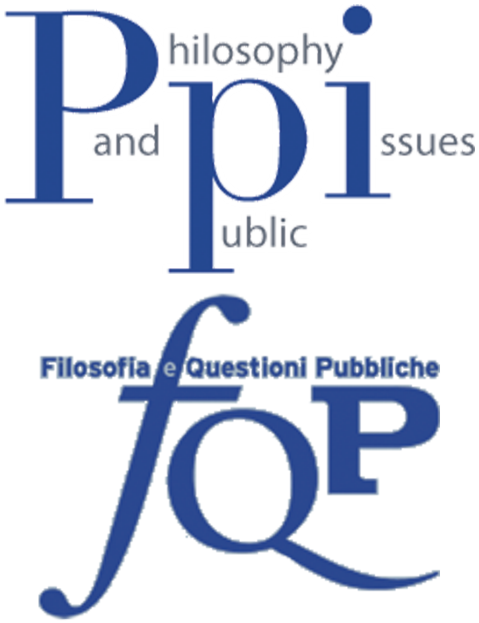
SYMPOSIUM

THE PHILOSOPHY OF PUNISHMENT



CRIMINAL LAW AND THE INTERNAL LOGIC OF PUNISHMENT

BY

KATRINE KRAUSE-JENSEN & RAFFAELE RODOGNO

 Philosophy and Public Issues (New Series), Vol. 5, No. 1 (2015): 135-171
Luiss University Press
E-ISSN 2240-7987 | P-ISSN 1591-0660

[THIS PAGE INTENTIONALLY LEFT BLANK]

Criminal Law and the Internal Logic of Punishment

Katrine Krause-Jensen & Raffaele Rodogno

We argue that punishment has an essentially retributive core that carries its own retributive type of logic or reasons. In particular, we show that punishment is something that we understand as in principle always being assessable in terms of deservingness and that this is ultimately to be understood in terms of moral culpability and nothing else. These features make up what we call the internal logic of punishment. The practice of punishment, however, can also be assessed with a logic that is external to it. What this consists in is first and foremost determined by the aims and constraints of the punishing agent. For the modern liberal state these are typically understood in terms of deterrence, incapacitation, rehabilitation, and, arguably, the expression of condemnation. The idea that punishment has its own internal logic has a number of consequences with regard to criminalization, to the extent, that is, that the latter involves punishment. For one, purely instrumentalist justificatory accounts of punishment will not work as they fail properly to consider the retributive core of punishment. Next, we consider what follows from the fact that by inflicting punishment, the state takes it upon itself to mix these two logics, the internal and the external, together. In particular, we bring forward some tensions that arise when the state mixes

Philosophy and Public Issues (New Series), Vol. 5, No. 1 (2015): 135-171
Luiss University Press
E-ISSN 2240-7987 | P-ISSN 1591-0660

the internal logic of punishment with certain modern, liberal aims and constraints that are external to punishment.

I

The Question and its Method

The criminal law, with its emphasis on punishment, is generally assumed to perform a number of functions. Deterrence and incapacitation are the most obvious ones. One way to envisage the criminal law is as an attempt to regulate the behaviour of those who inhabit its territory through deterrence and incapacitation. State punishment is the particular form of deterrence and incapacitation that behaviour regulation takes when handled by the criminal law. State punishment, however, is not the only mode of deterrence and incapacitation. The state may for example impose non-punitive sanctions in order to deter certain forms of behaviour, or restrain a person to an isolated space in order to disable her from spreading a dangerous disease. Given this fact, one legitimate question arises: is the criminal law simply another mode for the regulation of behaviour in the hands of the state or does the criminal law, with its emphasis on punishment, have a distinctive character or distinctive aims? What answer one gives to this question is important insofar as it will favour or disfavour answers to another urgent question: what kinds of conduct should be subject to the distinctive mode of control that is the criminal law?[1]

[1] R.A. Duff, L. Farmer, S.E. Marshall, M. Renzo, and V. Tadros "Introduction," in R.A. Duff, L. Farmer, S. E. Marshall, M. Renzo, and V. Tadros (eds.), *The Boundaries of the Criminal Law* (Oxford: Oxford University Press 2010), p.6.

