Ethic Theory Moral Prac (2013) 16:1029–1047
DOI 10.1007/s10677-013-9402-y

The Expressivist Account of Punishment, Retribution, and the Emotions

Peter Königs

Accepted: 7 January 2013 / Published online: 16 January 2013

Abstract This paper provides a discussion of the role that emotions may play in the justification of punishment. On the expressivist account of punishment, punishment has the purpose of expressing appropriate emotional reactions to wrongdoing, such as indignation, resentment or guilt. I will argue that this expressivist approach fails as these emotions can be expressed other than through the infliction of punishment. Another argument for hard treatment put forward by expressivists states that punitive sanctions are necessary in order for the law to be valid. But this justification of punishment, too, is unconvincing. There are no good reasons to assume that we have to resort to punitive measures in order to vindicate the law. I will then raise the more general worry whether there is any intelligible link at all between moral emotions such as indignation, resentment or guilt and retributive behaviour. I will finally conclude with some sceptical remarks on the moral worth of retribution.

Keywords Punishment · Retribution · Emotions · Guilt · Indignation · Hard treatment

1 Introduction

Philosophers of punishment are confronted with what John Leslie Mackie called the paradox of retribution: "The paradox is that, on the one hand, a retributive principle of punishment cannot be explained or developed within a reasonable system of moral thought, while, on the other hand, such a principle cannot be eliminated from our moral thinking" (Mackie 1982, 3). In recent years, it has become popular to solve this paradox by justifying retribution in terms of its expressive or communicative function.[1] Punishment serves the purpose of expressing indignation and of conveying one's disapproval of the crime to the offender. The expressivist account of punishment thus has, or hopes to have, found a way of dispelling the common impression that the desire for retribution is nothing more than some brutal, archaic instinct.

[1]Theorists who stress the expressive or communicative function of punishment include Bennett (2008); Duff (2001); Hampton (1992); Kleinig (1991); Primoratz (1989); Tasioulas (2006); Skillen (1980). At least to some extent, this applies also to Andrew von Hirsch; see Hirsch (1993, 12–13). For a very good introduction to expressivism, see Brooks (2012, chapter 6).

P. Königs (✉)
University of Cambridge, Queens' College, Cambridge, UK
e-mail: peterkoenigs@gmx.de

Another, related, trend in the philosophy of punishment is the focus on emotions.[2] There is little doubt that emotions such as resentment, indignation and guilt play a major role in explaining and possibly in justifying the practice of punishment.

A natural way of combining the expressivist account of punishment with the focus on emotions is to argue that the purpose of punishment is to express our retributive emotions. An expressivist theory of punishment of this kind has recently been put forward by Christopher Bennett in his "Apology Ritual" (2008). In what follows, I will offer a critique of the expressivist justification of punishment and argue that it ultimately fails. I will then proceed to expound in a more general fashion why the prospects of understanding – let alone of justifying – retribution in terms of the emotions that issue in retributive behaviour are rather bleak. I will therefore conclude my paper with some sceptical reflections on the moral worth of retribution.

Before I start with my discussion of Bennett's expressivist account of punishment, I would like to make some clarifying remarks about what it means for an action to be expressive of emotion. This will help us better understand the underlying rationale of the expressive account of punishment and its flaws.

There are, of course, many ways that actions and behaviour can be expressive of emotions or that emotions may trigger actions (see Goldie 2000, 123–140). I will not discuss all of them. Rather, I would like to focus on two different ways emotions can issue in action.

First, there are what philosophers of emotion have come to call expressive actions (see Döring 2007). An expressive action is an action that is triggered by an emotion but not performed as a means to some further end. It is expression for expression's sake, as it were. This expression may take a non-symbolic form. I could, for instance, express an emotion by banging my fist on the table. Banging one's fist on the table is not specific to a particular emotion; it is symbolically indeterminate. It may be an expression of anger, but it might as well express sadness, rage, joy, surprise and plenty of other emotions. I therefore take it that some expressive actions are (largely)[3] non-symbolic. However, there are also symbolic expressive actions. Such actions express emotions in terms of a symbolic language and are therefore symbolically determinate. An obvious example is cursing. Shouting "shoot!" is an expression of a specific emotion (frustration) in a highly specific symbolic language (natural language). But there are also other symbolic languages, such as art or gestures. I could (if I could) express sadness by playing Chopin's *marche funèbre*, and I can show disapproval by giving someone the finger. These are symbolically appropriate expressive actions.[4] Note that some expressive actions are intentional (playing Chopin), while others are presumably not (banging one's fist on the table in an outburst of rage).[5]

Second, emotions may also issue in actions that are purposive. When we are in the grip of an emotion, we often perform actions that – unlike expressive actions – *are* a means to an end. We act in order to bring about a different, better state of affairs. A case in point is fear. When in fear, we fight or flee, and either action has the purpose of bringing about a safer

[2] See e.g. Bennett (2002), (2008); Ciocchetti (2009); Holroyd (2010); Mackie (1982); Moore (1987). Somewhat relatedly, Martha Nussbaum has stressed the role of emotions in judging whether we should have mercy with wrongdoers (1993).

[3] I should note that banging one's fist on the table is not completely non-symbolic. It does convey the information that I am experiencing a *fierce* emotion. Nobody would bang his fist on the table if he were only faintly surprised or slightly upset.

[4] To be sure, these actions need not be expressive actions. They *could* also be done as a means to an end. I might play piano sonatas because I need to practice or in order to impress my girlfriend.

[5] I will not discuss whether these actions can be accounted for within the classical belief-desire model. On this question, see Hursthouse (1991); Smith (1998); Goldie (2000, 123–140); Döring (2007).

situation. Many emotions come along with what Peter Goldie calls "primitively intelligible" desires. Such desires "cannot be better explained in virtue of anything else other than the emotion of which [they] are part" (2000, 128). The same point is made by Robert C. Roberts. He reckons that some actions are very easily explained because they are "transparent by virtue of the 'logic'" of the emotions that cause them (2003, 172). This applies to many emotions other than fear: When we envy somebody, it is primitively intelligible that we undermine this person's efforts. When we love something, it is primitively intelligible that we protect it. By the same token, it is primitively intelligible that we try to destroy what we hate.[6] Ironically, however, it is not always obvious whether an action is primitively intelligible or not. I shall say more on this at a later stage.

The ensuing discussion of Bennett's approach will show that Bennett conceives of punishment as a (non-purposive) expressive action. After showing why Bennett's theory is doomed to fail, I will consider whether the desire to punish can be understood as a primitively intelligible entailment of guilt or indignation.

2 Bennett's Expressivist Account of Punishment

In a nutshell, Bennett argues that the state should mete out punishment to offenders because punishment is the appropriate expression of our indignation, which in turn is the correct reaction to an offence committed by a morally responsible agent. Emotional reactions to moral offences thus occupy centre stage in Bennett's discussion of retributive punishment. Let us look more closely at Bennett's defence of retributive punishment and at the role emotions play in it.

