On the Possibility of Kantian Retributivism

Dimitri Landa

Abstract

One the most potent motivations for retributivist approaches to punishment has been their apparent connection to an ethical background shaped by the Kantian notion of morally autonomous and rational human agency. The present paper challenges the plausibility of this connection. I argue that retributivism subverts, rather than embodies, the normative consequences of moral autonomy, justifying a social practice that conflicts with the considered judgments that the proper recognition of moral autonomy would authorize. The core of my case is the analysis of whether a punishment should be understood as a restriction of a criminal’s freedom properly understood. I argue that the affirmative view faces serious difficulties that have not been, and are not likely to be, resolved by retributivist justifications that draw their support from Kantian moral theory.

1 Assistant Professor of Politics, New York University, e-mail: dimitri.landa@nyu.edu
I. Introduction

Despite scores of criticisms, retributive justifications of punishment have largely endured the test of time. Even when not billed by their authors as retributivist, some of the most prominent justificatory theories of the last several decades contain strong signs of retributivist inspiration and are frequently presented as responses to the objections of anti-retributivist critics. The most prominent of these objections, charging retributivism with “weird metaphysics” and with, at times, authorizing unreasonable and unnecessarily high levels of punishment, are met with rationalizations appealing to some of our strongest felt moral convictions.

The details of what counts as retributivism vary from one justificatory theory to the next, with some theories, such as Mabbott’s (1939), eschewing a close connection between retributivism and morality, and others placing such a connection at the center of the account. The primary focus of this paper is on theories that treat retributivism as a moral requirement, and, more narrowly, on theories that seek to ground that requirement in Kant’s moral theory. The core claim of the “moral retributivist” approach may be captured by the proposition that “when breach of the law involves moral guilt the application to the offender of the pain of punishment is itself a thing of value” (Hart 1968, p. 8). Somewhat more precisely, theories that adopt this proposition characterize retributivism by two defining premises: the existence of agency capable of distinguishing moral right from wrong and the moral sufficiency of the criminal act for the imposition of punishment. The first of these makes the retributivist practice of repayment or recompense non-vacuous, while the second identifies the domain to which this practice is addressed. Perhaps the most enduring motivation for these premises has been their apparent connection to an ethical background shaped by the Kantian notion of morally autonomous and rational human agency. Indeed, it is often implied that retributivism is worth defending precisely
because it captures something critical about this notion of agency, something that is missing from utilitarian and other rival justifications of punishment.

My aim in the present paper is to challenge this presumption. I argue that retributivism is a social practice that conflicts with rather than embodies the considered judgments that the proper recognition of moral autonomy would authorize. This critique of retributivism, then, differs from the utility-based critique in its acceptance of the Kantian moral theory. Rather than arguing that retributivism is excessively Kantian, I argue that it and, at times, Kant himself are, in a relevant sense, not Kantian enough.

There is little doubt that Kant himself believed retributivism to be compatible with his moral theory. Still, both the prominence of retributivism as a practice and the profound influence of Kantianism on contemporary moral philosophy suggest that the compatibility of its conceptual core with retributivist justifications of punishment is significant quite apart from whether, as a matter of historical record, Kant himself arrived at the conclusions urged in this essay. This understanding defines both the purpose of the arguments advanced below and the sense in which they could be thought of as “properly Kantian.”

Since the basis of my case against retributivism lies in the connection between retributivist justifications and background moral arguments, I begin by locating the conceptual support for retributivism on the map of Kantian moral theory. My sketch of this map is focused on the notion of freedom, which exercises the dominant conceptual pull in that theory and yields normatively trumping criteria of moral justifiability. The section that follows this initial investigation considers an issue that should be viewed as central to justificatory theories that rely on the Kantian account of freedom: whether punishment should be understood as a restriction on a criminal’s freedom or as an expression of it. I argue that the latter view faces serious
difficulties that have not been, and are not likely to be, resolved by retributivist justifications that draw their support from Kantian moral theory. My strategy in this section is, thus, pragmatic: I conclude that punishment restricts the freedom (autonomy) of the criminal because the strongest available arguments to the contrary must be rejected as unconvincing. In the next, and final substantive section, the preceding conclusions motivate the consideration of explicit tests of the moral justifiability of punishment.

II. Kantian Moral Argument and Justifications of Punishment

The concept of freedom, dually conceived in its “negative” and “positive” guises, forms a cornerstone of Kant’s ethical theory as developed in The Groundwork and in The Metaphysics of Morals. Negatively construed, freedom is the inalienable ability of the will “to work independently of determination by alien causes” (Kant 1964, p. 114), including both internal desires (“inclinations”) and external coercive forces. In Kant’s positive conception of freedom as autonomy, the will’s ability to be independent in this fashion is transformed into its property of “being a law to itself” (Kant 1964, p. 108) – of arriving at maxims of behavior solely through the exercise of reason itself. Positive freedom/autonomy is, thus, the capacity to set ends, to commit oneself to moral principles and to a life plan, and to adhere to these even in the face of self-interest or attractive short-term ends.