Most broadly, the relevant theoretical landscape can be divided between, on the one hand, *instrumentalists* who see the criminal law as *just* another tool for the regulation of behaviour in the hands of the state with no distinctive character or aims of its own; and, on the other, *non-instrumentalists* who take it that the criminal law, with its emphasis on punishment, does have distinctive character and/or aims. Of course, a more fine-grained curving of the conceptual space brings forward important distinctions that somewhat soften the contrast between instrumentalists and non-instrumentalists. This is for example achieved by hybrid views that admit both instrumentalist and non-instrumentalist elements such as Duff[2] and, to a lesser extent, by non-consequentialist, instrumentalist views such as Tadros.[3] In this paper, however, we concentrate on what separates instrumentalists from non-instrumentalists rather than on what unites them, and what does separate them in the end is their respective stance on the question of the distinctiveness of the criminal law.

Note that though the question sounds descriptive, many philosophers are inclined to read it in normative terms. In other words, even if the question asks what, if anything, *is* distinctive of the criminal law, philosophers typically understand it as asking what *ought* to be considered as distinctive or, better, what would be distinctive of the criminal law on an *ideal* account of the latter. While we are also ultimately concerned with this philosophical understanding of the question, in this paper our main concern is to elucidate one of the fundamental concepts of the criminal law, i.e., punishment, and to do so *not* by offering another ideal

2 R.A. Duff, *Punishment, Communication, and Community* (New York: Oxford University Press 2001).

3 V. Tadros, "Criminalisation and Regulation," In R.A. Duff, L. Farmer, S. E. Marshall, and M. Renzo (eds.), *The Boundaries of the Criminal Law* (Oxford: Oxford University Press 2010).

account of the latter but by an interpretation of punitive practices that emphasizes their psychological underpinnings. As we are to explain, such analyses do have some normative import and are hence relevant to the philosophical question at hand.

The analysis of punishment that we present below is in family with historical and interpretative analyses of the criminal law (Lacey 2009). Unlike standard conceptual analyses, these analyses are not aimed at straightening out inconsistencies through systematization. They start by identifying the (disparate) elements of a certain concept or practice (e.g. the criminal law) in order to offer an understanding of its current form by illustrating the circumstances in which it came about, the purposes that it served in those circumstances, and how it evolved, survived, and adjusted to new contexts. The analysis we intend to offer can be seen as complementing this type of interpretative, historical analysis by focussing on some of the psychological features that cannot be ignored in order to achieve a correct understanding of our concept of punishment and the practices that it animates, such as the criminal law. At the most basic level, the idea behind this type of analysis is that practices such as punishment (in its various forms) are not merely the product of different cultural contexts but also the expression of specific psychologies. Creatures whose psychological profiles significantly differ from ours will likely develop practices and concepts different from ours. Hence, for example, in a world made of creatures who unfailingly obey the law, while it may still be necessary to legislate in order to co-ordinate behaviour, coercive practices such as criminal punishment may be superfluous and would in all likelihood fail to arise.[4]

[4] J. Finnis, *Natural Law and Natural Rights*, 2nd ed. (Oxford: Oxford University Press 2011), pp. 269-270.

Clearly the approach described here is not as such a normative approach. Yet, if correct, analyses that embody these approaches do impose constraints on ideal accounts of the criminal law. This is not because our approach hinges on the controversial assumption that we can derive an 'ought' from an 'is'. That this is not the case can be shown by looking at the type of constraints that we have in mind and the way in which they interact with normative or ideal theories. The constraints we have in mind are of two kinds.