First, however, it is important to understand how the expressivist theory of punishment differs from the communicative approach. As Bennett himself points out, his account of punishment is expressivist but not communicative (2008, 188–189). Unlike the expression of one's thoughts or emotions, which can be a solitary activity, communication is a social activity, which presupposes that there is somebody to whom we communicate (Duff 2001, 79). The communication involved in communicative punishment is, of course, primarily directed at the offender, but it is not wholly unidirectional. By punishing wrongdoers, we communicate blame to them, but we also give them the opportunity to communicate their remorse to the public and to make public apologies. As Thom Brooks puts it: "We communicate our disapproval; offenders communicate their remorse." (Brooks 2012, 104) Antony Duff, the most prominent advocate of the communicative approach, thinks of the offender not merely as an addressee of blame but "as an active participant in the process who will receive and respond to the communication" (Duff 2001, 79). Notice also that on the communicative account of punishment, expressing moral resentment has a further purpose. Duff says: "Punishment […] aims at the goals of *repentance*, *reform*, and *reconciliation*. These goals are to be pursued by a communicative process of imposing penitential burdens on offenders" (2001, 107, original emphasis). Unlike Duff, Bennett does not think that punishment has a further purpose. On his account, punishment has no purpose other than that of expressing condemnation: "[T]he fundamental job of the criminal sanction is not to

[6] Note, however, that not all emotions entail primitively intelligible actions. If we grieve at the loss of a beloved person, our grief does not give us any end for action at all. This is because we cannot (unlike in the case of fear and envy) change the world in such a way that our grief is soothed, see Döring (2007, 385). Another emotion that does not give rise to any primitively intelligible desires is the emotion of surprise.

induce repentance or to achieve moral reconciliation between offender and community: its job is simply to express proportionate condemnation" (Bennett 2008, 148). Punishment, then, is justified because it is the appropriate way of expressing blame, not because it is an efficient means of communication. Expressive punishment is unilateral, whereas communicative punishment is reciprocal.

I will briefly consider Duff's communicative account at a later point.

Bennett's account is, however, not as straightforward as it first seems. In his discussion of how we ought to express our condemnation, Bennett takes a detour via the appropriate reaction of the *offender*. Bennett maintains that the offender ought to feel guilt and that he should therefore incur suffering. His undergoing of penance is the appropriate expression of his appropriate guilt. On Bennett's account, then, we "can communicate our condemnation by putting into symbols, not how indignant or outraged we are, but how *sorry* we think the offender ought to be for what he has done" (2008, 146, original emphasis).[7]

Why does Bennett take the detour via the offender's appropriate emotions and *his* appropriate expression? Why not assert straightaway that punishment is the appropriate expression of condemnation? Bennett gives two reasons in support of this move:

> Firstly, if we express our condemnation through symbols of outrage and indignation then we will be led to think about doing things to the offender that are angry and aggressive, even violent. But these are not things that the decent state should consider doing to its citizens […]. Second, thinking about expressions of outrage and so on seems to say the wrong thing to and about the offender. Expressions of outrage emphasise the distance between the offender and the community of decent persons. […] It might be more adequate from a symbolic point of view if the language of punishment communicated rather what the offender will have to do in order to resume her place in the polity. Rather than emphasising distance, it emphasises a process of reconciliation (2008, 147).

Both reasons warrant some discussion. The first reason could strike one as odd, given that it is arguably the nature of punishment to be violent. Notoriously, the state holds the monopoly on violence, and the state's infliction of punishment on renitent citizens is, at least in some sense of 'violence', the prime example of violent state action. To be sure, some forms of punishment (e.g. imprisonment) seem less violent or aggressive than others (e.g. corporal punishment).[8] However, every form of punishment is ultimately the authoritative infliction of harm upon the offender through state power and can therefore, I think, rightfully be called violent. Perhaps Bennett's point would be better put by saying that outrage and indignation dispose us to do things to the offender that are *cruel*, and that *this* is something that a decent state should not do to its citizens. His second argument is doubtful, too. Moral outrage and indignation need not necessarily emphasise distance. Rather, these emotions can show that we still think of the offender as a responsible moral agent and as part of the moral community. Holding someone responsible does precisely not create a distance between the offender and the community; it reminds the offender that he is still part of the community and expected to comply with its norms.

Be that as it may, I do not think that much depends on whether we take the offender's or the community's reactive attitudes as the point of departure. In fact, I think that (the community's) indignation is what we may call the mirror-emotion of (the offender's) guilt.

[7] It should read "express" rather than "communicate", though.

[8] I am grateful to an anonymous referee for pressing me to clarify this point.

Guilt represents *oneself* as morally culpable or blameworthy, while indignation represents somebody *else* as culpable. And both emotions give rise to the desire to punish the offender, be it oneself or somebody else. Whether we construe punishment as the appropriate expression of guilt or of indignation, it eventually comes down to the same thing.

Let us now move on and assess the cogency of Bennett's justification of punishment. At the heart of Bennett's account of punishment are two relations of appropriateness. First, we ought to exhibit the appropriate reactive attitudes toward wrongdoing. The offender should show guilt and remorse, the community indignation and resentment. Second, these appropriate responses should be given appropriate expression, namely through blame and punishment. In Bennett's own words: "[W]e should look at the emotions *appropriate* to cases of wrongdoing and how these emotions are *appropriately* expressed" (2008, 145, original emphasis).

Accordingly, a critique of Bennett's account may focus on either of these two relations of appropriateness. Everyone who is acquainted with the debate about expressive and communicative accounts of punishment will (correctly) anticipate that my criticism will focus on the second relation of appropriateness. However, before I do so, let me make some remarks on the first component of Bennett's theory.

It should be noted that, strictly speaking, we should not be concerned with which emotions are an appropriate response to wrongdoing but with which emotions one ought to feel. Arguably, emotions are subject to an ethics of emotion. That is, besides the question of whether some emotion is appropriate or not, there is also the distinct question of whether we ought to have this emotion.[9] To be sure, there is presumably an ethical presumption in favour of having appropriate emotions. Usually, we ought to have precisely those emotions that are appropriate. We should, for instance, feel pity when pity is appropriate or admiration when admiration is appropriate. Sometimes, however, it is considered unethical to have an appropriate emotion. Envy and jealousy, for instance, are thought to be emotions that one ought not feel, even when they are appropriate. And sometimes it is improper to be amused by a joke, even if the joke is funny (see D'Arms and Jacobson 2000). At times, you may even be expected to have an emotion that is patently inappropriate. It has been maintained, for instance, that if you kill or severely hurt somebody through no fault of your own, you should feel guilty despite your being blameless (e.g. Moore 1987, 205).

For his expressive justification of punishment, Bennett should ask whether one *ought* to feel guilt or indignation, not merely whether these emotions would be appropriate to feel. For if we ought not feel the relevant emotions in the first place, then we arguably ought not give them expression either. However, it does not seem particularly promising to launch an attack on Bennett from this direction. For on the face of it, it is not only appropriate but also right to feel guilty when one has committed a moral wrong or to feel indignant at a wrong done by somebody else.[10] Still, let me mention one possible objection to Bennett's account that rests precisely on the idea that some emotions are unethical. It is the Nietzschean objection that the human urge to punish is really motivated by *ressentiment*. Punishment would then not be the expression of guilt or indignation, but of less respectable emotions such as envy, sadism, fear and the like.[11] If we find Nietzsche's suspicion plausible, we should be very wary of expressive accounts of punishment, for punishment may turn out to be the expression of base and spiteful emotions, i.e. of emotions that we ought not have in

[9] On the ethics of emotion, see Neu (2010).
[10] Unless one thinks that we lack freedom of the will and are therefore not to be held responsible. I do not want to discuss this issue at this point.
[11] For a discussion of the Nietzschean objection, see Moore (1987).

the first place. However, I will not pursue this objection further here. Even if it is not wholly without plausibility, the case against the expressive justification of punishment should not rest on a theory as controversial as Nietzschean psychology.