Although there remain a number of ambiguities in Kant’s account of the relationship between freedom and universal moral law, it is clear that in that account, as well as in the accounts of sympathetic commentators seeking to reconcile these critical features of Kantian moral theory, freedom serves both as (1) a condition of possibility and (2) a key normative objective of binding moral law. Correspondingly, human beings are viewed both as capable of
moral actions, because freedom is the most critical descriptive feature distinguishing them as beings capable of practical reason, and as subject to the normative force of the universal moral law derived from the aspiration of freedom as autonomy. The self-conscious capacity for autonomous choice implies the uniquely human condition of being always “an end in itself.” It alone can ground a possible categorical imperative, since “relative” ends (i.e., ends that derive their value exclusively from “their relation to special characteristics in the subject’s power of appetition”) “can provide no universal principles, no principles valid for all rational beings and also for every volition – that is, no practical laws” (Kant 1964, p. 95). Thus we arrive at the Categorical Imperative, which enjoins us to treat humanity “never simply as a means, but always at the same time as an end” (Kant 1964, p. 96) – i.e., as free, self-directed, and ultimately autonomous.

The primary concern of Kant’s and of Kantian ethics and politics is the individual choice of ends that embodies this ideal of autonomy. But while ethics is the domain in which autonomous choices may be promoted directly, the political sphere – the domain of the state – cannot have such an expansive mandate. Its goal, in the “doctrine of right” Kant puts forth in the *Metaphysics of Morals* is, effectively, to “mimic” ethics in the political world. But the just state is powerless to enforce autonomy given our limited knowledge of each others’ intentions and the general irrelevance of coercion for a genuine choice of one’s ends. Nonetheless, a just state can promote autonomy indirectly by, inter alia, maximizing the possibility of autonomous choices. It may do so by maximizing the possibility of free choices that are consistent with like choices by others, since autonomous choices would be a subset of those. Maximizing the possibility of free choices entails measures that need not require the use of the state’s coercive powers – for example, when the state coordinates citizens’ expectations or adopts social policies that aim to promote respect for human dignity. But insofar as a just state uses tools like penal practice, it
resorts to measures that are unequivocally coercive, and herein lies a *prima facie* puzzle: if freedom is at once the empirical basis and the object of normativity, how can a restriction on it (in particular, a restriction by means of positive legislation) ever be normatively defensible?

Kant’s answer to this question is referred to below as The Test of Justifiable Coercion (TJC): “...[I]f a certain use of freedom is itself a hindrance to freedom in accordance with universal laws (i.e., [juridical and ethical] wrong), coercion that is opposed to this (as a hindering of a hindrance to freedom) is consistent with freedom in accordance with universal laws, that is, it is right” (Kant 1996, p. 25). The moral defensibility of the “hindering of a hindrance to freedom” as invoked in TJC goes to the heart of the relationship between the state and morality in Kant’s, and more generally in Kantian, moral and political philosophy. Actions that violate other citizens’ exercise of freedom do not necessarily impinge on their individual autonomy, but when that exercise is universalizable (consistent with the like exercise by others), those actions violate the necessary condition for individual autonomy. Insofar as that is the case, they fail the universalizability test themselves, and cannot be properly autonomous. In a contrasting case, individuals may choose actions that are consistent with what behavior in accordance with autonomous choices would require but do so under threat or another kind of coercion. Although consistent with the observable (external) behavioral demands of autonomy, such actions would not, of course, be autonomous, but they are precisely the sort of actions that are the object of a just state under the doctrine of right. “A hindering of a hindrance to freedom” is an action that *enables* (though, of course, it does not *ensure*) the autonomy of the one whose freedom was hindered by the illegal action, and it does not, in and of itself, necessarily violate the autonomy of the one who was doing the hindering or even his or her freedom under the universal laws. In this sense, “a hindering of a hindrance to freedom” advances freedom in accordance with universal laws, and has at least the potential to be a pure gain with respect to autonomy as well.
TJC re-asserts the centrality of considerations of freedom in determining the content of moral action and identifies a sense in which a coercive action may be consistent with freedom under the universal laws. Still, these implications do not settle the question of whether satisfying TJC is sufficient to justify retributive penal practice in the face of the architectonic implications of freedom as the normative object of moral law. The sufficiency of such a test surely depends on the possibility that a coercive measure “opposed” to the criminal action may, in principle, be consistent with freedom and on the congruence of the justification of punishment with the conception of the responsible wrongdoer. But it also depends on the extent to which the moral-theoretic basis for that justification can be consistent with the preference for a distinctly retributive measure over alternative practices that may also be invoked in promoting the ideal of a just state but that may be no less, and possibly more, consistent with freedom.

Consider, in particular, the relationship between the retributive penal practice and the practice of what is referred to below as (coercive) preemption. The relevant distinction here is between, on one hand, negative sanctions whose gravity is determined primarily, if not exclusively, by the gravity of the crimes and whose immediate purpose is to extract a proportional repayment from the criminal, and, on the other hand, measures that are driven by assessments of the likelihood of the agent’s repeat offence. As such, the latter may range from the deprivation of freedom to enforced participation in remedial moral instruction to responses that may entail no coercion at all. Given that the point of retributivism is that the punitive measure is desirable over and above its preemptive effect on the criminal, what might be thought of as the strong version of retributivism (or retributivism proper) concerns precisely the endorsement of the former kind of measures in spite of the availability of the latter. To fix concepts, then, let us initially reserve the term “punishment” for the retributive measure alone, and let us suppose that there are practical differences between the retributive and the preemptive
measures and that both of them are available to the society. (These assumptions are relaxed later in the paper, where I consider the possibility of weaker versions of retributivism, which may endorse retributive measures as the feasible second-best.) The point of departure for the main arguments of this paper is, then, the following question: if freedom is the normative object of moral law, can a practice of (retributive) punishment be morally sustained given the possibility of less punitive preemptive measures?

