1. INTRODUCTION

The "classic debate"\textsuperscript{1} between retributive and utilitarian theories of punishment has dominated discussions of punishment since Jeremy Bentham's time (if not Plato's). "Intractable" may best characterize that debate. In this paper I will argue that hope of resolving the impasse between utilitarians and retributivists is to be found in an understanding of law as a system of rules and principles for regulating human behavior so as to achieve and maintain the conditions of basic trust in a community.

Though I cannot fully argue for this conception of law here, a few introductory words are necessary. The theory of law which is here presupposed is a functionalist one: the purpose of law is to secure the conditions of basic trust in a community. Though functionalist accounts of law have most often been associated with natural law theories,\textsuperscript{2} it is possible to speak of law's function within a positivist framework.\textsuperscript{3} I mean to understand functionalism as compatible with positivism. The result is a conditional claim: if the principal function of a legal system is to create and maintain conditions of trust in a community, then those who violate the law must be punished. Any legal system which is understood as performing this function must adopt a retributivist approach to punishment. I shall assume, moreover, that the legal systems of western pluralistic democracies meet this functionalist assumption. Since it is a contingent matter whether a legal system performs any particular function for legal

\begin{thebibliography}{9}
\end{thebibliography}

positivists, we must recognize that those which serve some other purpose may require a different practice of punishment which is not essentially retributivist. If the function of law was to secure only coordination, for example, then reparations rather than punishment might be a more appropriate response to those who violate the law. "Punishment" within certain aboriginal communities may serve as an example.

The main task of this paper is not to defend this conception of law; that must wait. Rather, it is to argue that this conception of a legal system entails retributivism. Insofar as the purpose here attributed to law is a legitimate one, moreover, such an account must justify a retributive theory of punishment.

2. PUNISHMENT AND RETRIBUTION

This paper is offered as a defense of retributivism. Before the details of that defense can be developed, though, some background work is required. First, we need a definition of punishment, since it is the practice of punishment that requires justification. Let us say for the purposes of this discussion that punishment is an avoidable loss intentionally imposed by legal authorities upon an offender for an offense. The offense, upon which such a loss is conditioned, is a violation of a legal rule or principle. The determination that such an offense has occurred must, furthermore, be established by some public process of inquiry.

To say that offenders are punished for violating legal rules or principles does not adequately explicate the nature of an offense, however, and providing such an explication is the second piece of background work to be done. We need to know what kind of wrong an offense involves, which justifies the imposition of punishment on those who do such wrongs. Actual harm produced by the offense, though a typical element of a criminal offense, is neither necessary nor sufficient, of course, to characterize the wrongness of criminal behaviour. Some mere attempts, which produce no material harm, are nonetheless rightly subject to punishment; and sometimes considerable harm is done, but through accident or mistake, and so is

4 I am grateful to Nathan Brett for helpful discussions about the nature of punishment.
not properly punishable. The wrong which legal offenses involve, and which justifies punishment cannot, then, be equated with harm done.

Harm done is not the only explication of legal wrong available, and others will be considered below. In this paper I want to argue that the violation of conditions of basic trust in a community is the characteristic wrong which criminal behaviour involves, and it is this upon which the justification of punishment rests. I shall elaborate upon this in what follows.

Before proceeding to that central task, however, we also require an account of retributivism. This is not so easy to provide. John Cottingham claimed in 1979 to have identified nine distinct varieties of retributive theories. That number would have to be increased to account for more recent developments in the philosophy of law and punishment theory. On the traditional conceptions that Cottingham looks at, what unites retributivists at the most general level is the belief that offenders deserve punishment. They deserve punishment, moreover, because they have behaved culpably in committing an offense; their culpability justifies the imposition of a loss upon them, moreover, independently of any good consequences that punishing may bring about (be they deterrence, reform, education, or what have you). Because the harm characteristic of an offense is necessary for the justification of punishment, retributivist theories are often characterized as backward-looking, for they look back to the past crime rather than ahead to any possible good consequences of punishing in order to determine if punishment is justified. These characterizations imply that a past offense is not only necessary but also sufficient for punishment to be justified; in fact, the latter is not a step which a retributivist need take (as will be discussed shortly when we compare obligatory with permissive versions of retributivism). Yet the central intuition of all retributivists is that offenders deserve punishment because, and perhaps just because, they have committed a past offense.

The notion of desert which is central to traditional retributivist theories has been terrifically difficult to defend, however, and all desert-based retributivists face the problem of attempting to move

---

from the claim that, because offenders deserve punishment, it is somehow permissible or required that the state ensure that they get what they deserve. Rather than pursuing the line of defense based in considerations of desert, I propose to follow instead the path cleared by Mark Michael. The traditional ways of characterizing retributivist theories, and distinguishing them from utilitarian justifications of punishment, have been challenged recently by Michael. He argues that utilitarian theories ought to be identified as such insofar as they hold that punishment is justified just in case punishing causes persons to act in morally desirable ways. The connection between punishment and its good effects is considered to be both causal and contingent. A retributivist, on the other hand, sees the good of punishment as neither contingently nor causally related to the punishment. Rather, the good that punishment brings about is internally and necessarily connected to punishment itself. Michael characterizes retributivism this way: “A retributive theory will claim that the relation which holds between punishment and the event or state of affairs which justifies it is what is known as the ‘by’ relation.”6 On Michael’s understanding of retributivism, the punishment and its justifying event/state of affairs begin simultaneously and are non-causally related.