However, as adumbrated above, the second relation of appropriateness is more critical. Ever since Joel Feinberg's seminal paper on the expressive function of punishment, it has repeatedly been maintained that we could (and should) find alternative ways of expressing disapproval (see e. g. Boonin 2008, 176–9; Hanna 2008; Hart 1963, 66; Holroyd 2010). Feinberg argued, convincingly I think, that it is part of the *definition* of punishment that it has an expressive function (1965).[12] However, he also indicated that there may be alternative ways of expressing reprobation: It may be true that, "[g]iven our conventions", hard treatment is the symbolically adequate expression of reprobation. He adds, however, that "[p]ain should match guilt only insofar as its infliction is the symbolic vehicle of public condemnation", implying that our present, somewhat cruel conventions may be replaced by more humane ones (1965, 423). This is the challenge that proponents of expressive accounts of punishment have to meet.

Of course, Bennett is aware of this problem and does not fail to give it some discussion. I do not think, though, that his response to the challenge is satisfactory. Let us look at how he addresses Feinberg's challenge.

Bennett contends that making amends is the appropriate way of expressing guilt. And since, on Bennett's account, the community's expression of condemnation should mirror the offender's expression of guilt, punishment is the appropriate reaction to moral wrongdoing.

Now, first, Bennett's talk of "amends" (Bennett 2008, 148) is slightly misleading. To make amends means to compensate for the damage one has caused. But punishment goes beyond mere compensation. However, this detail should not be a problem for Bennett's theory. As a matter of fact, when we feel guilty we do not merely feel the desire to compensate for the damage we have caused, but also to incur some additional burden. If, say, I feel guilty about having killed my neighbour's cat, I will not only get him a new one, but I will also apologise and invite him to dinner. Therefore, at least in this respect, taking the expression of guilt as the point of departure for making sense of punishment is perfectly in order.

Second, and more importantly, it remains unclear why there should not be ways of expressing guilt other than by undergoing penance in form of self-inflicted hard treatment. Bennett is at great pains to show that hard treatment is more than merely a conventional way of expressing guilt or condemnation. His argument goes as follows: When one feels guilty as a result of some wrongdoing, one acknowledges that one no longer has the full dignity that law-abiding citizens have. And while inflicting suffering on well-behaved citizens would be incompatible with their dignity, inflicting it on wrongdoers is appropriate because they have (at least to some extent) forfeited their dignity (Bennett 2008, 116–117). And this is why we should think of hard treatment as an intrinsically logical (rather than merely conventional) way of expressing reprobation

Bennett's argument is, I think, seriously flawed. To see why, it is very important to keep in mind that Bennett's account of punishment is expressivist, i.e. that the expression of guilt or condemnation has no further purpose. We punish wrongdoers not because they deserve it *simpliciter*, but because it is the appropriate expression of our condemnation. Therefore, even if it is tempting, we must be careful not to read Bennett as suggesting that wrongdoers deserve to suffer on account of their having (partly) forfeited their dignity. If Bennett thought that wrongdoers deserved to suffer, then he would (or at least should) have said so

[12] For a complete and very perceptive definition of 'punishment', see Zimmerman (2011, 1–21).

straightaway and made this the basis of a non-expressivist retributive theory of punishment; his expressivist account of punishment would then be obsolete altogether. Bennett, however, is not concerned with desert in its strict sense, but with which behaviour is a symbolically adequate expression of guilt. Now, if guilty people do not deserve to suffer (or if we are agnostic about it), why is inflicting suffering on oneself an intrinsically adequate expression of guilt? Arguably, it would be intrinsically adequate only if one deserved to suffer on account of having forfeited one's dignity in the first place. But Bennett does not seem to argue for this claim. First, if he made this claim, he would give a full-fledged moral argument for retribution in its own right, which would make the expressivist argument superfluous. Second, if he really wanted to put forward a retributive argument for punishment based on desert, he would have to say more on dignity, desert and suffering. We must therefore assume that Bennett is not arguing that wrongdoers *deserve* to suffer. But if they do not deserve to suffer on account of having forfeited their dignity, it remains opaque why it should be intrinsically appropriate to make them suffer on account of their lack of dignity.[13]

Note that I am not disputing that punishment *is* a symbolically appropriate way of expressing moral emotions such as guilt and indignation. There is no point in denying that punishment is a widespread and widely understood way of expressing condemnation. What I am calling into question is Bennett's claim that it is an intrinsically logical means of expressing reprobation. And if Feinberg's original suspicion is correct that the symbolic appropriateness of hard treatment is merely conventional, then we should presumably try to substitute some less ghastly convention for it.[14]

Now, one possibility to consider is that hard treatment is neither intrinsically appropriate *nor* merely conventional. Maybe humans have an innate desire to engage in punitive action against wrongdoers, be it oneself or someone else. If this were the case, punishment would be a natural (rather than conventional or intrinsically logical) expression of guilt or indignation. But even if this were true, more would have to be said on why we should go for this natural way of expressing condemnation rather than for a conventional alternative. It is widely acknowledged that when emotions are expressed in symbolic expressive actions, they may be expressed in many different ways (see e.g. Döring and Peacocke 2002, 82). And since there is arguably a presumption against using painful ways of expressing condemnation, why go for punishment?

So far, I have been mainly concerned with the expressivist account of punishment and have left aside the communicative approach. However, if disapproval or indignation can be *expressed* in ways other than through hard treatment, it is natural to surmise than it can also be *communicated* in ways other than through hard treatment. Communication differs from expression only in that it is directed at an addressee. When we communicate, we do not merely *express* a message (or emotion), we convey it to somebody else. There is, therefore, no reason to think that communicative theories of punishment are any less vulnerable to the conventionalist objection than the expressive account of punishment. If there are non-punitive symbolic languages available through which disapproval and indignation may be *expressed*, then they are also suitable for *communicating* disapproval and indignation.

However, curiously enough, this does not apply to Antony Duff's communicative account of punishment, which is arguably the most prominent communicative theory of

[13] It might be thought that moral disapprobation, e.g. guilt, already includes some notion of desert. Maybe guilt is not possible without the offender possessing some measure of desert. However, I do not think we have to assume such a strong link between desert and moral disapprobation. To be sure, they both have a common basis, namely culpable wrongdoing. But it seems conceptually possible to think of someone as guilty without at the same time implying that he is deserving of punishment.

[14] Unless, of course, one thinks wrongdoers *deserve* to suffer. But this is a claim that must be argued for.

punishment. Duff is, of course, aware of Feinberg's challenge and does not fail to address it. Why then does Duff take hard treatment to be the best way of bringing about repentance and reform in the wrongdoer? Because hard treatment "is a way of trying to focus his attention on his crime. […] As fallible moral agents, we need such penances to assist and deepen repentance" (Duff 2001, 108; see also Duff 2003, 390). Whether this is true or not is a purely empirical question. What is remarkable, however, is that Duff's communicative account of punishment is actually much more than merely communicative. In fact, Duff's labelling of his theory as 'communicative' is slightly misleading. The purpose of punishment is precisely not merely to communicate (i.e. to convey) a message to the wrongdoer, but to bring him to engage in a process of repentance and reform. Therefore, the fact that there are other equally *appropriate* ways of expressing and also of communicating resentment need not worry Duff. For Duff chooses punishment over other ways of communicating blame not because it is more appropriate but because it is (allegedly) more effective in making the wrongdoer understand why he acted wrongly.[15] Therefore, Duff's approach is not open to the conventionalist challenge. It is, however, open to an empirical challenge. If there are non-punitive measures that are better suited to bring about repentance, reform and reconciliation, his case for punishment collapses. I concur with Narayan (1993, 177) and Hanna (2008, 47) that we have good reasons to doubt that Duff can meet the empirical challenge. First, as Narayan rightly observes, hard treatment may be counterproductive and deflect the offender's attention from his wrongdoing. Offenders could experience their punishment as a humiliation and, as a consequence, feel anger and resentment rather than remorse.