A key element in the Kantian attempts to supply a positive answer turns on Hart’s (1968) distinction between “the general justifying aim of punishment” and a justification’s “principle of distribution.” As a recent sympathetic account of Kant’s retributivist view puts it, “for Kant... punishment itself has a larger purpose – the demonstration of due respect for free, equal and autonomous agents” (Holtman 1997a, p. 6). Another interpretation echoes this point: “the law generally, and particularly the criminal law, as an instrument of social power, is meant to maintain a condition of the greatest freedom possible for each individual to the extent that that freedom may coexist with like freedom for all” (Byrd 1989, p. 198).

In these accounts, retributivism proper may be best thought of as an implication of the arguments about the distribution of punishment (e.g., Holtman 1997a; Byrd 1989; Murphy 1987; Scheid 1983). The essence of this view is that the Kantian categorical treatment of morality disallows the kind of case-by-case judgment about punishment that turns on considerations other than the criminal act itself to determine the appropriate social response to that act. In particular, considerations that are agent-specific, or that invoke the effects of punishment on the probability of committing other crimes or on the general welfare of members of the society, etc., would contradict ‘a priori based law’ and involve us in the treatment of human beings as means only, directly violating the command of the moral law (e.g., Mundle 1954; Hart 1968; Byrd 1989).
The argument, then, proceeds something like the following: “In order to avoid this charge [of treating human beings merely as means], we must show that there is something about the offender, apart from the consequences of punishing him, which justifies us in canceling his right not to be made to suffer. Retribution, insofar as it stresses desert, fills this void... The crucial point... is that a morally tolerable account of punishment must include the idea that the offender, in virtue of committing an offence, deserves to be punished” (Burgh 1982, p. 195, emph. added).

These arguments for retributivism, then, are advanced precisely on the grounds of retribution’s compatibility with respect for autonomy. But they stop short of proving the necessity of retributivism on the grounds of autonomy: retributivism may satisfy an important moral requirement, but there is no prima facie reason to believe that it is the only approach to dealing with criminal action that would. In particular, by the logic of these arguments, punishing everyone, irrespective of whether they committed a crime – with the punishment “grim-triggered” by a criminal act – satisfies the same moral requirement, too. But, more importantly, so does not punishing anyone at all. Further, insofar as a (coercive) preemption that stops short of punishment prevents criminals from taking actions that will be at odds with their freedom under the universal laws (a precondition for autonomous actions), the state adopting such measures may reasonably be construed as acting to minimize the criminals’ non-compliance with the external behavioral aspects of what would be their autonomous choices – that is, to minimize the deviation from the choices that criminals would make if they were choosing autonomously. (In fact, as I argue below, the system of preemption may be particularly attractive in actually enabling autonomous choices by the criminals.) It would seem wrong, then, to say that a system of preemption treats criminals as means only.8
What a Kantian moral defense of retributivism appears to require is not just an explanation of why the assignments of punishment are to be implemented in a particular way, but an argument that retributive punishment scores at least as high with respect to the furtherance of the “general justifying aim” as any other practice, including non-punishment. Since autonomy is a (if not the) defining characteristic of personhood, and since criminals cannot lose it by committing a crime, they, like all other persons, are entitled to be treated as subjects who retain their normative status as ends in themselves and so as claimants on the general justifying aim. A Kantian retributivist must argue that retributivism respects human freedom, including the freedom of criminals – understood, of course, in its positive sense – at least as much as, if not more than, other systematic responses to crime that are consistent with universal laws.

This realization provides the analytical starting point for the arguments developed below. Its immediate implication is that a punishment that is at odds with moral considerations that center on the freedom of all involved individuals should be considered prima facie morally indefensible. I argue that, properly construed and against the background of the supporting moral theory, this point severely undermines retributivism as a morally defensible framework of punishment.

The development of this thesis hinges on making the case that punishment is a curtailment of a criminal’s freedom properly conceived. The opposite view – that punishment is, instead, an expression of a criminal’s freedom – is often implied but, surprisingly, rarely supported by a sustained argument, despite some evidence of the recognition of the importance of this thesis for both the substance and the rhetoric of retributivism. As the foregoing discussion underscores, the consideration of the compatibility of criminal punishment with the freedom of the criminal is of paramount importance for determining the possibility of a morally
consistent Kantian retributivism. If such compatibility could be demonstrated, punishment would become less a matter of tragic moral choices and retributivism more an interpretive idea than a moral dilemma. Still, if punishment is, after all, a restriction of freedom (in particular, of freedom as autonomy), then the moral-theoretic centrality of freedom makes justification of its curtailments much more contingent and nuanced.

III. Is Punishment a Restriction of Criminal’s Freedom?

The arguments that punishment is not a curtailment of freedom vary across justificatory approaches. Fundamentally, though, these arguments rely on a conception of freedom that corresponds broadly to the Kantian notion of freely-willed autonomous choice, and draw their conclusion from the claim that the criminal freely wills her punishment. The central importance of the criminal’s being “an end” makes establishing that claim a moral prerequisite for the compliance of the practice of punishment with moral law: “as rational beings, they [human beings] ought always at the same time to be rated as ends – that is, only as beings who must themselves be able to share in the end of the same action” (Kant 1964, p. 97).

