing scapegoats to deter crime would be self-defeating because it would actually undermine deterrence, since potential criminals are deterred by the worry that *they* will be punished for their crimes, not that scapegoats may be punished in their place.

Further, when contemplated from the standpoint of original deliberation about punishment, a principle that aims to deter by scapegoating is self-deceiving and in a sense self-contradictory. Since I may well be among the deceived, I would be volunteering to be deceived about a fundamental principle of society which I had chosen myself. Self-deception on this scale undermines one's status as a rational agent in a way that is similar to consenting to slavery and should be recognized as self-contradictory for similar reasons.

It may be objected that the structure of the original position I rely on in this approach is arbitrary in assuming an equal probability of finding oneself among the punished and among the unpunished. Certainly it is not as tidy as the notion of original position deliberation that Rawls applies to distributive justice. In that context we can apply "maximin" reasoning because there is a unique candidate for the worst-off position, the poorest.[12] In the context of criminal justice, however, there are two candidates for the worst-off position, criminals and those they threaten. The purpose of original position deliberation is to proceduralize Kantian fairness, and it seems to me that the fair thing to do in dealing with a socially necessary practice that benefits some at the expense of others is to assume equal probability of being on both sides.

It may also be objected that a system of criminal justice that dispenses with desert cannot be called a system of punishment at all, because punishment involves desert as a conceptual matter.[13] I am not sure that this is right, but if the dominant usage of "punishment" takes it to

imply blame, then I would in one sense be happy to go along with this, and call this an account of criminal justice but not punishment, but I fear that there is a hazard of euphemism here which is more pressing than the hazard of violating norms of usage. Referring to a practice as *punishment* highlights that it cries out for justification, but calling it *criminal justice* does not, and philosophers who take skepticism about imputing blame seriously ought to keep the demand for justification clearly in view.

It is therefore important to be clear what this approach can and cannot do as far as justifying punishment goes. It is an approach that weighs consequentialist considerations but makes them morally significant only insofar as they derive from the rational consent of each party to punishment, and this means that the consequentialist considerations derive from a non-consequentialist foundation in respect for persons. It differs from the way that Kant draws consequentialist considerations out of respect for persons, in that it shows how strict guidelines for whom we should punish, and the kind and quantity of punishment we should impose, can rise directly from this foundation rather than by way of an interjected retributive principle. It also differs from Kant's approach in that it treats criminals as ends by respecting their rational consent rather than by imputing blame and punishing retributively.

But it might reasonably be objected that this can only be a partial justification of punishment, because nothing but desert could alleviate the basic unfairness of the fact that some are punished while others are not. I think this is correct, and it highlights the importance of the fact that the original position deliberators in this story would set the abolition of punishment as their top priority, and would endorse after-the-fact coercive measures only as temporary solutions to maintain the coercive protection of right en route to abolition. I think that this is in the spirit of Kant's best thought on punishment, with which we began: that "the more legislation

and government agree with this idea [of a perfect state], the less frequent punishment will become," and that we must strive "to bring the legislative constitution of human beings ever nearer to [this] possible greatest perfection" (A316-17/B373-74).

---

**Notes**

1. For a helpful discussion of this point, see Allen W. Wood, *Kantian Ethics* (Cambridge: Cambridge University Press, 2008), 85.

2. The details of Kant's explanation of establishing a condition of right appeal to acquired rights to property (MM 6:312), but this appears to be an expository matter rather than a necessary feature of his argument, and I think it obscures its basic structure.

3. This point is emphasized in Don E. Scheid, "Kant's Retributivism," *Ethics* 93, no. 2 (Jan. 1983): 262-82.

4. For the view that retributivism can be "constructed" along these lines, see Susan Meld Shell, "Kant on Punishment," *Kantian Review* 1 (March 1997): 115-35; and Jane Johnson, "Revisiting Kantian Retributivism to Construct a Justification of Punishment," *Criminal Law and Philosophy* 2, no. 3 (Oct. 2008): 291-307. For the view that it cannot, see Paul Gorner, "The Place of Punishment in Kant's *Rechtslehre*," *Kantian Review* 4 (March 2000): 121-30.

5. This explanation of the consequentialist principle's contribution to Kant's justification of punishment is related to, but distinct from, explanations by some recent scholars that draw on Rawls's distinction between justifying a practice and justifying actions based on rules internal to the practice. See John Rawls, "Two Concepts of Rules," *Philosophical Review* 64 (1955): 3-32. Their notion is that, for Kant, consequentialism explains why it is morally appropriate to have a practice of punishment, and the rules internal to the practice are retributive. But since Kant's notion of the duty to coerce provides no

guidance about means, it has no necessary connection to punishment, and it therefore cannot provide what H. L. A. Hart calls the "General Justifying Aim" of punishment in any sense in which it does not also provide the general justifying aim of ticklebots (*Punishment and Responsibility: Essays in the Philosophy of Law*, 2nd ed. [Oxford: Oxford University Press, 2008]). For this approach, see B. Sharon Byrd, "Kant's Theory of Punishment: Deterrence in Its Threat, Retribution in Its Execution," *Law and Philosophy* 8, no. 2 (Feb. 1989): 151-200; and Thomas E. Hill, *Human Welfare and Moral Worth: Kantian Perspectives* (Oxford: Oxford University Press, 2002). For a critique, see Allen W. Wood, "Punishment, Retribution, and the Coercive Enforcement of Right," in *Kant's "Metaphysics of Morals": A Critical Guide*, ed. Lara Denis (Cambridge: Cambridge University Press, 2010), 111-29.

6. See Wood, "Punishment, Retribution, and the Coercive Enforcement of Right," 121.

7. Jean-Christophe Merle also emphasizes the importance of respect for persons, in "A Kantian Critique of Kant's Theory of Punishment," *Law and Philosophy* 19, no. 3 (May 2000): 311-38.

8. Matthew Altman also addresses the limits of Kant's response to Beccaria – he argues that even if murderers deserve to die, the fact that we cannot be sure we will not be wrongfully accused of murder gives us reason to exclude capital punishment from the social contract. See Matthew C. Altman, "Subjecting Ourselves to Capital Punishment," in *Kant and Applied Ethics: The Uses and Limits of Kant's Practical Philosophy* (Malden, Mass.: Wiley-Blackwell, 2011), 117-38.

9. John Rawls, *A Theory of Justice*, rev. ed. (Cambridge: Harvard University Press, 1999), 215-17.

10. Here I draw on a longer discussion of this approach to punishment, in Benjamin Vilhauer, "Punishment, Persons, and Free Will Skepticism," *Philosophical Studies* 62,

no. 2 (Jan. 2013): 143-63. Sharon Dolovich proposes a similar approach, but not in the context of free will skepticism. See Dolovich, "Legitimate Punishment in Liberal Democracy," *Buffalo Law Review* 7, no. 2 (Jan. 2004): 314-29.

11. This point is central to the objection developed in Dimitri Landa, "On the Possibility of Kantian Retributivism," *Utilitas* 21, no. 3 (2009): 276-96.

12. Rawls, *Theory of Justice*, 130-38.

13. See Anthony Quinton, "On Punishment," *Analysis* 14, no. 6 (1954): 133-42; and more recently**,** Wood, *Kantian Ethics*, 208.