Second, even if Duff's claim that hard treatment helps the offender focus on his crime were correct, this is not enough to *justify* hard treatment. For he would also have to show that the intentional infliction of suffering is uniquely suited for this purpose, or at least better suited than possible alternatives. This is highly unlikely. What is it about the intentional infliction of suffering that supposedly makes it better suited to bring about regret and reconciliation than non-punitive means? Particularly, it is unclear why punishment should fare better than non-punitive restorative measures (such as victim-offender dialogues) that are especially designed to raise the offender's awareness of the wrong he has done and to make reconciliation possible.

3 Hard Treatment as Vindication of the Law

In the preceding section, I argued that the conventionalist challenge still stands. In this section, I will further strengthen the conventionalist objection by criticising another argument for hard treatment that has been advanced by expressivist theorists of punishment. It was first put forward by Igor Primoratz and then, at least to some extent, endorsed by Bennett (Primoratz 1987, 1989; Bennett 2006, 291–293, 2011, 287–289).[16] This argument is also supposed to defuse another, very natural objection to the expressivist account of punishment: Why should we express condemnation in the first place? Unlike, for instance, Antony Duff, proponents of expressivism face the problem that they cannot name a purpose

[15] Duff claims, however, that punishment is an "intrinsically appropriate" (Duff 2001, 89) way of bringing the offender to repent. I do not think, though, that this claim accords with Duff's actual justification of hard treatment.

[16] Bennett defends a weaker version of Primoratz' argument. While Primoratz claims that it can justify hard treatment, Bennett thinks it shows "that the state should do *something*", but not necessarily that it should inflict punishment (Bennett (2006, 292), emphasis added).

of expressing reprobation. Punishment simply has no purpose besides being the expression of guilt and indignation. This is a very feeble justification for a practice as unkind as punishment. Even if the conventionalist objection could be met, i.e. even if the symbolics of expressing indignation were not conventional, expressivism may *still* fail to justify hard treatment. For it is not enough to show that 1) indignation cannot be expressed other than through punishment, and that 2) there is some vague *pro tanto* reason for expressing one's appropriate emotions. Expressivists must also argue that this expressivist rationale trumps all other considerations that speak against inflicting harm. This is a formidable challenge given that there are overwhelming reasons against inflicting suffering.[17] With the exception of retributivists who think that wrongdoers just *deserve* to suffer, legal theorists usually consider the infliction of suffering a last resort that stands in need of special justification. And it does not seem that the desire or right to express one's emotions could serve as such a justification. If you find out that your partner has been cheating on you, would it be okay to express your rightful anger by, say, beating up your partner or by locking her or him up for a couple of months or years? Or would it be okay, upon learning that you have been fired, to express your anger by spontaneously kicking the cat that happens to be purring around your feet thus causing her serious harm? Presumably not.[18] Rather, you should try to find alternative ways of expressing your anger, and if – as we stipulate – there aren't any, you should contain yourself and refrain from expressing your anger altogether. It seems, then, that the principle not to inflict suffering has priority over the right to express one's emotions. This seems particularly true in the legal context, given that the harm involved in legal punishment is often extremely severe. This objection has also been advanced by David Boonin (2008, 172–176). Interestingly, however, Boonin raises this objection against Duff's account, which is precisely not expressivist but communicative. So while Boonin's objection is generally valid, it cannot be raised against Duff. It is true, though, that Duff faces a similar problem. He has to explain why the aim of bringing about the three 'R's (repentance, reform, reconciliation) is so important as to justify the infliction of suffering (cf. Hanna 2008, 44). However, Duff seems much better equipped to overcome this objection than Bennett. While the value of expressing one's reprobation seems negligible, it is easier to see how the value of regret, reform and reconciliation could override the presumption against suffering.

Let me now turn to Primoratz and the idea of hard treatment as vindication of the law. Like Bennett, Primoratz favours a purely expressivist justification of punishment (Primoratz 1987, 217). He then goes on to argue, however, that we must express our condemnation through punishment in order to vindicate the law[19]: "If actions of a certain kind do not revoke such a response [i.e. blame and punishment] from society, that goes to show that no rule prohibiting such actions is accepted as a valid standard of behaviour" (1987, 217).

Primoratz and Bennett may be right that the flouting of norms must provoke public condemnation in order for the laws to be valid. Irrespective of the ontological question of whether a norm would still be a norm if it could be flouted without provoking condemnation,

[17] I here agree with the anonymous referee who urged me to consider this independent argument against expressivism.

[18] I concede, though, that there might be other reasons why we would object to such behaviour. Maybe, the infliction of harm should be a privilege of the state, and it is *therefore* not okay for you as an individual to express your anger through inflicting harm. My examples, however, are not supposed to be ultimately conclusive.

[19] Whether a justification of punishment that invokes an 'in order to' still counts as "intrinsic expressionism" (Primoratz (1989, 200)) is doubtful, though. But I do not want to pursue this question here. See Primoratz (1989, 203).

it is certainly true that such indifference would be very awkward, to say the least. Let us therefore grant that for there to be valid laws in the first place, breaches of the law must be publicly condemned. However, Primoratz must show more than this. He must explain why condemnation must take the form of punishment, and I do not think he succeeds. Let us look at Primoratz' arguments in favour of this claim:

In one of his arguments, Primoratz confuses *explanans* and *explanandum*: "[I]f there are to be rights sanctioned by the criminal law, if some acts are to be offences, if there is to be criminal law at all – there has to be punishment" (1987, 218)). This, I think, is perfectly true, but sadly by definition so. *Of course* there would be no criminal law if there were no punishment. But this is because criminal law is by definition the legal institution that metes out punishment.[20] And clearly, the justification of punishment should precede the justification of criminal law. The reason why there should be such a thing as criminal law is that there should be punishment, not vice versa. We must therefore come up with an independent rationale for punishment.

Elsewhere, Primoratz contends that "merely verbal condemnation is not likely to reach its immediate addressee and to be fully understood by him. […] So if society's condemnation of their misdeeds is really to reach [the criminals], if they really are to understand how wrong their actions are, it will have to be translated into the one and only language they understand", which is punishment (1989, 200). Note that the focus of the argument has now shifted. Punishment now seems to have the function that Duff thinks it should have, namely that of making the offender understand that he acted wrongly. But this goes beyond merely reaffirming or vindicating the law in order for it not to be "empty". And it simply does not seem that the offender has to understand how wrong his action was in order for a prohibition to be valid.