Firstly, consider this. Ideal or normative accounts of the criminal law are not at freedom to provide any account of punishment whatsoever but must provide accounts in which punishment can still be recognized as such. By contributing to an analysis of punishment, we set the frame, and hence the limits, within which any normative or ideal account of the criminal law can legitimately operate, if it is to be recognizable as such.

Secondly, the type of analysis provided below is substantive. In particular, we will show that punishment is retributive at its core. This is to say that, if normative or ideal accounts of the criminal law are to justify punishment at all, they must justify a practice that has some distinctive retributive elements. This may set significant constraints on the normative theorist, for, as we will see, justifying retributive punishment may be especially arduous in the context of liberal politics. Yet this is not to say that the best normative or ideal account of the criminal law must be at least partly retributive, as one may of course be abolitionist and/or defend as non-punitive interventions as the only justified form of legal regulation of behaviour.

Given the centrality of punishment to the (disputed) distinctiveness of the criminal law, we will dedicate the bulk of the paper to the analysis of punishment and then examine its implication for the criminal law. More in particular, we begin by

focussing on generic punishment (Section 2) on the assumption that legal punishment is indeed a specific form of generic punishment.[5] To anticipate a little, we argue that punitive action of any kind is something that originates in our emotional life and in particular in our sense of justice and the emotions that are in family with anger (Section 3 and 4). Punishment, as a concept and a practice, has a distinctive emotive logic that involves distinctive retributive and condemnatory features. Any punitive practice worth of its name cannot escape this fact (Section 5). With this understanding of punishment in hand, we return to the dispute between instrumentalists and non-instrumentalists (Section 6 and 7).

II

Punishment

It is standard fare for legal theorists to start one's account of punishment with the claim that it consists in the infliction of some burden, deprivation, harm, or hard treatment or more generally, of a *disvalue*, or something that a given community recognizes as such. While this is indeed an essential feature of punishment more in general, it would be misguided to think that it is all there is to it. In particular, punishment is essentially historical. To illustrate this consider an act that consists in the intentional infliction of pain on someone. Whether this act amounts to an instance of punishment or an instance of assault will essentially depend on what comes before it. We claim that for it to count as an instance of punishment, it must be understood

[5] This should be an uncontroversial assumption, as it only claims that nothing is to count as an instance of punishment, legal or otherwise, unless it displays those features that are agreed to be necessary for anything to count generically as punishment.

as a *reaction* or a *response*. More in particular, punishment is the infliction of a disvalue as a reaction or response to a *perceived injustice, wrongdoing*, or *violation*. Any infliction of harm that cannot be understood as a reaction of this kind is likely to be understood as an assault, a wrong, or a violation of some kind rather than as a punishment.

Envisaging punishment as a reaction to perceived wrongdoing has one important consequence. In reacting to an (alleged) misdeed by inflicting a disvalue on the alleged wrongdoer, it is quite clear that we are sending the message back to him that his action was unwelcome. This is not an accidental feature of punishment: we want (perceived) wrongdoers to undergo something negative as a response to the wrong they are perceived as having committed. This does indeed come very close to the idea that punishment is in its nature (rather than in its justification) condemnatory.

Punishment then is essentially a reaction to perceived wrongdoing. But it is also more than simply that. It is an undeniable feature of our practice that it comes in unmistaken normative language, the language of justice and deservingness. Whether punishment is just is in general evaluated in terms of desert.[6] Punishment is something the innocent does not deserve, something that the wrongdoer deserves, and, in fact something that he deserves to greater or minor extent depending on the gravity of his deed. In other words, punishment is always

[6] Note that the deservingness relation between the (punishable) act and its (punitive) response does not on its own determine whether it is right or just all things considered to inflict punishment (and what kind and amount of punishment) nor whether there is most reason or it is most rational to inflict punishment (and what kind and amount of punishment). In short, the question of the deservingness of punishment is separate from the question of its infliction (and as we will argue later regulated by distinct 'logics').

considered to be either deserved or undeserved and when deserved there are concerns about the amount, kind, or intensity of punishment that the wrongdoer deserves.