Some examples of recently defended retributivist theories may help to make this “by” relation clearer. For Herbert Morris, Jeffrie Murphy and Michael Davis, for example, we restore a certain equilibrium between benefits and burdens by punishment; for Jean Hampton, we annul (falsify) the claim of superior value made by criminals by punishing them; on my view, we restore the conditions of trust in the community by punishing.7 What event/state of affairs, to which punishment is internally related and which has sufficient value to justify punishment, will depend upon how one characterizes

the wrong which crime involves: for Morris, Murphy and Davis, legal offenses necessarily involve the taking of unfair advantage by offenders over those who obey the law. The fair distribution of benefits and burdens which crime upsets is restored by punishment. For Hampton, the wrong that crime involves is the assertion by the criminal, through his acts, of his superior value over his victim. We negate that false claim of superiority on the part of the criminal, and reaffirm the equal value of his victim, by imposing a defeat upon the criminal in the form of punishment. For me, the harm of legal offenses is to be found in their violation of the conditions of basic trust in society, and we restore trust by punishing the offender.

I think Michael’s account of the distinction between retributive and utilitarian theories of punishment is a useful interjection in the current debate. For the purposes of this paper, I will accept Michael’s reformulation of retributive theories, and argue that my theory counts as retributivist so understood. In a book-length development of my view, I would also want to argue that it can be understood as the view that criminals deserve punishment, and so it is retributivist in the traditional sense as well; I can do no more than offer this as a promissory note here.

Another distinction within the literature on punishment is worth taking note of before we proceed. It is important to distinguish between what has been called “negative” and “positive” retributive theories. Negative retributivism is the claim that it is a necessary condition of justified punishment that it be imposed only on the guilty. That is, negative retributivists insist that the commission of an offense is necessary for the imposition of punishment to be justified. Positive retributivism is the view that an offense is sufficient to justify punishment. There are two types of positive retributive theories. Weak (or permissive) positive retributivism holds that it is

---

8 This is not to deny that this description of retributive theories generates some classificatory difficulties. R.A. Duff, for example, has developed a theory of punishment which holds that punishment is justified just in case it aims at the moral reform of the criminal. The goal of moral reform is thought to be internally, non-contingently and non-causally related to punishment, moreover, and so Duff’s account might qualify as a retributive theory on Michael’s model; yet Duff insists that his is not a retributivist theory. Cf. R.A. Duff, Trials and Punishments (Cambridge University Press, 1986).

permissible to punish offenders, just in virtue of their having committed an offense. One could, in defending permissive retributivism, insist that whether one should actually punish is conditional on further contingent considerations of a consequentialist kind. Strong (or obligatory) positive retributivism holds that it is obligatory to punish offenders, just in virtue of their guilt. I shall defend strong positive retributivism. Legal wrong-doing requires punishment.

Now that we have some notion of what retributivism is, as well as the definition of punishment that I propose to use, the reader may be thinking that my definition of punishment begs an important question: whether only the guilty can be punished. Since one of the principal objections to utilitarian theories of punishment is that they allow, or even require, those innocent of a legal offense to be punished (even if only under rather bizarre circumstances), a definition of punishment which implies that only the legally guilty can be punished seems to answer what is a substantive moral objection to utilitarian theories by definitional fiat. This might be the kind of "definitional stop" that H.L.A. Hart warned philosophers of law against many years ago.10

No doubt stipulative definitions can be abused. And it is true that my definition of punishment has the implication that only those guilty of a legal offense can be punished. It seems to be unproblematic that it does so, however. For it is simply an analytic truth on any adequate definition of punishment that legal punishment is for a legal offense. The innocent may be "victimized" by the penal system, but they cannot be "punished".11 Moreover, it is real punishment – the infliction of a loss upon an offender for an offense – that those engaged in the classic debate seek to justify with their competing theories. The question they ask is when, if ever, punishment is justified; they do not argue over whether the imposition of a loss upon a person known to have committed no offense is justified. If


utilitarianism is sometimes committed to saying that such losses can be justified, this may speak against that theory as a moral theory, but not against utilitarian theories of punishment particularly. In closing off that objection to utilitarian theories of punishment by definitional stop, I have if anything made the position of the opposition stronger, and surely that kind of question begging is unobjectionable.

But my definition of punishment might be thought to beg the question in a different way, for it commits us to a kind of retributivism: namely, negative retributivism. I accept negative retributivism on conceptual grounds here; I will offer a more substantive argument for it later in the paper.

It is true, though, that even negative retributivism is a substantive thesis about the justification of punishment, and so someone might think that acceptance of it on mere conceptual grounds, in a paper designed to defend retributivism, is objectionably question begging. Part of the reason for my easy acceptance of negative retributivism is that virtually everyone accepts that it is a necessary condition of justified punishment that it be imposed only on those guilty of a legal offense. The lines of debate are not drawn between those who think that guilt is necessary for punishment to be just and those who think it unnecessary. What people disagree about is whether legal guilt is sufficient for punishment to be justified; they argue about the merits of some positive version of retributivism. I offer this paper as a contribution to that debate.

3. TRUST AND THE LAW

I shall limit my discussion in what follows to criminal law for the most part, though I believe that what I say of it can be applied mutatis mutandis to contract and property law as well. Tort law, on the other hand, is importantly different on my account, and would require separate discussion. I shall only briefly indicate how I see this difference later in this section.

The criminal law is a purposeful code of rules. The purpose of this law is to secure certain forms of behaviour, by making the doing of certain acts illegal. By making an act an offense, society expresses its condemnation of that act; by attaching penalties to the commission of that act, the law provides an additional reason for individuals to
refrain from doing it (distinct from any other moral or prudential reasons they may already have to do as the law requires). Together the denunciation and threat of penalty are designed to guide human behaviour, to reduce if not eliminate the incidents of that act being performed.