Finally, Primoratz reckons that laws must appear empty to the victims if their infringement provokes merely verbal condemnation: "*they* would surely see purely verbal condemnation of crime, however public and solemn, as half-hearted and unconvincing", especially as the state "would be seen as desisting from activating its apparatus of force and coercion, which is surely one of its essential, defining features"(1989, 200). Primoratz' reasoning is fallacious in two ways: First, the state need not desist from making use of coercion. If retributive justifications of punishment fail, the state may mete out punishment on non-retributive, i.e. forward-looking grounds. This need not be half-hearted and may be a very forceful way of showing how much it cares about the safety and wellbeing of its citizens. Second, victims will see purely verbal condemnation as half-hearted only if there is the social convention of punishing wrongdoers in the first place. If it were customary to express reprobation in different purely non-punitive ways, there is no reason to assume that victims would consider this inappropriate. And as I argued above, we have good reasons to assume that other ways of expressing indignation are available. Bringing these two objections together: If the state reacted to breaches of the law by expressing condemnation in appropriate, yet non-punitive ways and by making an effort to prevent further crimes, why should we (or victims) think that legal prohibitions are empty or invalid?

What can we conclude from this? While Primoratz (and Bennett) may be right that there must be *some* public reaction to breaches of the law in order for laws not to be meaningless, hard treatment is dispensable. There may be good reasons for expressing reprobation, but there are no good reasons for expressing it through punishment. The conventionalist objection still stands: why go for punishment if there are alternative ways of expressing

[20] This is more obvious is other languages. Think of ‚*droit pénal*', ‚*diritto penale*', '*derecho penal*', '*Strafrecht*', and so forth.

reprobation? If forgoing punitive measures would undermine the normative status of norms, we would have an independent reason to inflict hard treatment. But it does not.

We must therefore conclude that the expressivist attempt to justify punishment as the only appropriate emotional reaction to wrongdoing fails.

Notice that my critique of expressivism does not entail that punishment should not be expressive. I have argued that expressivism cannot provide a justification of punishment. If, however, punishment can be justified on other (e.g. instrumentalist) grounds, it seems that we ought to punish in ways that are *also* expressive of our reprobation.[21] It is not implausible to assume that some modes of punishment express our moral emotions better than others. If they are equal in all other relevant respects, we should opt for the mode of punishment that is the best expression of our indignation. That is, the expressivist rationale alone might not provide a conclusive justification of punishment, but it could specify how punishment should be meted out if called-for on other grounds. Similar considerations hold for the communicative account. Even if the communicative rationale alone cannot justify punishment, we may use punishment – if justifiable on other grounds – in such a way that it fulfils communicative purposes. Thus, the expressivist or communicative function of punishment can be of interest to legal scholars even though it offers little in the way of a conclusive justification of punishment. We must bear in mind, however, that expressive or communicative considerations will always be parasitic upon an independent (e.g. instrumentalist) justification of punishment.

Having shown that expressivism fails to justify legal punishment, I will now proceed to examine whether punishment may be understood and possibly justified as a primitively intelligible entailment of retributive emotions. Unlike in the theories of Bennett and Primoratz, the urge to inflict retribution would then be purposive rather than *l'art pour l'art*.

4 Is Retributive Behaviour Primitively Intelligible?

We saw that emotions can be expressed in many different ways and that hard treatment can therefore not be justified as the intrinsically logical expression of indignation or guilt. This objection, however, does not apply to emotions-based actions that are a means to an end, i.e. when we act on emotions in order to bring about a very specific state of affairs. If, say, I envy my neighbour for his beautiful cat, I will try to change the world in such a way that he no longer has a cat that is more beautiful than mine. To be sure, I may do this in various ways. But still, the range of options is strongly limited because my action must serve a very specific purpose. Now, if the infliction of hard treatment on wrongdoers is of this kind, retribution may be made sense of. For if we should express our guilt (or indignation) by meting out punishment because we should bring about the situation in which the wrongdoer suffers, the conventionalist objection no longer applies.

Purposive and non-purposive expressions of emotions may easily be confounded. This is problematic as the prospects of justifying punishment vary significantly depending on how we construe the expression of moral emotions such as guilt and indignation. Let me quote one remark by Bennett that shows how different ways of expressing an emotion can be confused:

> Emotions issue in characteristic forms of behaviour: as John Skorupski observes, fear disposes to flight, anger to attack, grief to mourning. This behaviour is not purposive

[21] I am indebted to an anonymous referee for pointing this out. See also Brooks (2012, 107).

> in the sense that we intend it for some further goal. In order to understand or explain it we need not ask for what end it is carried out. Rather it is appropriate as the issue of the emotion itself. We do not regard the behaviour in which mourning consists to be rationally suspect merely because it does not seem to be carried out for any further purpose: we are grieving, and this is what mourning consists in. The same goes for the expression of blame (Bennett 2002, 151).[22]

Bennett fails to account for a crucial difference between expressing grief through mourning and expressing fear through fleeing. *Pace* Bennett, fleeing is, unlike mourning, quite obviously not without further purpose. And it will make a big difference whether we construe expressive punishment as analogous to the fear/flight dyad or to the grief/mourning dyad.

Bennett, to be sure, construes expressive punishment in analogy to the grief/mourning dyad, and I argued above that such a purely expressive account of punishment is vulnerable to the conventionalist objection. Interestingly, however, these objections may be countered if we conceive of expressing reprobation through punishment in analogy to the fear/flight dyad.

The main objection – namely that there are suitable ways of expressing indignation other than through punishment – would then no longer apply: If we are in a state of fear and want to bring about a safe situation, the range of available options is limited. I agree that in such a situation flight is a primitively intelligible action.

At this point, it will be instructive to draw on Sabine Döring's remarks on the 'ought-to-be' and the 'ought-to-do' that emotions involve. Emotions are evaluative states. They represent the world as being in a certain way. As evaluations, emotions involve an ought-to-be. When we are disappointed, for example, we feel that the course of events ought to have been different. Or when we are pleased, we feel that the world ought to be just as it is. Sometimes, the evaluative ought-to-be does not entail a normative ought-to-do. If we are, say, nostalgic about Paris in the 1920s, we feel that the world ought to be like back in the days. But since there is no way we can bring back the old times, no ought-to-do follows. Often, however, the ought-to-be does entail an ought-to-do, as the fear/flight dyad shows: We feel that we ought to *be* safe, and so we judge that we ought to *flee*. The ought-to-do can then be rationalised in terms of the ought-to-be. Emotions can provide us with (at least *prima facie*) reasons for action (Döring 2007).

Goldie's notion of primitive intelligibility is, I think, best understood in terms of the ought-to-be/ought-to-do distinction. A desire, or action, is primitively intelligible if it is entailed by the ought-to-be that the relevant emotion contains. It is difficult to see how actions that spring from emotions could otherwise be primitively intelligible.

Could punishment possibly be made sense of in this way? Does indignation or guilt entail the ought-to-do of inflicting suffering on the wrongdoer? Is this retributive behaviour primitively intelligible? Or to put the question another way: Does inflicting suffering on wrongdoers "change the world in such a way that it fits the emotion" (Döring 2007, 385)?

Such an approach to punishment, unlike Bennett's, would be genuinely retributive. In the purely expressivist justification of punishment, the logic of desert has been replaced by the logic of expression. If, however, emotions entail an "ought-to-inflict-punishment",

[22] The reference is to Skorupski (1993, 136). Bennett's talk of blame as an emotion strikes me as odd, though. I do not think you can be in the emotional state of blame. Blame may refer to an action (e.g. to a speech act of the kind "I blame you for x.") or to a purely cognitive state (the belief that somebody is responsible). Emotions that represent somebody as blameworthy are, for instance, indignation, guilt, anger or resentment. On anger, see Roberts (2003, 202–221).

wrongdoers can properly be said to *deserve* punishment. This second emotions-based approach to punishment might therefore be a better way of making sense of the idea of retribution.