This normativity, however, should not be understood as a sheer linguistic feature of punitive practices. The nature of our discourse should rather be taken to express the phenomenology of our punitive responses. When we punish, or are undergoing the impulse of doing so, our perception of a wrong, and unjustified harm or slight, is accompanied by a feeling of injustice. What is more, we tend to feel entitled to, justified in, or righteous about our punitive attitude. If we perceive that the offender has gotten what he deserved, our feeling of injustice will subside: justice has been done. But if we perceive that the offender has "gotten away with it", the sense of injustice—feelings that the state of affairs ought to be rectified, that the offender deserves to pay for his misdeeds—will linger on for some time.

The immediate reaction that someone deserves punishment is generally modulated by two factors: wrongdoing (which could also amount to an omission) and moral responsibility or culpability. If you committed a wrong act but were not at all culpable (you killed someone while unwillingly hypnotized), no one would say that you deserved to be punished for it. A combination of these two factors, e.g., the severity of the wrong committed and the degree of culpability determines our thoughts about the amount, kind, and intensity of punishment that is deserved.

It becomes clear from these features that our reactive punitive attitudes are *retributive* in nature, for retribution is nothing other than the idea that one should get what one deserves, where this is uniquely determined by the culpability and severity of one's perceived wrong. Given our purposes, this is an arresting conclusion, as it implies that the criminal law has at its core a

practice whose retributive and condemnatory nature would render sufficiently distinct from other modes for the regulation of behaviour available to the state, whose nature is neither retributive nor condemnatory. Instrumentalism would at this point begin to look simply off track, as a view offering an account of something other than punishment.

This central conclusion, however, is in need of greater argumentative support. After all, one may object that the argument so far consists in a mere appeal to intuition and phenomenology. What we would have so far is an arbitrary collection of claims about punishment that include, among others, the idea that it is retributive. In section 3, we will therefore attempt to show that the characterization of punishment provided here is not a mere collection of disparate features but that of a unitary phenomenon. On our account, what gives unity to these features is the fact that punishment is, in some sense to be explained, based on our sense of justice and the emotions in family with anger. In section 4, we show how through this account we gain evidence to the effect that punishment is indeed retributive.

III

The Emotive Account of Punishment

Consider once again our characterization of punishment as a reaction to the perception of a wrongdoing or injustice accompanied by feelings of injustice, and followed by the intentional infliction of disvalue on the perceived wrongdoer. When punishment is considered, more holistically, in these terms, it mirrors important elements of our sense of justice and the emotive basis to which this is often associated. Providing support

for this claim will be our main concern in this and the next section.

Consider the way in which we take norm violations to differ in kind. Norm violations come, as it were, in different colours. More specifically, while we may cognize some norm violations as *injustices* (or *wrongdoings*), we may cognize other violations as *imprudent* or *impolite* actions, i.e., the violations of, respectively, prudential norms (e.g., getting drunk the night before an important exam), or of norms of etiquette (e.g. burping loudly at the dinner table). The capacity to cognize norm violations as injustices is part of what we shall call the *sense of justice*. As we saw, this type of cognitions are integral to understanding any action as an instance of punishment as opposed to a simple act of aggression. There is, however, more to the sense of justice. The idea is that, most often, perceived injustice does not leave us cold but is rather accompanied by distinctive feelings and action tendencies. These perceptions are in other words intimately connected to our capacity to react emotionally and, in particular, to react with anger and related emotions such as resentment, indignation, outrage or fury, and moral outrage.[7] Let us henceforth refer to these emotions as a group as *justice-related emotions*. The thesis we propose is roughly that our concept of

[7] With regard to the specific nature of this connection, R. Rodogno "Robots and the Limits of Morality," in M. Nørskov (ed.), *Social Robots: Boundaries, Potential, Challenges* (London: Ashgate 2016), pp. 39-55, argues in favor of a constitutive claim to the effect that it is precisely these emotions that enable us (developmentally and phylogenetically) to cognize certain norm violations as injustices or wrongs. Note that this thesis is compatible with cases in which a subject perceives injustices unaccompanied by the relevant emotive reactions. For the purposes of this paper, however, it will not be necessary to assume this constitutive connection; a less controversial thesis to the effect that the sense of justice and the anger-related emotions are connected in some way (causally, statistically, or constitutively) will do.

punishment and the practices that it animates are phylogenetically and developmentally dependent on the sense of justice and the capacity to experience these emotions.