The requirement that criminals be able to share in the end of punishment, and, correspondingly, that punishment be willed freely by the criminal, must, however, be understood correctly. Its substance is not that any individual will is a constraint, but rather that only that will which is itself universalizable is: “every will... is restricted to the condition... that it be subjected to no purpose which is not possible by a law which could arise from the will of the passive subject itself” (Kant 1996, p. 108). Requiring that the criminal will her punishment freely, thereby acting as a rational co-legislator in the kingdom of ends and not as a criminal, ensures that the relevant decision is made by the criminal not on behalf of her will qua criminal but rather on behalf of her will qua co-legislator. The fact that the appropriate criterion for this
decision (universal moral law) is public and known makes the criminal’s decision itself public and open to criticism, preventing the possibility of her abuse of first-person authority.¹¹

I examine this strategy of justification in the remainder of this section by considering two distinct types of arguments, appealing, respectively, to revealed preference and to moral obligation.

**Revealed Preference Arguments**

According to the revealed preference argument, the criminal’s preference for punishment is an implied content of her will when she wills a crime. The preference for punishment is, in effect, revealed through the criminal act itself.

In one version of this argument, the crime is the revelation of a preference for the appropriate punishment because that punishment is simply the application of the universalization test of the Categorical Imperative to the criminal herself.¹² By this logic, for example, in willing a murder, a criminal wills a universalization of her maxim, and hence a capital punishment for herself. However, the requirement of willing punishment as an autonomous and rational person may be regarded as an *eo ipso* disqualification of this version of the argument. This is so because the requirement of willing punishment cannot be understood to entail the assertion that in willing a crime, the criminal is acting morally, for in that case the Categorical Imperative test would imply the existence of a contradiction. If the criminal is not acting morally, it is difficult to see how her decision¹³ could have normative moral force in convincing her that she really does prefer punishment or in convincing the society responding to her action that she prefers it, potentially despite her denials.
This criticism does not extend to the second version of the revealed preference argument, which holds that, by virtue of human beings’ status as autonomous dignity-bearing choosers, their criminal choices must be interpreted as expressions of their rational preference for the associated punishment. The justification offered for this view avoids the criticism raised against the first version of the argument by relying on a conception of preference derived exclusively from the a priori status of human beings, rather than from their (criminal) choices. It does so by appealing to the important element of our moral intuition invoked in the claim that persons have a “right to punishment.”

The argument proceeds as follows. A contractarian or mutual-advantage picture of social coordination is, first, used both to substantiate the assertion of individual preference sovereignty and to assure acceptance of a penal system, so that non-punishment appears to be explicitly rejected in the presence of punishment as a contractarian support. The social responses to crime could, then, arguably be captured by two options: reducing the criminal act to an unchosen individual characteristic – a kind of “illness” (calling, in the extreme, for mental “therapy”) – and treating the criminal as a responsible chooser deserving of punishment. Given these choices, and given that punishment arguably better preserves the human dignity of the chooser, human beings who value their dignity must prefer it (Morris 1968). Punishment is, thus, not only willed, but willed in a morally relevant way.

This argument hinges on two critical premises. The first premise is that (retributive) punishment would be chosen as a necessary provision of social organization. I consider and cast doubts on the validity of this premise below. For now, however, consider the second, and arguably defining, premise of this version of the revealed preference account: that the criminal
must view her exception from punishment from the standpoint fixed by the dichotomy of social responses, punishment and “therapy.”

Although this version of the revealed preference argument is congenial to the Kantian preoccupation with respect for universal human dignity – explicitly including in this case the dignity of the criminal – its force is distinct from that of the standard Kantian proposition that one denies one’s own dignity by acting immorally. Unlike that proposition, the revealed preference argument sees the criminal as denying her dignity “directly” – through being subjected to a practice (“therapy”) that literally detracts from it. Consider, however, a criminal who is interested in evading punishment, effectively free-riding on social organization. (Because, by supposition, there exists an effective system of feasible preemption, “free-riding on social organization” here does not, at least prima facie, fail the universalization test.) If the evasion is possible, then in refusing punishment the criminal may not need to subject herself to dignity-denying “therapy.” If the moral justifiability of punishment is meant to be a function of the existence of the “preference for punishment,” then an attempt to evade, signaling the absence of this preference, must undermine the justifiability of punishment. As long as individuals cannot commit to non-evasion, either in or via the first step of the argument, the problem of the moral justifiability of punishment is not resolved.

The basic thrust of the moral argument against evasion is clear: if the first step of the revealed preference inference is, in fact, a kind of social contract and that social contract calls for (retributive) punishment in the event of a criminal violation, then attempting to evade a just punishment would be failing to act as a co-legislator in the kingdom of ends. Note, though, that if such an argument is to have bite, then its implication for the criminal freely willing her punishment will, for a Kantian, be effectively independent of the revealed preference inference.
She is under a moral obligation to will her punishment because, as a free-willing coauthor of the social contract, she has associated crime with punishment – regardless of whether the criminal act itself is understood to express any kind of preference. In short, to the extent that the revealed preference argument works, it appears to do so because of an argument from moral obligation – an argument which, if successful, would seem to be sufficient to establish the desired conclusion. I consider that argument next.

*Moral Obligation Arguments*

One could think of the argument that the criminal is under a moral obligation to will her punishment as consisting of two steps. In the first step, any rational individual acting autonomously assents to the association of punishment with a given crime from an appropriately specified standpoint. In the second step, that assent is invoked as normatively binding on the criminal. The criminal must accept retribution as a will of her rational self, and hence as her own free (autonomous) will. This two-step structure corresponds naturally to the so-called “reciprocity” or “fair play” justifications of retributivism (Rawls 1999/1964; Morris 1968; Murphy 1994/1970).\(^{15}\) Whether these justifications posit the existence of an implicit, actual, or hypothetical contract between an individual and her society, they capitalize on the obligating status of promise or of enjoyed mutual advantage. A criminal act violating these obligations creates a moral burden of repayment on the criminal, which is thought to be annulled (only) by punishment.