Which acts should be prohibited in this way? In the case of the criminal law (as well as in the law of contracts and property), I suggest, those acts which, if performed by any sizable minority or with any frequency, would undermine basic trust between members of a community: trust in each other, in the first instance, and trust in the legal system itself as an institution designed to maintain trust secondarily.

There is a danger that this point may be misunderstood. We are here concerned with understanding more fully the functionalist theory of law which serves as the backdrop to the defense of punishment, rather than with the defense of punishment itself. This is important, because my claim is not meant to imply that individuals may justifiably be punished not for what they have done, but because of the adverse effect on community trust of the combined actions of groups of individuals. To punish some for the actions of others would be unacceptable on my view of retributivism, which takes a past offense committed by the individual to be both necessary and sufficient to justify punishing her. Rather, we are here further exploring the nature of an offense. Our question is which violations of trust are sufficiently important to warrant being prohibited in law and can justly be punished. The answer here offered is that those acts which violate the conditions of basic trust between members of a community are properly punishable.

In claiming that securing basic trust is the central purpose of law I am not, of course, making any claim to originality. Among those who held this view are Thomas Hobbes and David Hume. Hobbes has vividly described the conditions in which persons live without trust: it is the state of “warre”, in which individuals are driven to anticipatory violence by diffidence, and the life of man “solitary, poore, nasty, brutish and short”.

While one may disagree with Hobbes that it takes the establishment of Leviathan, a sovereign

---

with absolute and indivisible power, to overcome such a state, one can nonetheless appreciate the plausibility of his description of what life would be like without basic trust. One could, more moderately and with Hume, insist that some forms of trust can and do arise prior to the establishment of political society and the creation of a legal system. It is because we see that such trust, and the relations it makes possible, are "infinitely advantageous to the whole and every part" of society that we seek to reinforce the conditions of trust through the use of law.\footnote{David Hume, \textit{A Treatise of Human Nature} 2nd Edition, ed. L.A. Selby-Bigge (Oxford: Clarendon Press, 1978), Book III, Part II, Section II.} We can agree with Hume that this is the central purpose of law.

In doing so, we shall be following the lead of a great many contemporary philosophers (especially contractarians), economists, and others interested in rational choice and game theory, who have demonstrated that a lack of trust frequently results in outcomes of interactions that are (Pareto) sub-optimal for the participants in strategic choice situations.\footnote{The literature here is vast, and sufficiently well-known that a comprehensive bibliographical reference is not required. I have been especially influenced by the work of David Gauthier: \textit{Morals by Agreement} (Oxford: Clarendon Press, 1986) and \textit{Moral Dealing} (Cornell University Press, 1990). Cf. the collection of essays in \textit{Rationality in Action}, ed. Paul K. Moser (Cambridge University Press, 1990).} While the recognition that an absence of trust leads to sub-optimal outcomes lacks some of the rhetorical force of Hobbes's description of the state of nature, such results are surely relevant for our understanding of the law. For the general emphasis in such work has been on demonstrating, first, that trust is necessary to secure peace and the benefits of cooperative interaction among people, and, second, that a system of coercive rules is necessary to secure trust (at least some of the time).

This emphasis on securing the conditions of basic trust in a community needs to be clarified, however, for trust has both subjective and objective conditions. The purpose of the law is to maintain the objective grounds of trust.\footnote{The only defense of punishment grounded explicitly in trust that I am aware of in the literature is David Hoekema’s "Trust and Obey: Toward a New Theory of Punishment", \textit{Israel Law Review} Vol. 25, Nos. 3–4, 1991. Hoekema concentrates only on the subjective conditions of trust, however, and so I think his project is implausible.} In any complex society, coercive rules will have to secure trust among those who have only temporarily
limited interactions with each other, and between whom no special relationships or tutistic interests bind, as well as among those more intimately connected. In order to understand the relation between law and trust, however, we need to say a bit more about trust itself.

Trust is a relation between three relata: the trustor, the trustee, and that which the trustor entrusts to the care of the trustee. While I will not offer a comprehensive analysis of trust, we can say that trust of another individual, as well as trust in our institutions, has certain general characteristics. Trusting others requires that we willingly place some interest or good of our’s into the hands of the trusted, when doing so involves some risk of being disappointed and having our interests thwarted or neglected, or our expectations disappointed. At this level we trust particular individuals with specific interests. Our trust can be mutual or one-sided; it can be between people of varying degrees of intimacy and power. Trust always involves a willingness not to monitor too closely or too frequently that the trustee is acting with due concern for our interests, and so whenever we trust we place ourselves in a position of vulnerability, for the trustee can disappoint our expectations and fail to competently care for the interests we have placed in his hands and under his power. For this reason, Baier remarks, philosophers such as Hobbes have thought that trust was a pathological emotion.

Yet pathological or not, we do often trust people. We stand in a relation of trust to a great many people: intimates, associates and strangers, and in the relation of both trustor and trustee. Some work has been done on explaining the subjective conditions of trust, i.e., those conditions which actually cause individuals to trust others. The subjective cues by which we come to judge some people to be trustworthy and others to be untrustworthy vary enormously: shifty

---

16 Cf. Annette Baier, “Trust and Antitrust”, Ethics 96: 2 (1986). Baier has recently collected a number of her essays on trust (among other topics) in Moral Prejudices: Essays on Ethics (Harvard University Press, 1994). I have benefitted enormously from reading Baier’s work on trust, and I borrow freely from her in what follows, both in terminology and in content. I doubt that Baier would agree with much of what I say in this paper, however. I have also benefitted from reading the essays collected by Diego Gambetta in Trust: Making and Breaking Cooperative Relations (New York; Basil Blackwell, 1988). My debt to the authors in the Gambetta volume will not be quite so obvious as that I owe to Baier in what follows.

eyes, an unwillingness of another to meet our gaze, knowledge of a person’s past behaviour, institutional roles, cultural norms and restricted choices about whether to trust all influence our willingness to trust. Colleagues in the social sciences tell me that we tend to trust others who share our ethnic or racial heritage more readily than others, and that the same is true with respect to class.