We can begin articulating and defending this thesis, by considering some classical characterizations of anger. On Aristotle's much discussed account, for example, anger is an impulse, accompanied by pain, to a conspicuous revenge for a conspicuous slight directed without justification towards what concerns oneself or one's friends.[8] Somewhat similarly, Roberts argues that in anger we construe the situation in these terms:

> S has culpably offended in the important matter of X (action or omission) and is bad; I am in a moral position to condemn; S deserves (ought) to be hurt for X; may S be hurt for X.[9]

From these classic examples, there appears to be a similarity between the characterization of punishment, on the one hand, and that of anger, on the other. First, just as punishment is a reaction to perceived wrongdoing or injustice, anger involves cognitions to the effect that someone has culpably (and hence unjustifiably) offended or attacked you and yours, or violated an important norm.[10] Second, just as in punishment we take

[8] Aristotle, *Rhetoric,* in *Complete Works. Revised Oxford Translation*, vol. 2. edited by J. Barnes (Princeton, NJ: Princeton University Press 1984), 1378a31-1380a4.

[9] R. Roberts, *Emotions, An essay in Aid of Moral Psychology* (Cambridge: Cambridge University Press 2003), p. 204.

[10] It must be noted that the psychological literature is actually divided on one aspect whose presence is decisive to the argument to come. Unlike the account that we will offer, neo-associationist accounts such as those found in L. Berkowitz and E. Harmon-Jones—"Towards and Understanding of the Determinants of Anger," *Emotion* vol. 4, n. 2 (2004), pp. 107-130—dismiss the idea that *other-accountability* and *unfairness* would necessarily characterize the formal objects of anger or, as they would rather say, that they are necessary "determinants" of anger: when angry, it does not follow that we perceive or cognize someone as culpable of an unfair action or wrong. On this account,

ourselves to be righteous in wanting to inflict a disvalue on the offender, in anger we take ourselves to be in a moral position to condemn. Third, in both cases we take it that the offender *deserves* to receive some disvalue. Fourthly, while anger typically involves action tendencies to the effect that the wrongdoer be hurt or be

---

the perception that one's goal is being frustrated or averted may on its own give rise to anger. Given the centrality of culpable wrongdoing (or unjust intentional behaviour) to our idea of punishment, anger understood in this sense would not qualify as a good explanation for it. Psychologists, however, are divided on this issue as attested by the appraisal approach championed by C.A. Smith and L.D. Kirby, "Appraisal as a pervasive determinant of anger," *Emotion* vol. 4, n. 2 (2004), pp. 133-138). In a recent study involving 832 high-school subjects, Kuppens et al. (2007)—P. Kuppens, I. Van Mechelen, D.J.M. Smits, P. De Boeck, and E. Ceulemans, "Individual differences in patterns of appraisal and anger experience." *Cognition and Emotion* vol. 21 (2007), pp. 689–713—have shown that while goal frustration is always a necessary determinant of anger, a large number of subjects will not experience anger unless they perceive the situation as involving other-accountability and unfairness. In other words, due to individual differences with regard to a number of dispositional traits, while goal-obstacle is necessary and sufficient for many, it is necessary but insufficient for many others who also need to appraise the situation as involving others-accountability and unfairness. Importantly, the authors further found that if a situation is perceived as involving these three elements – goal-frustration, other-accountability, and unfairness— almost all participants reported experiencing anger. Some may find it hard to identify, for example, the anger you feel when cheated by someone and wanting justice to be done with the emotion that you feel when inattentively tripping on a table leg and wanting to kick the table. We would find it natural to understand the first as an instance of anger at the wrongdoer and the second as an instance of irritation or frustration. For our purposes, however, we need not decide which camp, neo-associationism or appraisal theory, is right about this. Instead, we will call the reader's attention to the fact that anger is here designated as being in family with emotions such as resentment, indignation, outrage or fury, and moral outrage, all emotions that non-controversially involve other-accountability or the attribution of intentionality. We will simply take the type of anger relevant to our emotive account of punishment to be similar in this respect to other emotions that are in family with it.