To frame the evaluation of this argument, note first that a key implicit condition underlying it is that the agency identifying the moral injunction against a particular criminal act is, *in effect*, distinct from the agency committing that act. This dualism is well-ingrained in
contemporary moral-theoretic discourse, and Kant himself implicitly authorizes it in relation to willing punishment in *The Metaphysics of Morals*: “as a co-legislator in dictating the penal law, I cannot possibly be the same person who, as a subject, is punished in accordance with the law; for as one who is punished, namely as a criminal, I cannot possibly have a voice in the legislation” (Kant 1996, p. 108). The dual co-legislator/criminal identity need not imply that the criminal was not acting freely in choosing the crime and so cannot be held responsible (cf. Wolff 1973, Singer 1983), but it does say something important about the status of the criminal’s perfectly rational autonomous agency. If morality demands that the act be designated as criminal, the perfectly rational moral agency whose judgment supports this designation would never choose to commit that act, and so the punishment as its price would never be autonomously chosen either.

Given that, upon committing a crime, the criminal does not lose her humanity (and hence, her spacio-temporally identical rational autonomous self, which was obviously “unrealized” in the criminal act, but is nonetheless present), the punishment for the criminal act is visited not only on the chooser of that act, but on the morally autonomous chooser as well. Since the latter could never have entered into the exchange of crime for punishment, the posterior assignment of punishment would, in effect, be holding her accountable for the actions of her immoral counterpart. Even if, from the standpoint of the integrity of the rule of law, the contractual view of the relationship between crime and punishment is reasonable and conforms to our basic intuition about what it means for a social practice to be fair, the moral force of fair play does not bind the rational proxy of the criminal to will her punishment as the previously agreed upon outcome of the broken promise. A truly punitive measure, such as imprisonment, will, inevitably, have the effect of reducing the choices available to the criminal agent as the autonomous chooser and so of undermining her “rights-sensitive autonomy”—her ability to choose an authentic morally admissible way of life (Meyer 1987, pp. 267-8). The force of
justifications based on a criminal’s willing punishment as a performance of a putative contractual obligation will, then, be undermined as well.\textsuperscript{18}

That the punishment of a (phenomenal) criminal imposes a limit on her ability to do (authentic) good, even as it also imposes a limit on her ability to do ill, has far-reaching consequences that are rarely recognized, yet as a proposition, it should not be controversial. Positing the conflicting identities of the criminal implies neither psychological abnormality nor indefensible metaphysics. It requires only a recognition that, to paraphrase Kant’s famous claim, “the timber of humanity” is, indeed, likely to be “crooked” – that human beings are more often than not torn between conflicting motivations, and that even criminal action does not signify a summary and final capitulation of moral autonomy to the immoral impulse. To deny the autonomous motivations of the criminal their own moral agency vis-à-vis the social contract, while accepting the contractarian defense of punishment as the satisfaction of the moral burden of repayment, is, in effect, to privilege the criminal’s immoral (and non-autonomous) motivations. This move should be particularly unappealing to a Kantian; indeed, the opposite ranking of motivations stands behind the normative force of the concept of freedom in establishing the moral relevance of the rights of the criminal in the determination of punishment. To speak of the relation of the morally autonomous agency of the criminal to punishment is, thus, less to reify a moral-theoretic fiction than to reaffirm the motivational complexity of actual human choice already implicit in the claim of inalienable autonomy.\textsuperscript{19}

It is instructive to see why this dual identity argument cannot be effectively met by adopting a strongly cognitivist view of society’s normative principles as the substantive content of the intersubjective rationality of its members. If criminal punishment is to be one such normative principle, rather than an inexplicable act of social violence, the criminal must
recognize the punishing group as her own “moral community” (Oldenquist 1988, p. 469). She must view the punishment “not merely as a negative feeling or abstract judgment but as the community’s [and so, by extension, her own] committed opposition” to the offence (Hill 2000, p. 198). The implied inference is that, since no rational being would choose to deny her own identity (here defined by such substantive communal membership), a rational member of society who finds herself in a situation with a socially prescribed punishment must necessarily will it.

To see that this is not a satisfactory way of getting around the dual agency argument, note that in appealing to the moral self of the criminal, the injunction to will the punishment here is still subjecting that moral agent to the punishment along with its criminal alter ego. Even if the moral ego finds itself in agreement with the motives of those with the “committed opposition” to the offence, does it really make sense to argue that it is thereby committing itself to the punishment? To be sure, one may argue that the strongly cognitivist interpretation somehow rules out the very possibility of an internal conflict interpretation of criminal action and so preempts the dual identity critique altogether. But that argument is a double-edged sword since the appropriate explanation for a disjunction between the criminal’s actions and the normative-rational language of her social milieu would then be that the society failed to establish the universal identification of a given behavior with criminality. But that would undermine the retributivist figure of a responsible wrongdoer – as someone who truly knows the difference between right and wrong and still chooses the wrong – and, with it, the fiction of her rational assent to punishment.

A Kantian argument need not, of course, rely on a contractarian structure. Indeed, one may argue that a criminal’s willing her punishment is a moral obligation because it alone captures the demands of universalizability imposed by the duties of moral legislation in the
kingdom of ends. As the purpose of this kind of argument would be to demonstrate that this act of willing is necessary, it must show that acting differently would be unethical. In particular, it must establish that not willing one’s punishment necessarily implies acting under a maxim that could not be willed to be a universal law because the practice of punishment is a necessary complement of a system of laws.