Intimate interpersonal relationships, as well as institutional arrangements, often make those involved in them more willing to trust. Toward intimates, of course, the scope of our trust is great: we trust them not only with a great many things that we value, but we grant them extensive discretionary powers in caring for those things. When we do so because we have experience which indicates that those who are trusted both care about our interests, at least in part because they are our’s, and desire to protect and advance them, and show good judgment in their use of their discretionary powers in doing so, such trust is reasonable rather than pathological. Such experience makes reasonable our belief that our intimates will not deliberately or carelessly act in such a way as to put our interests in jeopardy, and so make trust reasonable. Trust among intimates is particularly risky in the absence of good reasons to trust, however, both because the range of interests that are entrusted to the trustee tends to be extensive and because we cannot be too vigilant in monitoring the trustee’s performance in caring for what matters to us (constant monitoring being a sign of mistrust). Due to the intimacy of our relationships, moreover, the ways in which they can violate that trust is also greater than with others, because of their more extensive knowledge of facts about us that they can exploit. When such trust is violated, the intimacy of the relationship typically cannot be maintained.18 Should the trustee fail to care about what matters to

18 This may seem too strong, for counter-examples come readily to mind: abused wives often stay with their abusive husbands, for example, indicating that some relationships can withstand substantial trust violations. I suspect, though, that such cases are explainable as cases in which the abused (and the abuser) believes that it is lack of ability (self-control, etc.) that leads to the abuse, rather than a genuine conflict of interests or lack of concern. Such cases, involving as they do shortcomings at the level of exercising discretionary power properly, may seem not to warrant complete withdrawal of trust. These cases differ from those in which the abuser’s actions are motivated by lack of concern and affection for the abused, or by interests which genuinely conflict with those of the abused. In the latter sorts of situations, the relationship simply cannot be maintained in most cases.
us, or should she care only because our interests happen to converge for other reasons, or should she act carelessly or ineffectively in her attempts to advance our interests, these facts would undermine the subjective grounds of continued trust.

Among associates we also maintain various trust relationships. We trust teachers to educate our children rather than to abuse them, colleagues not to take credit for our work, etc. Those with whom we deal on an on-going basis in society are in a position to harm our most vital interests, yet we trust them not to do so. Some of these forms of trust are institutionalized: we trust teachers with our children, bankers with our money, mechanics with our cars, doctors with our bodies, clergy with our secrets, just because they occupy certain institutional positions within society. The institution is defined, at least in part, by the trust relations it involves. When these trusts are violated, it is especially shocking.

Trust between intimates and associates may be subjectively reasonable: we may have good reasons to think that our fellows are trustworthy. These reasons must be grounded in the experience of functioning webs of trust, however, and these seem to depend upon a certain coincidence of interests between the trustor and the trustee (as least with respect to those interests that are entrusted to the care of the trustee). Among intimates, we expect that the interests of the trustor will matter to the trustee, at least in part because they are the trustor’s; we expect that affection results in a coordination of interests which makes trust reasonable. In the case of associates, the coordination of interests may be mediated by institutional roles, but such a coordination is nonetheless still crucial. Through the coordination of interests and the shared expectations to which institutional practices give rise, individuals can come reasonably to believe in the trustworthiness of their associates.

This all points to two important things. First, because subjective trust is often maintained even when it is clearly not reasonable, subjective trust cannot be used to ground the defense of punishment being offered here. Second, the differing responses to these cases is reflected in different aspects of enforcement as it is actually practised. What Wesley Cragg calls the “enabling” function of enforcement, which is designed to enable those who might not otherwise be able to obey the law to do so, is an important part of policing; this is appropriate when lack of control and ability are at issue. Cf. Wesley Cragg, The Practice of Punishment: Towards a theory of restorative justice (Routledge, 1992). Such a response may be inadequate in cases of genuine conflict of interests, however.
The scope of our trust vis-à-vis strangers is less extensive, but it is nonetheless considerable. We trust strangers, on a daily basis, not to violate our interests in bodily security, to leave us in possession of our non-bodily goods, to deal with us fairly and honestly in economic transactions, etc. But what reason do we have to trust strangers? We lack the subjective conditions for trusting, for we lack the experience upon which the belief that others are trustworthy depends. We cannot assume that they are motivated by benevolence and affection toward us, or that they care about our interests because they are our’s. Nor can we assume that our interests will not come into direct conflict with those of others; sometimes we do face zero-sum games, and even more often, we find ourselves in choice situations which can be modelled by the prisoners’ dilemma. Furthermore, it must often be the case that strangers are not sufficiently knowledgable about what matters to us to choose means appropriate for assisting us (or at least not hindering us) in the pursuit of our goals, even if they were motivated to do so. It would seem, then, that trusting strangers cannot be subjectively reasonable, at least without much more knowledge than we typically have concerning both their past behaviour and their current motives, as well as their level of competence. Among strangers, then, we lack the bases for subjective trust.