inflected some disvalue, punishment involves the actualization of these tendencies. Finally, though this is not Aristotle's claim, we could articulate Aristotle's remark on the painful nature of anger by saying that the negative feeling at issue here is that which we feel when we perceive that an offender has not paid his due, and "justice was not done".

Now we take it that these similarities between anger, on the one hand, and our understanding of punishment, on the other, are not an accidental matter. We rather take it to suggest the thesis already mentioned above according to which:

> *Thesis*. The sensitivity to injustice and the capacity to react emotionally thereupon is necessary to understand the concept of punishment and the practices that it animates.

The claim here is that in the absence of the relevant emotive basis there would be no logic or intuitiveness in the flow from evaluations of culpability to inclinations to inflict disvalue, the flow embodied in justice-related emotional processes presented above. How could the idea of responding to perceived harm with the infliction of harm (or disvalue) be intelligible to anyone deprived of the capacity of such emotional processes? Suppose, for a moment that we could establish beyond doubt that a punitive social practice that responded to harm with harm could be shown to have no deterrent or educational effect whatsoever. While some (but by no means all) of us may thereby take such a practice to lack justification, most of us will still understand the practice, find it intuitive or intelligible, or displaying a certain logic or point. This intelligibility, we claim, is due to our sense of justice and our capacity to experience justice-related emotions.

The thesis we propose is not to the effect that each and every occurrence of punishment requires a corresponding occurrence

of anger in those who impart the punishment. That thesis would be quickly rejected by the existence of institutionalized forms of punishment, such as punishment by the state, in which state officials will often inflict punishment without necessarily feeling justice-related emotions. Our thesis does not try to establish a one-to-one connection between occurrences of punishment, on the one hand, and occurrences of justice-related emotions, on the other. It is a thesis about the genesis and, from there, the nature of the concept of punishment and the social practices that it animates.

The thesis is rather to the effect that grasping the concept 'punishment' requires certain emotional capacities because this concept is itself a product of these capacities. If correct, it would follow from this view that those individuals (human or otherwise) who lack this capacity or whose capacity is damaged or impaired will fail to understand our punitive practices. Individuals are born in social contexts that include punitive practices whose content was actively and progressively shaped through the ages by our ancestors' sense of justice and justice-related emotions. These practices, with their specific content, are already in place whenever any individual is born. As they develop their social, affective and cognitive skills, individuals come to learn about and understand the ambient punitive practices. Those individuals (humans or otherwise) unequipped with the relevant capacities will struggle to make sense of them.

In the remaining part of this section, we present and articulate five auxiliary theses, which, if true, would lend inductive support to *Thesis*. Whether there is indeed any evidence in favour of these auxiliary theses will be discussed in the next section.

Consider first:

> *Thesis 1.* Individuals who lack the capacity to feel justice-related emotions (or whose capacity has been significantly hindered or damaged) exhibit significantly different patterns of behaviour as a response to injustice.