An emotional approach to punishment roughly along these lines has been advocated, e.g., by Jeffrie Murphy and Michael Moore. Murphy argues that retribution can be justified in terms of what he calls retributive hatred. Retributive hatred is the appropriate emotional reaction to serious wrongdoing and issues in punitive behaviour. The idea of retribution can thus be understood in terms of an appropriate retributive emotion: "If hate is sometimes justified, then the desire to hurt another must sometimes be justified" (Murphy 1988, 94). While Murphy focuses on the *victim's* emotional reaction to wrongdoing, Moore's justification of retribution – like Bennett's – rests on what the *wrongdoer* ought to feel, namely guilt. Moore maintains, maybe rightly, that our "emotions are our main heuristic guide to finding out what is morally right" (Moore 1987, 189; see also 201). He then goes on to argue that our "feelings of guilt […] generate a judgement that we deserve the suffering that is punishment" (1987, 215).[23] And since the feeling of guilt is generally a reliable guide to moral truths, retributive punishment is morally warranted.

In what follows, I will not discuss Murphy's or Moore's account in particular. Their accounts have been convincingly criticised elsewhere (see e.g. Duff 2001, 23–27). Rather, I want to assess in more general a fashion whether the desire to inflict retribution on wrongdoers can be accounted for by looking at the emotions that give rise to this desire.

As I said at the outset, it is not always obvious whether the behaviour an emotion triggers is primitively intelligible or not. Many might be tempted to think that retributive behaviour is as natural a consequence of resentment as flight is of fear. If this were true, retributive behaviour could be made sense of and possibly justified in terms of retributive emotions. However, this first impression is false. Unlike the desire to flee when in danger, the desire to inflict punishment when in a state of indignation is not primitively intelligible, at least not in any obvious way. The suffering of the wrongdoer just does not seem to change the fact that he has consciously flouted moral reasons, and it is precisely the conscious flouting of moral reasons that calls for indignation. By the same token, inflicting pain on oneself does not reduce the guilt on has incurred by committing a wrong. The suffering of the wrongdoer just does not seem to bear a relation to the fact that he has flouted moral reasons. The fact that we are used to seeing resentment and guilt issue in punitive behaviour should not mislead us into believing that there must be an intrinsically logical connection.

While there is no obvious relation between indignation and the desire to punish, there might well be less obvious ways in which the desire to punish turns out to be a perfectly intelligible entailment of indignation. In order to find out whether there is a primitively intelligible, if not obvious, link between indignation and the urge to punish, I will consider Robert Nozick's, Herbert Morris's and Jean Hampton's defences of retributive punishment. As they all argue that punishment serves to correct an imbalance, their theories might give us hints as to how punishment may be a way of making the world fit our retributive emotions. Of course, I am focussing here only on a limited class of retributive theories of punishment, namely on those that may help us make sense of our retributive emotions. Other retributivists

[23] On the same page, he states that "to feel guilt *is* to judge that we must suffer." (emphasis added). This claim, however, is patently false. First, emotions are cognitions, but they are not judgements in the strict sense. Unlike judgements, they do not enter inferential relations. (see Döring (2007)). Second, the emotion of guilt (or indignation, resentment, etc.) does not directly represent the wrongdoer as deserving suffering. Rather, it represents the wrongdoer as having culpably committed a moral wrong. The judgement (or cognition) that the wrongdoer deserves to suffer is not part of the emotion itself.

prefer to base retributivism on a more primitive notion of desert (e.g. Kershnar (1995, 2000) and Tasioulas (2006, 297)). I am not dealing with this simple notion of desert in this section of my essay, though I will discuss it in the last section.

Nozick observes that wrongdoers have "become disconnected from correct values" (1981, 374). The wrongdoer has flouted moral reasons; he has chosen "not to give them effect in his life" (1981, 375). On Nozick's account, punishment has the purpose of reconnecting the wrongdoer with the correct values. Note, however, that this does not mean that punishment is supposed to improve the character of the wrongdoer or to make him repent. Punishment aims *not* at "recognition of the correct value" by the wrongdoer or at its internalisation for future action (1981, 375). But what does it then mean to reconnect the wrongdoer with the correct values? Nozick contends that the rationale behind the concept of retribution is that it "gives significant effect in his [the wrongdoer's] life to correct values" (1981, 387). The wrongdoer failed to give moral considerations effect in his life (by flouting them), but they can retroactively be given effect through punishment. For it is trivially true that the correct values have an effect on the wrongdoer if we punish him for flouting them. Thus, Nozick reckons, by inflicting punishment on wrongdoers "the imbalance is rectified" (1981, 384).

Can Nozick's defence of retribution make sense of our desire to punish when we are in a state of indignation? Is inflicting punishment really a suitable and intelligible way of soothing our indignation? It does not seem so. It is certainly correct that our indignation is directed at the fact that the wrongdoer has flouted the correct values. But the problem is that there is simply no way of undoing his misdeed. Nozick correctly stresses that retributive punishment aims not at the recognition and internalisation of the correct values on the part of the wrongdoer. Retribution is backward-looking; we are upset at the wrongdoer's *past* flouting of the correct values. But why should inflicting punishment be a correction of this moral imbalance? Why not give effect to the correct values in the wrongdoer's life by blaming him or by rewarding him? Nozick falsely assumes that just *any* effect of the correct values on the wrongdoer's life can correct the imbalance that his flouting of the correct values has caused. But clearly, this will not do. What we want is that moral values have a very specific effect on people's lives: They should acknowledge them and then act according to them out of duty. This is the kind of effect that moral reasons should have on people's lives. And once somebody failed to give moral considerations this effect, there is no way of retroactively compensating for it. To make the same point by drawing on the ought-to-be/ought-to-do distinction: When we are in a state of indignation, we feel that somebody ought to have acted morally. It is the flouting of moral reasons that gives rise to our indignation. But retroactively giving moral considerations just *some* other random effect in the wrongdoer's life is no way of undoing his flouting of them. Therefore, no ought-to-punish follows form the ought-to-have-acted-morally. Indignation is more akin to grief than fear in that there is no way of making the world fit the emotion.

But let us also consider Herbert Morris's approach (Morris 1968). Maybe he can help us make sense of our retributive emotions. On his account, punishment has the purpose of rectifying the balance of benefits and burdens in society. Punishment is, obviously, meted out when laws are infringed upon. Laws have the function of providing goods that benefit everybody, such as security and liberty. For such a system to work, everybody has to incur the burden of complying with the law, often against one's will. In Morris's picture, then, a law-breaker is a free-rider who benefits from the legal system without himself paying his fair share. Breaking the law gives one an unfair advantage over one's law-abiding fellow citizens, and punishing the free-riders is a way of restoring the fair balance of benefits and burdens.

Can Morris' defence of retribution make the desire to inflict suffering on wrongdoers intelligible? I do not think so. To be sure, if we were indignant at the unfair distribution of benefits and burdens that results from a breach of law, inflicting some harm on wrongdoers would be primitively intelligible. In this case, there would indeed be a way of changing the world in such a way as to make it fit our emotion. Unfortunately, however, Morris seems to miss the point of retribution. The problem is well summarised by Jean Hampton: "The idea that punishment is simply the taking away of the *advantages* which rapists or murderers have by virtue of being unrestrained presupposes that […] we object to them only because, if performed by everyone, they would be collectively harmful" (Hampton 1988, 116, original emphasis).[24] Morris simply fails to acknowledge that we object to misdeeds because they are *inherently* wrong. Irrespective of whether a murderer has gained a comparative advantage or not, we feel that he ought to be punished. Morris's benefits-and-burdens approach just does not get at the heart of the matter. From this follows that Morris's approach cannot help us make sense of our retributive desires. Nozick is right: We are indignant at wrongdoers because they have flouted moral values. To confer a disadvantage on wrongdoers does nothing to rectify the situation.