Even among those more intimately related, there may be subjective reasons not to trust, and our trust in any of these relationships need not be unlimited, of course. While some intimate relationships – those between partners, friends, parent and child – could not survive without extensive trust, one may qualify that trust if the trustee has shown poor judgment in the use of her discretionary powers, or if some set of our interests are in serious conflict with those whom we trust. We are well-advised, even when dealing with those who occupy positions of trust within our institutions, to check for ourselves that that trust is well-placed; thereafter, some judicious monitoring may also be warranted. When dealing with strangers, we are well-advised to limit our trust in even more drastic ways.

The criteria for reasonableness of trusting, and the conditions needed to sustain such trusts, vary enormously across the categories of intimates, associates and strangers (rough and ready as they admittedly are). Furthermore, these various forms of trust can be violated in many diverse ways. The child’s trust can be violated by not receiving
some promised benefit or gift, as well as by being physically abused or neglected. A wife who commits adultery violates the trust of her husband; if she then knowingly or even carelessly infects him with HIV, contracted from her lover, she violates his trust in a different way. A stranger can violate our trust by purposely giving us false directions in response to our request for assistance in navigating about an unfamiliar town, just as he can by holding us up.

These examples point to something important for my purposes, and bring us back to the question of which violations of trust merit the attention of the law, for only the latter violations of trust are illegal and properly punishable. There are a great many violations of trust, among people of varying degrees of intimacy, that are not illegal (and should not be). What distinguishes the former from the latter? One might think that the difference is to be identified according to the harmfulness of the betrayal. The wife who infects her husband with a deadly virus, the abusive or negligent parent, and the robber all act in ways which are more harmful than the parent who merely does not keep her promises vis-a-vis benefits she will confer upon the child, or the wife who cheats on her husband, or the stranger who misdirects us just for fun. I think this would be a mistake; though those violations of trust which are properly punishable will frequently be those which cause considerable harm, the crucial difference is not to be cashed out in terms of the harmful consequences of violations of trust. For the stranger who misdirects us may cause us more real harm than the one who robs us. We can easily imagine such circumstances: we are seeking directions to a hospital, where our aged father is dying, and because of the misdirection of the stranger we arrive too late to say our final good-byes, or we miss an important business meeting resulting in a substantial financial loss. As a robber, on the other hand, our stranger may not actually harm us all that much: handing over the little bit of money that we have in our wallet may not be any considerable hardship in itself, and if we believe that the thief will refrain from harming us provided we hand the money over, we may not even be particularly traumatized by the incident. More generally, there can be violations of trust that result in no palpable harms to others, that are nonetheless properly punishable, such as the dissemination of hate literature. A person who distributes hate literature does so with the hope of inciting hate and violence against
some identifiable group; it may be a matter of luck that some such material falls on deaf ears, and results in no real harm being done.

To understand which violations of trust require punishment and which do not, we need to move beyond the subjective conditions of trust. For all of the betrayals of trust considered above equally undermine the belief that those whom were trusted were in fact deserving of that trust, yet not all are properly dealt with by the law. To understand the difference we need to understand how the law serves basic trust, which is objective rather than subjective. That is, on this functionalist theory of law, the law must serve to make trust in others objectively reasonable.

This characterization of law, as an institution that makes trust objectively more reasonable, requires that the law function really to produce meta-trust, trust in trust. In the examples given above what distinguishes cases of trust violations that are punishable from those which are not is that only the former involve violations of meta-trust, as well as the particular subjective trust that the violator disappointed. Not all specific violations of trust undermine meta-trust. Those violations that make mistrust more objectively reasonable than it would otherwise be, especially mistrust of those other than just the violator, are betrayals with which the law ought to be concerned. The worst cases involve violations of trust that make anticipation and preemptive violence more likely. But far short of this extreme, many violations of trust undermine the objective conditions of trust; they make mistrust more reasonable, both for the victim and often for third parties as well. The spouse who is infected with HIV, the child who is abused, and the victim of the robber, all have good subjective reasons to mistrust those who have treated them so badly. But if society acquiesced in their being so treated, this would make everyone less trustworthy in an objective sense. Knowing that our fellows have allowed us to be victimized without complaint and protest and condemnation, or that they are unwilling to assist us in providing protection against further abuse, would make trust of them, and not just of our violator, less objectively reasonable (and in most cases subjectively less reasonable as well). The climate of trust between men and women, for example, is not determined on an individual or subjective level only. Even if a woman happens to avoid being abused or sexually assaulted by the men with whom she
must live, through a combination of skilful judging of character and a lot of good luck, she cannot have objectively grounded reasons to trust men if she knows that men can abuse and assault women with impunity. If such behaviour is tolerated by the members of her community, moreover, then knowledge of this fact will influence the climate of trust in such a way so as to make subjective trust more difficult to achieve or sustain. In this way the objective conditions of trust can influence subjective trust. The objective conditions of trust depends on a risk-assessment model, while subjective trust depends upon a willingness to trust; the subjective can (and should) be influenced by the objective, however, and so we can say that it is violations of the objective conditions of trust, such that if they go unpunished then the subjective conditions of trust are made less reasonable between the victim and others besides just the violator, that the law must be concerned to prevent.