Consider, then, the proposition that the practice of punishment is such a complement. Since it does not refer to an intrinsic property of laws, demonstrating its validity is, in effect, equivalent to demonstrating that, as a practical matter, crimes cannot be stopped unless there are punishments. Unless this argument is narrowly circumscribed to apply only to a given agent’s own likely future crimes (already accounted for by the effectiveness of the system of preemption), it is a standard deterrence argument. Endorsing such an argument in the presence of effective preemption must, for a Kantian committed to treating individuals always also as ends, imply a performative contradiction. Indeed, the presence of preemptive measures must make that argument even less satisfying than the classic deterrence argument for punishment.

The dual identity argument articulated above suggests that one cannot avoid this conclusion by retreating to the claim that the threat of punishment is necessary to enforce laws—a justificatory strategy favored by some retributivist accounts (e.g., Goldman 1979; Quinn 1985). To be sure, it seems unequivocal that respect for human dignity and enhancement of the possibility of victims’ autonomy may be served by mechanisms that discourage criminal actions. As co-legislators in the kingdom of ends, we would recognize the value of such mechanisms and seek ways to take advantage of them. However, insofar as they call for attaching a credible threat of punishment (in the above, retributive, sense) to criminal actions, they give rise to a key conceptual difficulty. While it may be true that credible threats are sufficient, and punishment
itself may be just the credibility inducer, treating the threat and the practice as one presents a “moral time-inconsistency” problem. The decisions to threaten and to punish occur at different times, and once the crime has been committed, one is forced to consider the question of punishment on its own terms. Even if, by establishing the moral status of “fair warning” (e.g., Quinn 1985), the threat-based retributivist arguments were somehow to succeed in resolving the moral time-inconsistency for the individual qua singular agency, they would not therefore do so for the dual agency whose autonomous ego was, in the instant of the crime, overcome by its immoral counterpart. What argument can one give to that autonomous self? Given the feasibility of effective preemptive measures, it would seem to be one of the existence of some value of deterrence that goes beyond preempting repeat crimes. Once again, such a strategy must, for a Kantian, present a moral problem.

Relaxing the Assumption of Feasible Preemption

The central conclusion implied by the preceding arguments is that the strong version of the Kantian defense of retributivism cannot be sustained: the commission of the crime is not sufficient for punishment. At the least, the endorsement of retributivism must require that the maintained premise that there exists an effective system of preemptive measures must be false. As such, if the criminal must freely will her punishment, it is not only because she committed a crime, but also, and necessarily, because the social and technological progress that could bring about such a preemptive system has been too slow, or because the society chose not to invest into bringing it about, etc. This realization has the effect of depriving retributivism of a considerable part of its moral force, but might appear to leave room for a weak version of Kantian retributivism: the criminal must freely will her punishment because retributivism is the best we
can do for now, given the absence of an effective system of preemption. Can this weak version of Kantian retributivism be sustained?

Suppose, then, that because the effective system of preemption is not available, we must, to support a system of laws, resort to some practice of punishment. Given the dual agency construction, the claim we must evaluate is that the autonomous egos of criminals are, therefore, under a moral obligation to will punishment – that failing to will one’s punishment necessarily means acting under a maxim that could not be willed to be a universal law. With this formulation in mind, the dual agency argument points to a key tension in the fabric of the Kantian account. While it seems relatively easy to embrace the claim that, given the impossibility of successful preemption, it may be incumbent on the criminal’s autonomous self to accept the punishment for the sake of meeting the universalizability test, it seems much more difficult to maintain that punishment is not an infringement on the freedom of that autonomous self. In order for that to be true, it would have to be the case that the freedom of the autonomous self of the criminal is fully served by the punishment. Unless one accepts an account of that self that denies it the possibility of multiple moral pursuits, it is a claim that seems difficult to maintain.

Unless one is prepared to undertake much more serious revisions of the core features of Kantian moral theory, the conflict that this argument identifies is resolved by granting that punishment is a restriction of the criminal’s freedom that may, nonetheless, be sometimes justified. Coupled with the moral priority of the considerations of individual freedom, this admittedly pragmatic conclusion motivates a view of punishment as a fundamentally moral challenge to the justificatory enterprise. Consequently, the task of such an enterprise should be the development of tests of moral justifiability that help navigate moral conflicts that entail tragic trade-offs of values. It should not be the elaboration of constructions aiming to justify an abstract
notion of punishment as an expression of our moral self-understandings, which has been the preferred path for Kantian justifications of punishment. The remainder of this paper takes some preliminary steps in advancing the former approach.

IV. Ethically Justifiable Punishment?

From Kantian Retributivism to Kantian Prevention

Kant’s test of justifiable coercion (TJC) is a carefully worded criterion that, as indicated above, amounts to something substantially short of retributivism. A “hindering of a hindrance to freedom” may be a necessary condition for punishment, but not a sufficient one. Consider the conjunction of propositions asserted above:

(1) Considerations of freedom are trumping in ethical justifications of coercion.
(2) Respect for humanity is exceptionless.
(3) Punishment is an infringement on a criminal’s freedom.