Let me finally turn to Jean Hampton's defence of retributivism. On her account, too, retributive punishment is meant to correct an imbalance. Hampton thinks of her defence of punishment as "expressive" (1992) and "communicative" (1988, 123). This is misleading, though. On her account, the purpose of punishment is neither merely to express something nor to communicate something to somebody. Rather, punishment has the purpose of *showing* something. Her defence of retribution should therefore, I think, be referred to as *demonstrative*.

Hampton claims that by flouting moral norms, offenders assert moral superiority over their victims. By disregarding morality, they elevate themselves over others and thus deny the truth that all human beings are of equal moral worth. Retributive punishment has the purpose of correcting this false claim about the victims' relative worth. Notice, now, that it is not enough to merely give "a ticker-tape parade after the crime to express our commitment to his value" (1988, 128). Also, the purpose of punishment is not to communicate the moral truth to those who denied it. Hampton makes clear that we should mete out punishment even when we can be sure that the message will go unheard (1988, 131). Thus, punishment has neither an expressive nor a communicative function. Rather, the purpose of retributive punishment is to nullify the evidence of the alleged inferiority of the victim: "The punishment is therefore a second act of mastery that negates the evidence of superiority implicit in the wrongdoer's original act." (1988, 129) By punishing the wrongdoer, we *show* that he is in fact not superior to his victim, precisely because we have been able to subdue him. It is in this sense that punishment corrects an imbalance that is established by a crime.

I think we can agree with Hampton that by committing a wrong, offenders often assert or imply their superiority over their victims. They assert or imply that morality does not apply to them, that their victims have no moral claims on them. This, I think, is the reason why we react to crimes with indignation. However, it does not seem that subjecting the offender to punishment helps the problem.

First, by committing a wrong, offenders might assert their moral superior worth, but their assertion is usually not evidence of it. If I recklessly commit a murder out of greed, would anyone take this to be evidence of my superior moral worth? Obviously not. It is merely

[24] For further critical discussions of Morris's approach, see e. g. Dolinko (1991); Mackie (1982, 5).

evidence that I falsely take myself to have special moral privileges or that I am unimpressed by moral considerations. And, most importantly, it is evidence that I was *able* to commit the murder. Second, and relatedly, by subjecting the wrongdoer to punishment, we do not show that wrongdoer and victim are of equal moral status. Rather, we show that we are *able* to subject the wrongdoer to punishment: „The wrongdoer can't take her crime to have established or to have revealed her superiority if the victim is *able* to do to her what she did to him."(1988, 129; emphasis added) The evidence we nullify by meting out punishment is not evidence of the moral superiority of the wrongdoer. No such evidence was established by the crime in the first place, and "mastering" the wrongdoer does not show that he is of equal moral status as the victim. The evidence that we really nullify is the evidence of the wrongdoer's superior power. By subjecting the wrongdoer to punishment, we show that he cannot commit a crime with impunity. Hampton observes: "To inflict on a wrongdoer something comparable to what he inflicted on the victim is to master him in the way that he mastered the victim." (1988, 128) This is true. However, mastery is a sign of superior power, not of superior moral worth. Inflicting punishment on wrongdoers is of course an effective means of showing that one is more powerful than the wrongdoer. But it is not a way of demonstrating the equal moral value of all human beings.[25]

It does not seem, then, that Hampton's theory can help us make sense of retribution and our retributive emotions. When we are indignant at a wrong committed, we resent the offender's intentional flouting of moral values. This is what appals us. To show that we are more powerful than the wrongdoer, that we are able to subdue him, does not address the problem.

We may conclude that the desire to inflict suffering on wrongdoers is *not* primitively intelligible. *A fortiori*, punishment cannot be morally justified in terms of our emotional reaction to wrongdoing. Unlike fleeing when in fear, punishing wrongdoers out of indignation just does not help the problem. While guilt or indignation may be appropriate and ethically called-for responses to wrongdoing, the desire to mete out punishment remains completely unintelligible. Therefore, we may state quite generally that the prospects of justifying retribution in terms of the emotions that give rise to retributive desires are fairly bleak. It remains wholly opaque why moral emotions such as indignation, guilt or resentment issue in punitive behaviour, let alone why they *should* issue in it. And the burden of proving that the urge to inflict suffering can be rendered intelligible is clearly on the advocates of retribution. Note also that emotions can equip us at best with *prima facie* reasons for actions, i.e. with reasons that might be outweighed by other considerations. Even if our desire to inflict punishment were intelligible in terms of our retributive emotions, we would still have to ask whether the reasons that these emotions provide are overriding. However, since our punitive urge is not primitively intelligible, we can rule out the emotional approach to punishment straightaway. We do not even get to the point of asking whether the reasons that our emotions provide are overriding or not.

It seems a mystery, then, why we have the desire to inflict retribution upon wrongdoers when we are in a state of indignation or retributive hatred. In the remainder of my essay, I would like to consider another way making sense of retribution, which, however, cannot serve as a justification of punishment.

[25] For a similar critique of Hampton, see Gert et al. (2004). Also, we might ask: Why is it so important to show that all human beings have equal moral worth? Is it really so important as to justify the systematic infliction of suffering?

5 How, then, can Retributive Behaviour be Accounted for?

To my mind, the most promising approach to understanding retribution was hinted at by John Leslie Mackie, whom I quoted at the outset of my paper. Mackie conjectured that there may be an evolutionary explanation for the tendency to engage in punitive behaviour (Mackie 1982, 1991). As retributive behaviour is very likely to be conducive to social cooperation, it is only natural to surmise that it may have been naturally selected. This conjecture has been corroborated in a recent study by Robert Boyd et al. (2003). They have shown how altruistic punishment can naturally evolve even in large groups. A similar naturalist approach to retribution has been proposed by Tamler Sommers (2009).

If the lust for retribution cannot be accounted for within the confines of philosophy, it is only natural to look elsewhere for a solution of the "paradox of retribution". Precisely because the desire to inflict punishment on wrongdoers is not primitively intelligible, an evolutionary explanation is appealing. The evolutionary explanation of retribution will not be discussed in any more detail here. However, let me make some remarks on the philosophical significance that a plausible evolutionary account of retribution may have. For some might think that such an explanation of the practice of retribution has no implications whatsoever on the moral worth of retribution. I do not think this is correct, for the explanation may serve as what Folke Tersman has dubbed a debunking explanation: "Consider a fact *F* that is offered as evidence for theory *T*. A debunking explanation of *F* is an explanation that does not entail that *T* is true or significantly likely" (Tersman 2008, 395). It is a widespread intuition that wrongdoers deserve to suffer (fact *F*). This intuition is then often taken to be evidence for the correctness of the claim that wrongdoers deserve to suffer (theory *T*). If, however, we can give an explanation of this intuition that does not speak in favour of the truth of the intuited claim, the intuition is debunked. An evolutionary explanation of our retributive urges could serve precisely as such a debunking explanation. It would show that the idea of retributive desert is, *contra* John Tasioulas, not "a basic norm of justice" (Tasioulas 2006, 297). Rather, the norm could be accounted for in terms of the increase of fitness that it brought about. The reason, then, why we think that offenders should suffer is that the retributive urge once served the purpose of sustaining cooperation and thus of increasing individual fitness. And obviously enough, the increase in fitness that retributive inclinations may have brought about in the past does not speak in favour of the truth of the moral principle of retribution. Similar evolutionary arguments designed to debunk moral intuitions have recently been put forward by Peter Singer (2005) and Sharon Street (2006). The general idea is that if we can explain moral intuitions in terms of their evolutionary genesis, we have little reason to believe that they are true. Making these evaluative judgements, and acting on them, was once conducive to reproductive success. This is why we have come to have these intuitions in the first place. But the property of increasing reproductive success is unrelated to the property of being true. If we have these intuitions solely because they increased our fitness, it would be a huge coincidence if they were also true.