The most direct influence that the law has on promoting meta-trust depends upon its coercive nature. By attaching penalties to certain behaviours the law provides a serious disincentive to engaging in those acts. We trust the stranger beside us on the subway not to attack, at least in part, because we know (and we can assume that he knows) that such behaviour is illegal and likely to be punished. The penalty, as well as the more positive aspects of the law as an educative tool in societies, makes it less likely that our fellow traveller will be motivated to harm us. The same sort of consideration, though, leads us to trust restauranteers not to inadvertently poison us with spoiled food, for we know that the law provides for routine health and safety standards which, if followed, protect us from such an occurrence. More generally, the law serves to promote an atmosphere in which specific trust relations can be achieved and sustained by providing a background of shared and relatively stable expectations about how others will and will not act. By doing so, the law makes it less necessary for individuals to have knowledge of the past behaviour and current motives of those whom they trust. Furthermore, the law relieves individuals of the burden of exercising discretionary power in not thwarting the interests of others in a great many cases; provided that the laws are well-framed, obedience is all that is required. Likewise, the law makes special ties of community and assumptions of benevolence or altruism less needful. The law
makes subjective trust easier to achieve and maintain; it makes trust more reasonable.

It is not enough for the law to serve its central purpose, of reinforcing basic objectively-grounded trust, merely that a threat of sanctions be attached to violations of trust. If the rules of law are to provide a public expression of condemnation of those forms of behaviour that undermine trust in a community, the community must support the imposition of those sanctions. The point here is not merely that the threat of a sanction would not be convincing unless it was supported by the community, and so it would not serve to influence behaviour in the way it is designed to do, though that it likely true. More importantly, though, those who violate the conditions of community trust violate more than just the trust of their victims. The withdrawal of trust that such a violation warrants must, then, also be withdrawal from more than the victim.

Just as there must be reasons for our subjective belief that another is trustworthy in order for our trust to be reasonable, we can have objective reasons to believe that trust is unreasonable. The person who violates the criminal law, in the absence of excusing conditions, demonstrates conclusively that there are reasons not to trust her. That is, an offender who commits a crime, when the usual conditions of *actus reus* and *mens rea* are present (together with their required connection), demonstrates conclusively her lack of trustworthiness. Such offenders demonstrate both a willingness and ability to violate the conditions of basic trust in society.

But why do such displays of untrustworthiness merit punishment? Why, that is, must society actually impose a loss upon the offender for breaking trust? In the first place, the reason is that punishment is necessary to reaffirm for the members of the community the commitment to basic trust. We restore the objective conditions of trust by punishing. Here the distinction between the objective and subjective conditions of trust is central. Whether punishment in a particular case actually restores the willingness of the victim or others to trust the offender (or others) is a contingent matter. There is no reason to believe that punishing offenders necessarily affects the extent to which members of a community actually trust anyone. Yet the objective conditions of trust are restored through punishing offenders, independently of the effects which punishing has on the willingness
of individuals to actually trust. For the objective conditions of trust are tied to risk-assessment, which is necessarily affected by the willingness and ability of societies to punish offenders. I have at least hinted above that I think the restoration of the objective conditions of trust will typically influence the subjective willingness of individuals to trust, but that is an independent matter.

Secondarily, it is necessary that when trust between members of society has been violated, trust in the law as capable of maintaining the conditions of trust be reaffirmed. Punishment serves this purpose. Those who commit offenses demonstrate that the ability of the law to maintain the conditions of trust in the community is not complete; punishing the criminal, however, serves to reestablish that trust and demonstrates that individuals need not adopt recourse to anticipatory violence as a means of protecting their interests against those who are willing to harm them.

If the criminal is not punished, this will demonstrate an unwillingness or inability (or both) on the part of the community to maintain the conditions of trust. It would undermine the conception of law as designed to maintain trust. In the absence of that conception of law, we should not be able to defend the reasonableness of trust in the face of willing violators, and so we have a general reason for claiming that punishment ought to be imposed upon those who violate the law, just in virtue of their offense.

This understanding of the role of law and the practice of punishment as fundamentally one of maintaining basic trust thus provides us with a justification of positive retributivism. It also provides a direct argument for negative retributivism. For if the law is to maintain trust in a community, punishment must be inflicted only upon those guilty of an offense. To institutionalize, or even allow in exceptional cases, victimization, would undermine the objective conditions of trust rather than promote them. This need not be true in the case of subjective trust, of course, which could be reestablished by punishing the innocent, and so we have another reason to insist that basic trust be construed objectively rather than subjectively.

This has been no more than a sketch of what form I think the most promising retributive theory of punishment would take. Yet it might be thought seriously incomplete, even for a sketch. For I have not even mentioned another feature of virtually every retributivist
theory: the requirement that punishment be proportioned to the culpability of the offense. If punishment is required just in virtue of a criminal’s having committed an offense, the severity of that punishment must be proportioned to the gravity of the offense. What does the requirement of proportionality come to on my view? What punishment do offenders deserve?

The level of punishment that offenders are thought to deserve on retributivist theories must be proportioned to the wrong they have done. Insofar as different retributive theories offer different conceptions of the wrong which is characteristic of criminal offenses, they also defend different conclusions concerning proportionality. Hampton, for example, argues that the wrong which crime involves is a demeaning of the value of the victim and inflation of the status of the criminal; punishment must reaffirm the equal worth of the victim and the offender, and so requires that the criminal suffer a defeat which will ‘bring him down’ from his self-proclaimed position of superior value to one of equality with his victim.

I have identified the wrong that is characteristic of crime as a violation of the conditions of basic trust in a community. Separate from any material harm which an offense may cause to the victim of the crime (which harm is properly remedied through the law of torts), every legal offense will involve a violation of the conditions of basic trust in the community (which violation is properly remedied by the imposition of punishment upon the offender by the community as a whole, or by the ‘state’ as the agent of the community).