If a legal sanction against a criminal amounted exclusively to a coerced cancellation of the criminal act without affecting her other possible actions, then such sanctions would be consistent with TJC as well as with (1)–(3). But, assuming that punishment is minimally continuous, it may satisfy TJC while failing to be minimally harmful with respect to the freedom of the criminal, and so imply a contradiction with at least one of the premises (1)–(3). Given (2) and (3), such a punishment would constitute bad faith in relation to the “always at the same time as an end” formulation of the Categorical Imperative. The effect of asserting the conjunction of premises (1)–(3) is the substantial undercutting of the normative force that could be exercised on behalf of the moral asymmetry between the criminal and the victim.

This conclusion intuitively suggests the value of the following further test of ethical justifiability:
The Test of Punishment (TP): *an increase in the coerciveness (severity) of punishment is justified only if it leads to an increase in the freedom of action of the prospective victim*.21

Curtailments of freedom that are consistent with both TJC and TP have a distinctly preventative flavor. Given the injunction that human beings are to be treated “never simply as a means, but always at the same time as an end,” enforcing such measures must have among its purposes prevention of an(other) offence by the same individual. The underlying preventative purpose of punishment is, thus, fundamentally distinct from (general) deterrence.

But the appeal of TP should not be overstated. While it offers some guidance as to what legal sanction cannot be justified, it is less helpful when, as in the paradigmatic case of punishment, a “hindering of a hindrance to freedom” amounts to a comprehensive rather than incremental constraint on the criminal. In such cases, TP provides little guidance as to what sanction can be justified, and even when it is violated, there may be reasons for moral defensibility if, given feasibility constraints, the sanctions in question are required as part of a background mechanism of prevention in a world of equal freedom.

If the arguments in this paper are correct, such cases entail not only conflicts between the autonomy of the actual or prospective victims, which may be best served by maximizing the negative sanction, and the *welfare* of the criminals, which would suffer as a consequence of such a sanction (cf. Sorell 1999), but also conflicts between the autonomy of the victims and the autonomy of the criminals. Addressing these conflicts requires that we make compromises between different violations of freedom – however unappealing to a traditional Kantian the prospect of making them may be. The injunction to treat humanity always also as an end appears to create a moral asymmetry between the criminal and her possible victim(s), ruling out the possibility of sacrificing the freedom of the victim for that of the criminal. However, if freedom
is a normative objective of moral law, this asymmetry stops short of offering a fully satisfactory moral consolation. The moral sentiment that befits the assignment of punishment seems to be regret at the sacrifice of value, rather than satisfaction at the accomplishment of moral rectitude.

The normative weight of this conclusion is to encourage a re-assessment of the proper role of the state in Kantian accounts of punishment. If the argument about the loss of freedom in punishment is correct, the historic influence of retributivism on such accounts has obscured the import that Kantian moral theory should place on punishment-minimizing justificatory solutions. More broadly, it has also de-emphasized the advocacy of social reforms targeting the causes of criminal behavior. To the extent that such reforms may diminish the probability of criminal offences (even at a substantial social cost), they must enter into the moral accounting of the justifiability of the state response to criminal action.

V. Conclusion

Kantian moral theory has undergone considerable change in the last century, disposing of indefensible metaphysical commitments and rigoristic doctrines that belie considered moral intuition, while retaining the core normative implications of moral autonomy. As the arguments presented above suggest, Kantian retributivism is a plausible target of this evolution as well. While the pitch of Kant’s rhetoric in scorning attempts to lighten the burden of punishment on criminals and the resoluteness with which he insisted on the need for equivalence between the crime and its punishment have become mainstays of the retributivist canon, neither of these rhetorical points has an uncontestable basis in Kantian moral theory. This fact should make jettisoning retributivism less damaging theoretically. At the same time, the conclusion that retributivism is unappealing because it ignores tradeoffs between considerations of autonomy focusing on the (possible) victims and those focusing on the criminals suggests strongly that the
caricature of Kantian moral theory as a theory in which “all good things must always go together” is wrong as a claim about the implications of fundamental ideas. Insofar as that caricature is, nonetheless, fair as a comment on the practice of applied Kantian moral theory, the re-assessment of that theory that the above arguments implicitly urge may, indeed, be quite substantial.

The inference that those arguments amount to a rejection of retributivism should, of course, be tempered by the fact that they are internal to Kant’s moral theory. As with all such arguments, their immediate presumptive target is the internal coherence of a theoretical construct, rather than the absolute and independent (im)plausibility of its constitutive elements. This means, among other things, that the strict implication of this paper’s arguments is not that retributivism is unjustifiable, but that, to the extent that one shares the basic commitments of the Kantian moral outlook, retributivism in punishment should not be a natural creed.

Where does this leave a Kantian? One normative consequence of the above arguments is the importance of doing something that would seem to be both unnecessary and repugnant to a committed retributivist: seeking options that could replace or mitigate punishment. A closely related implication is the importance of analyzing what the background system of prevention in a world of equal freedom would require. The features of such a system are unavoidably contingent on both technical information and aspects of particular criminal acts, and this contingency does not, in and of itself, violate the moral a priorism that protects against treating human beings as means only. The key to complying with it is in determining which contingent information should be interpreted as a part of the description of the relevant moral choices and which a part of the description of the identity of the chooser. While it would judge the variation in the latter to be morally irrelevant, universalization need not preclude sensitivity to the details of moral choices.
The determination of which aspect of the situation is properly seen as one type of descriptor rather than the other is itself a moral choice, and the requirement of moral a priorism does not free us of the moral responsibilities of making it. The task for a Kantian theory of punishment that is willing to give up retributivism is to make those choices.
Works Cited


Notes

1 Cottingham (1979) describes no less than nine “varieties of retributivism,” and his list is, surely, far from complete to date.