Thus, evolutionary theory may *explain* why indignation often issues in punitive behaviour, even though this behaviour is not primitively intelligible. However, such an explanation of retribution would not *vindicate* retributivism. On the contrary, a naturalistic explanation would cast serious doubts on the truth of retributivism. I have not discussed more straightforwardly desert-based accounts of retribution in any detail here, and this would be beyond the scope this essay. However, if retributivists do not provide an independent rationale of why wrongdoers deserve to suffer, their retributivism will be open to a naturalistic objection of this kind.

To be sure, a debunking explanation of our retributive urges would not imply that inflicting punishment is wrong. What it would show, though, is that there is presumably no purely retributive justification of punishment. Even the concept of retribution, which is

backward-looking, could be explained in terms of forward-looking considerations: We have acquired the urge to engage in retribution because it was conducive to cooperation. While the retributive reflex may have served its purpose in the past, it is obsolete today: Arguably, our forward-looking concerns – compliance with the norms of cooperation – are best served if we draw on the abounding criminological literature rather than on our coarse-grained retributive reflexes.

Of course, the evolutionary explanation should be taken with a grain of salt. However, absent better explanations of why wrongdoers deserve punishment, evolutionary approaches may be a promising alternative to consider.

References

Bennett C (2002) The varieties of retributive experience. Philos Q 52(207):145–163

Bennett C (2006) State denunciation of crime. J Moral Philos 3(3):288–304

Bennett C (2008) The apology ritual: a philosophical theory of punishment. Cambridge University Press, Cambridge

Bennett C (2011) Expressive punishment and political authority. Ohio State J Crim Law 8(2):285–318

Boonin D (2008) The problem of punishment. Cambridge University Press, Cambridge

Boyd R, Gintis H, Bowles S, Richerson PJ (2003) The evolution of altruistic punishment. Proc Natl Acad Sci 100(6):3531–3535

Brooks T (2012) Punishment. Routledge, London

Ciocchetti C (2009) Emotions, retribution, and punishment. J Appl Philos 26(2):160–173

D'Arms J, Jacobson D (2000) The moralistic fallacy: on the 'appropriateness' of emotions. Philos Phenomenol Res 61(1):65–90

Dolinko D (1991) Thoughts about retributivism. Ethics 101(3):537–559

Döring S (2007) Seeing what to do: affective perception and rational motivation. Dialectica 61(3):363–394

Döring S, Peacocke C (2002) Handlungen, gründe und emotionen. In: Döring S, Mayer V (eds) Handlungen, gründe und emotionen. Akademie Verlag, Berlin, pp 81–104

Duff A (2001) Punishment, communications, and community. Oxford University Press, Oxford

Duff A (2003) Punishment, communication and community. In: Matravers D, Pike J (eds) Punishment, communication and community. Routledge, London, pp 387–407

Feinberg J (1965) The expressive function of punishment. Monist 49(3):397–423

Gert HJ, Radzik L, Hand M (2004) Hampton on the expressive power of punishment. J Soc Philos 35(1):79–90

Goldie P (2000) The emotions: a philosophical exploration. Oxford University Press, Oxford

Hampton J (1988) The retributive idea. In: Murphy JG, Hampton J (eds) The retributive idea. Cambridge University Press, Cambridge, pp 111–161

Hampton J (1992) An expressive theory of retribution. In: Cragg W (ed) An expressive theory of retribution. Franz Steiner Verlag, Stuttgart, pp 1–26

Hanna N (2008) Say what?: A critique of expressive retributivism. Law Philos 27(2):123–150

Hart HLA (1963) Law, liberty, and morality. Stanford University Press, Stanford

Holroyd J (2010) The retributive emotions: passions and pains of punishment. Philos Pap 39(3):343–371

Hursthouse R (1991) Arational actions. J Philos 88(2):57–68

Kerhsnar S (2000) A defense of retributivism. Int J Appl Philos 14(1):97–117

Kershnar S (1995) The justification of deserved punishment via general moral principles. South J Philos 33(4):461–484

Kleinig J (1991) Punishment and moral seriousness. Israel Law Rev 25:401–421

Mackie JL (1982) Morality and the retributive emotions. Crim Justice Ethics 1(1):3–10

Mackie JL (1991) Retributivism: a test case for ethical objectivity. In: Feinberg J, Gross H (eds) Philosophy of law, 4th edn. Wadsworth, Belmont, pp 676–684

Moore MS (1987) The moral worth of retribution. In: Schoeman F (ed) The moral worth of retribution. Cambridge University Press, Cambridge, pp 179–219

Morris H (1968) Persons and punishment. Monist 52(4):475–501

Murphy JG (1988) Hatred: a qualified defense. In: Murphy JG, Hampton J (eds) Hatred: a qualified defense. Cambridge University Press, Cambridge, pp 88–110

Narayan U (1993) Appropriate responses and preventive benefits: justifying censure and hard treatment in legal punishment. Oxf J Leg Stud 13(2):166–182

Neu J (2010) An ethics of emotion? In: Goldie P (ed) An ethics of emotion? Oxford University Press, Oxford, pp 501–517

Nozick R (1981) Free will. In: Nozick R (ed) Free will. Harvard University Press, Cambridge, pp 291–398
Nussbaum M (1993) Equity and mercy. Philos Public Aff 22(2):83–125
Primoratz I (1987) The middle way in the philosophy of punishment. In: Gavinson R (ed) The middle way in the philosophy of punishment. Oxford University Press, Oxford, pp 193–220
Primoratz I (1989) Punishment as language. Philosophy 64(248):187–205
Roberts RC (2003) Emotions: an essay in aid of moral psychology. Cambridge University Press, Cambridge
Singer P (2005) Ethics and intuitions. The J Ethics 9(3–4):331–352
Skillen AJ (1980) How to say things with walls. Philosophy 55(214):509–523
Skorupski J (1993) The definition of morality. R Inst Philos Suppl 35:121–144
Smith M (1998) The possibility of action. In: Bransen J, Cuypers SE (eds) The possibility of action. Kluwer, Dordrecht, pp 17–41
Sommers T (2009) The two faces of revenge: moral responsibility and the culture of honor. Biol Philos 24(1):35–50
Street S (2006) A Darwinian dilemma for realist theories of value. Philos Stud 127(1):109–166
Tasioulas J (2006) Punishment and repentance. Philosophy 81(02):279–322
Tersman F (2008) The reliability of moral intuitions: a challenge from neuroscience. Australas J Philos 86(3):389–405
von Hirsch A (1993) Censure and sanctions. Clarendon, Oxford
Zimmerman MJ (2011) The immorality of punishment. Broadview Press, Peterborough