Punishment must then be proportioned to the need to restore the conditions of trust. I believe that this approach can provide some fairly determinate recommendations concerning the appropriate level of punishment for various categories of offenses. Consider a range of financial offenses, in which no one is personally menaced or physically harmed: suppose it is a case of corporate embezzlement in which the corporation is able to maintain all of its commitments to its customers, suppliers, shareholders and whoever else could be directly affected by its financial loss. Or suppose the offense is tax evasion, or failure to purchase various business licenses or permits. Such offenses are surely violations of public trust. What would be necessary to reestablish objective trust in the face of such violators? Perhaps the offender should be barred, for a determinate period of
time including life, from holding an institutional position of financial trust, or from operating a business, or prevented from receiving any future tax returns for which she might be eligible. She may be required to pay a fine and/or perform community service. She may be forced to submit to random audits and other invasions into her privacy in order to ensure future conformity to the laws that she has broken. Incarceration seems unnecessary, however, for such behaviour would be unlikely to lead to anticipatory violence in the community. It is necessary to ensure that such crimes don’t pay, to prevent an increase of free-ridership and the instability that can result therefrom, but monetary, supervisory, regulatory, non-custodial penalties seem sufficient for that purpose.

Lest it be thought that I am displaying some prejudice of privilege toward “white collar crimes” in what I have just said (which would be wrong on at least two counts), the theory of punishment I am proposing would often recommend non-custodial forms of penalty. Even more serious cases of robbery, break and enter and the like ought often to be dealt with by means of punishment other than incarceration. These crimes do more to disturb the conditions of trust, and make anticipation more reasonable. Yet they should be dealt with within the community whenever possible. Given the basic demographics of this group of offenders – poor, underemployed or unemployed, young and male – the interests of trust are not best served by incarcerating them in penal facilities like our modern prisons. Employment and educational programmes under suitable supervision seem more likely to facilitate the goal of reestablishing trust in communities when dealing with such offenders.

Retribution need not be retaliatory, and indeed it should not be on my view. Institutionalizing vengeance will not serve to promote trust. The view I am advocating would encourage the use of community-based approaches to non-custodial forms of punishment. To make such penalties effective, substantial reform of our penal institutions and practices would have to be undertaken. In particular, the range of intermediate penal establishments, between prison and probation, must be expanded.  

---

19 For an excellent selection of essays that explore many alternatives to incarceration see A Reader on Punishment, eds. Antony Duff and David Garland (Oxford University Press, 1994).
Recourse to prison cannot be eliminated, however, and I do not follow contemporary abolitionists in thinking that punitive penalties, including incarceration, can be or should be abolished. Violent criminals, through their acts, pose a direct threat to trust in society. Their offenses make diffidence and anticipation more objectively reasonable. Such offenders must be removed from the community by incarceration. First time offenders should be detained in facilities that provide opportunities for rehabilitation, including education, training, and psychiatric counselling, and should be released back into the community as soon as possible. The release of serious offenders must occur only gradually, however, through a decreasingly supervised series of community-based facilities. In the case of some repeat offenders of very serious crimes, incarceration may have to be for life.

I believe that concentrating on the violation of trust which legal offenses involves gives us the resources to answer both of the questions that R.A. Duff indicates must be answered by any adequate retributive theory: 1) Why do the guilty deserve to suffer (to have a loss imposed upon them)? Why is punishment the proper response to legal wrong-doing? 2) Why should the state (rather than God, say, or the victim) actually impose that suffering upon them?20

4. A COMPARISON WITH SOME ALTERNATIVE RETRIBUTIVE THEORIES

Though I follow the same argumentative strategy as some other contemporary retributivists, I think that the account offered here has certain advantages over the specific theories developed by Morris/Murphy/Davis and by Hampton. The theory developed by Morris, Murphy and Davis, which I shall call the "unfair advantage theory", argues that law is a system of restrictions, compliance with which benefits everyone. The cost of those benefits is individual restraint. Law-abidingness ensures that everyone bears a proportionate share of the burden; but those who violate the law take an unfair advantage over the law-abiding, for they enjoy the benefits of the law without bearing the costs in individual restraint. The

20 Duff, Trials and Punishment, 7.3.
criminal, on this view, commits the injustice of being a free-rider. But it is important to understand the nature of the advantage that criminals supposedly take from their wrongdoing. It is not any material advantage that is the harm characteristic of legal wrongdoing (the loot from the robbery, say, or feelings of satisfaction that criminals might take from assaulting someone they hate); rather, the advantage they take is just in not restraining themselves while others do. But this is not only descriptively questionable; it is morally repugnant. For on this view the rapist or the murderer deserve punishment just because they have taken unfair advantage of their law-abiding fellows. While we bear the burden of self-restraint (resisting the temptation the kill and rape), the murderer and the rapist enjoy the benefit of giving free reign to such desires in themselves. I take it that both the misdescription and the perniciousness of such an account will be obvious.

Furthermore, the unfair advantage theory might be thought to imply that all offenses are equally serious, since the wrong is the same in each case: free-riding. Davis, to concentrate only on the most recent proponent of this view, argues that this is not so, by appealing to the idea of an auction for licenses to commit crimes (a license functioning as a pardon-in-advance). How much money various licenses would fetch in the auction determines how serious the various crimes are. But this does not really help; indeed, it reinforces the morally objectionable underpinnings of the theory. For on this view, it would have to be the case that a substantial number of people would be willing to bid on licenses to rape, say, and pay quite a lot for that “privilege”, if rape is to be deemed more serious than tax evasion or theft, for example. If, as I would hope, many more people would be willing to bid on, and so bid up, the licenses that would provide substantial financial gains rather than rape or other violent crimes, the price of licenses to commit tax evasion and theft ought to be worth more than a license to rape, murder or maim. The conclusion would be that the former are more serious forms of wrongdoing than the latter.\(^{21}\)