Thesis 1 is indeed very close to our original thesis. However, while the latter focussed on the connection between the capacity to feel justice-related emotions and the *intelligibility* of the concept of punishment, the former focuses on the connection between that capacity and actual patterns of punitive *behaviour* (or lack thereof). Behavioural patterns will be taken as evidence that the correct type of understanding is in place. In particular, evidence to the effect that those whose relevant capacities are absent or impaired do not punish as much, or as hard, or at all, is indirect evidence to the effect that they lack the proper understanding of punishment and, hence, of the practices that it animates. Unfortunately, however, this thesis has to be laid to rest here, as no empirical evidence either in favour or against it seem to have been gathered to date.

With the next four theses, we shift focus from the *capacity* to feel justice-related emotions in general, to the way in which sensitivity to such emotions affects instances of punishment, and the way in which occurrences of the former affect and co-occur with occurrences of the latter. Hence:

> *Thesis 2.* Occurrences of justice-related emotions partly determine who and what is to be considered as deserving of punishment.

Clear evidence that the occurrence of justice-related emotions in a subject has an effect on the subject's judgements about who and what is deserving punishment is here taken as indirect

evidence in favour of our main proposal. The same goes for the next claim:

> *Thesis 3*. Individuals' sensitivity to justice and justice-related emotions affects the content, strength and frequency of their punitive attitudes.

You may be sensitive to justice in the sense that episodes that would typically not elicit for example anger in others, do elicit anger in you because you conceive that episode as unjust. And similarly, you may be more sensitive to justice than others in the sense that you typically experience justice-related emotions more intensely than them. As we understand the thesis, one should expect those who tend to feel for example anger more intensely (with regard to certain kinds of violations, or perhaps with regard to violations more generally) to hold harsher punitive attitudes. This thesis can be understood both at the level of single individuals or of entire groups, be they defined by culture, gender, or both. Hence it may be that due to certain cultural contingencies certain violations are experienced as for example more angering by certain groups as opposed to others. We should thereby expect these violations to be the object of harsher punitive attitudes on behalf of members of such groups.

These theses explicate three important senses in which punishment is based in justice-related emotions. Even though none of these theses implies that we must be undergoing an occurrence of justice-related emotions in order to experience punitive attitudes, the way in which we envisage punishment to be based in justice-related emotions would certainly involve the following:

> *Thesis 4*. Occurrences of punitive attitudes tend to co-occur with justice-related emotions.

The idea here is that punitive tendencies are a part of occurrences of justice-related emotions. Hence, occurrences of the latter will carry punitive attitudes in their stride. The inverse is also true but only typically so given that, as explained above, we may well make judgements about punishment in the absence of an occurrence of anger or other anger-related emotions.

Finally, given our understanding of punishment as inherently retributive, and given *Thesis 4*, we should expect the following final claim to be true:

> *Thesis 5*. Justice-related emotions co-occur with punitive attitudes that are retributive in nature.

As we are about to see, work in social psychology does indeed provide some evidence in support of theses 2-5, and thereby supports our claim that punishment is an emotionally based retributive concept closely connected to our sense of justice. In the next section, we provide a quick overview of the literature relevant to establishing this claim.

IV

The Psychology of Punishment

Social psychologists have begun to explore both people's explicit beliefs about the justification of punishment and their motivation in punishing. A number of studies is taken to show that ordinary people, while overtly declaring to punish on consequentialist as well as retributive grounds, in practice would tend to judge the appropriateness of punishment on various cases pretty much in accordance with retributive intuitions about just

deserts.[11] There is in other words a disconnection between what people take to be the justification of punishment and the way in which they tend to punish in practice.[12] Importantly, however, it looks like when deciding what punishment to impart on someone, individuals are driven by retributivist as opposed to consequentialist considerations.