2 Both the *Oxford English Dictionary* and Cottingham’s taxonomy of usage in justificatory theories of punishment (1979) identify this to be “the basic or fundamental sense of ‘retribution’.”

3 In one of his most strongly worded statements about their relationship, Kant (1949, p. 140) says that they “reciprocally imply each other.”

4 See Hill (1992) for an interpretation of the notion of an “end in itself” that emphasizes its dependence on the capacity for a self-chosen autonomous will.

5 Taken to the extreme, this is tantamount to saying that desert does not *automatically* imply a moral permission, let alone a moral duty, to punish. (Cf. Dolinko 1991). Indeed, as I argue below, even the assertion of the freedom-consistency of a punishment that prevents the commission of a crime by the punished subject is not unproblematic.

6 This practice of (possibly coercive) preemption bears close resemblance to what Sorell (1999) calls “paternalistic coercion” – a coercive practice to which a state might resort to bring citizens’ behavior in line with the behavior they would engage in had they been autonomous – including, in particular, behavior that is consistent with the laws of a just state. A key assumption that underlies both of them is the error theory of choice that Sorell argues “provides the background in Kant for the pursuit of autonomy” (1999, p. 27) but is often missed by Kant’s modern interpreters. The error theory provides a natural background for the claim, which figures prominently in the discussion below, that a criminal does not lose his or her humanity, and so his or her claim to autonomy, upon committing a crime.
Murphy (1987) provides evidence of considerable revisions and ambiguities in Kant’s own statements on punishment from his earlier and later works.

This echoes Sorell’s (1999, pp. 12-14) argument that “paternalistic coercion,” e.g., in taking cigarettes away from a smoker, is consistent with Kant’s Principle of Humanity because the smoker could reasonably share in the end of that action.

Among many points of textual support for this claim, see Kant’s characteristic argument about the inappropriateness of rape as a punishment for a rapist, because that would effectively imply the lack of respect for the rapist’s humanity (Kant 1965, p. 333). This is so because, as Kant insists in the *Groundwork*, freedom is a property of all rational beings “endowed with a will” (Kant 1964, p. 115). If it could be shown that the rapist were incapable of will (i.e., incapable of practical reason), respect for humanity would cease to be a constraint on his punishment. Of the more recent Kantian treatments of punishment, Morris (1968) and Holtman (1997a and b) develop arguments emphasizing the importance of punitive practices’ respecting the personhood of criminals.

See the explicit statements by Kant (1996, p. 25). Morris (1968, p. 497) maintains that a criminal, in choosing a crime, necessarily wills her punishment; withholding it, therefore, constitutes a denial of her freedom and of her claim to a rational nature. Similarly, Murphy (1994, p. 121) argues that with regard to punishment, “the criminal himself has no complaint, because he has rationally willed or consented to his own punishment.”

Cf. Murphy (1994).

Fleischacker (1992) presents the most sophisticated recent defense of this claim, based in part on his explicit rejection of the freely willed punishment criterion. As I argue above, however,
that criterion is called for both as the most consistent reading of the exegetical evidence and as an inference of the overall analytical framework of Kant’s moral theory.

13 Rather than, say, its consequences, which are strictly outside the purview of this argument.

14 Since moral law is the ratio cognoscendi of autonomy that is equivalent to human dignity (Kant 1949, p. 119fn).

15 The criticisms presented below could be seen as applying also to Nino’s (1983) “consensual theory of punishment,” which shares some of its features with the revealed preference argument as well.

16 Of course, “dualism” is a radical simplification, since there is likely to be a continuum of morally more and less acceptable motivations, but it captures the sense of multiplicity relevant for the discussion that follows.

17 Even if it can bind it to will a carefully circumscribed restitution. It bears emphasizing that the argument considered here is not whether the contractarian argument is, in principle, morally binding, but whether it is morally binding just because it is contractarian.

18 Recall that the maintained assumption is that the alternatives to punishment include less punitive preemptive measures, not merely therapy. This makes implausible the claim that the rational moral self may be under an obligation to will punishment as a way of constraining the immoral self: in the presence of effective preemptive measures, there would, presumably, be no criminal actions left to constrain.

19 The implications of the distinct conceptualizations of the “split-level self” for the meaningfulness of the notion of autonomy are the subject of a substantial volume of articles, catalyzed, in particular, by Frankfurt (1971). Although controversies on the precise nature of the division of self and the relationship between its components persist, there seems to be a general
agreement that a robust notion of autonomy requires the recognition of some such division (Christman 1988).

20 Deigh (1984, p. 197) makes a parallel point in his criticism of Morris’ argument: “To put the problem succinctly, to hold, as Morris does, that pardoning a criminal against his wishes constitutes a wrong done to him is to deny that the authority of those responsible for meting out the punishment includes discretion to forgo the sanction. And this should strike us as rather odd.”

21 The necessity of punitive measures is, of course, deduced not from the Categorical Imperative alone, but from its conjunction with (a) empirical facts about the actual intentions, moral psychology, and physical capability of the criminal; and (b) the extent of the deprivation of freedom required by the mechanism that is expected to prevent the crime. I explore some of these factors below.

22 In particular, the present argument neither addresses nor implies a criticism of the distinctly non-Kantian, naturalist versions of retributivism, such as that of Moore (1987), or of the theories like Mabbott’s (1939), which make no explicit connections between retributivism and morality.

23 In T.M. Scanlon’s (1998, p. 199) felicitous phrase, “even the most familiar of moral principles are not rules which can be easily applied without appeals to judgment.”