Hampton’s view is more attractive, both descriptively and morally. On her view, the wrong that is characteristic of criminal

behaviour is that of objectively demeaning the victim; criminal actions are objectively disrespectful of the victim's worth. Such an account will surely match our judgments of what is wrongful in such central cases as rape and murder better than the unfair advantage theory can. And it provides a morally attractive justification of punishment. For punishment is conceived as expressing our condemnation of such claims of superior value that the criminal makes through his or her actions, and reaffirms the equal worth of the victim. I have three reasons for preferring to concentrate on trust rather than individual value, however. First, Hampton's theory rests on a rigorous Kantian theory of human worth, which holds that all persons are intrinsically, objectively and equally valuable. Insofar as such a theory is itself controversial, the resulting justification of punishment will likewise be controversial. In the theory I have been advocating, on the other hand, all that must be conceded is that basic trust is necessary in society and that law serves to secure that trust. This seems much less controversial, and so provides a firmer foundation for the justification of punishment. Second, Hampton's theory seems unduly narrow in its scope; while it may capture many of our intuitions concerning the wrong done in violent crimes, it would have no application for "victimless crimes" (if there are, or should be, such), or for crimes against the state or social institutions. Insofar as various institutions operate and give rise to expectations among individuals in a society, actions which undermine those institutions could properly be seen as disturbing the conditions of trust in society and so could be included within the theory I have been advocating.

Finally, and perhaps most importantly, Hampton's view requires that we see the purpose of punishment as reestablishing the equal worth of the offender and the victim. The defeat (loss) imposed upon the offender is merely supposed to negate the claim to superior value implied by the offense, bringing the offender and her victim to a position of equal moral status. But why is equality the correct goal? If all that criminal behaviour merits is a response that brings about a state of equality, then criminals "have nothing to lose" by their offenses.
Admittedly, none of these are decisive reasons for rejecting Hampton's theory, or preferring that being offered here. Perhaps this debate will have to wait for its final resolution until both views are more fully developed.*

5. SOME SECONDARY ADVANTAGES OF THE VIEW PROPOSED HERE

The concentration on trust allows us to unify a number of otherwise disparate features of what is commonly assumed by those who believe that punishment is, under some conditions, justifiable. In particular, it allows us to explain a) why certain excusing conditions are and ought to be recognized in the law, and b) the intuition which many theorists share concerning the justification of punishing at least some unsuccessful attempts as severely as completed crimes.

a) The concentration on trust gives us a plausible justification of recognizing certain defenses and excuses in the law. Those acts which are deemed to be justified (e.g. killing in self-defense) preclude both conviction and punishment. In such a case a person acts in such a way as is not considered to be wrong, because the act does not disturb the conditions of trust. Likewise with cases of excuses proper (e.g. accident, mistake, provocation, duress and insanity, to use Hart's classification of them22), though what the person does is wrong, the absence of intention and the other conditions of mens rea make it the case that the action does not violate the conditions of trust. When a person commits an act that would otherwise be illegal by accident, mistake, etc., he does not thereby demonstrate his untrustworthiness. Accordingly, he does not merit punishment.

Completed crimes, when performed under one or other of the excusing conditions recognized in Anglo-American law, do not demonstrate the untrustworthiness of the person who performs the action. Completed crimes, done in the absence of any excusing conditions, do, on the other hand, demonstrate that the offender does

---

* Note added in proof: The author was saddened to learn of Jean Hampton's untimely death while this paper was in preparation for publication. Analytic philosophy has thus lost a compelling and insightful member.

not deserve to be trusted (indeed, he deserves not to be trusted). In
the most serious of criminal offenses, the offender violates the most
basic conditions of trust.

b) Likewise, some attempted but incomplete crimes demonstrate
that the person who attempts the crime is just as untrustworthy as
the person who succeeds at the same crime. This is true particularly
in those cases where failure to complete the crime is due to what
Hyman Gross calls “manifest impossibility”; failure in such cases
is attributable just to factors beyond the would-be offender’s ability
to control or predict. In these circumstances, failure is a matter of
luck.23 Because such would-be criminals demonstrate their untrust-
worthiness as fully as a successful criminal, such attempts deserve
to be punished as severely as completed crimes. Though Anglo-
American law has typically not followed the advice of its theorists
in this respect, and so has not typically punished mere attempts as
severely as the corresponding completed crime, it does treat many
attempts themselves as offenses (albeit carrying lighter penalties
than the complete offense). The view being advocated here can pro-
vide a defense of this practice, whereas any theory which claims
that punishment must somehow be proportionate with harm actually
done, unfair advantage actually taken, or the objective demeaning
of the victim actually resulting, cannot.

6. CONCLUSION

I have been able to do no more than indicate what form I think a
promising theory of retributivism would take in this paper. If we
accept that the law serves to secure the objective conditions of trust
within a community, I have argued, then offenders must be punished.
A number of questions remain to be addressed, however, even if this
trust-based account is accepted. What is the connection between the
objective conditions of trust and the subjective willingness to trust?
What is the connection between this justification of punishment and
more traditional desert-based accounts? What are other plausible
extensions of this theory: does it justify minimally decent samari-
tan laws, for example? Which, among the wide array of intermediate

penal institutions that are available, would best protect and foster the conditions of basic trust? These questions, and many more, remain to be explored.*

Department of Philosophy
York University

*I would like to thank Roger Shiner, Bob Bright, Paul Viminitz and Jill Flohil for comments on an earlier version of this paper, as well as participants at the Ontario Philosophical Society's Annual Meeting 1995, the Western Canadian Philosophical Association's Annual Conference 1995, and Philosophy Departmental Colloquia at the University of Toronto and York University.