Some of these studies measured and emphasized the presence and co-variation of anger in connection with the severity of punitive attitudes. Carlsmith *et al.*, for example, found that moral outrage ratings were a strong predictor of judgements about punishment and mediated the influence of retributive considerations on those judgments: seriousness of wrongdoing

[11] See, for example, R.M. McFatter, "Sentencing strategies and justice: effects of punishment philosophy on sentencing decisions," *Journal of Personality and Social Psychology* 36 (1978), pp. 1490–1500; R.M. McFatter, "Purposes of punishment: effects of utilities of criminal sanctions on perceived appropriateness," *Journal of Applied Psychology* vol. 67 (1982), pp. 255–267; D. Kahneman, D. Schkade and C. R. Sunstein, "Shared outrage and erratic awards: the psychology of punitive damages," *Journal of Risk and Uncertainty* 16 (1998), pp. 49–86; J.M. Darley, K.M. Carlsmith, and P.H. Robinson, "Incapacitation and just deserts as motives for punishment," *Law and Human Behavior* 24 (2000), pp. 659-683; K.M. Carlsmith, "The roles of retribution and utility in determining punishment," *Journal of Experimental Social Psychology* 42 (2006), pp. 437–451; K.M. Carlsmith, "On justifying punishment: The discrepancy between words and actions," *Social Justice Research* 21 (2008), pp. 119–137, K.M. Carlsmith and J.M. Darley, "Psychological aspects of retributive justice," *Advances in Experimental Social Psychology* 40 (2008), pp. 193–236; K.M. Carlsmith, J.M. Darley and P.H. Robinson, "Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment," *Journal of Personality and Social Psychology* 83 (2002), pp. 284–299; J. Baron and I. Ritov, "The role of probability of detection in judgments of punishment," *Journal of Legal Analysis* 2 (2009), pp. 553-590.

[12] Nadelhoffer et al challenged the evidence to the effect that we are retributivist in practice as gathered by the studies mentioned above but do provide evidence to that effect through another experimental set up.

and absence of mitigating circumstances tended to co-vary with reported anger and judgements on severity of punishment.[13] This we shall take as evidence in favour of Theses 2, 3, 4 and 5.

Unlike the studies mentioned so far, the studies that we are about to review share the feature of manipulating anger directly. Psychologists working in this area are usually interested in documenting co-variation and causal relations. As a result, they tend to conceptualize anger and punitive judgements/attitudes as separately operationalizable occurrences whose relation needs to be documented. As we shall see, while leading to interesting and useful results, this approach has limited exploratory power.

A useful study is Lerner *et al.*, in which anger was induced in some subjects but not in a control group in order to compare the two groups' respective punitive reactions.[14] More in particular, the experimenters induced anger by exposing individuals to a video displaying bullying behaviour. They then asked the subjects to rate the degree to which perpetrators of hypothetical harms (unrelated to those shown in the video) should be punished. The punishment ratings for the subjects in the anger induction group were higher than those for subjects in a control group, indicating that incidental anger has a causal effect on (spills over) judgments about punishment, in line with (some version) of Thesis 4.[15]

13 K.M. Carlsmith, J.M. Darley, P.H. Robinson, "Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment."

14 J.S. Lerner, J.H. Goldberg, P.E. Tetlock, "Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of responsibility," *Personality and Social Psychology Bulletin* 24 (1998), p. 563.

15 Psychologists are forced by their conceptualizations to understand punitive judgements as effects caused by the occurrence of anger. One may however see them as part of the same process as anger. Thesis 4 is cashed out in terms that are compatible with both views.

The link between anger occurrences, our sense of justice and punitive attitudes has also been studied in connection with the vast literature on the so-called Ultimatum Game.[16] In the standard version of this game, one individual (proposer) controls an amount of money (say $10) and makes an offer to another individual (responder) on how to divide the $10 between the two individuals. Both individuals know the amount being divided and the rules of the bargaining. The responder can either accept or reject the offer. If the offer is accepted, the sum of money is divided as proposed and the bargaining ends. If the offer is rejected, both individuals receive nothing and the bargaining ends. As Srivastava *et al.* explain:

> The game–theoretic, sub-game perfect equilibrium, prediction is that a proposer should offer the smallest unit of currency and the responder should accept. The rationale is that an income maximizing individual would accept any offer since something is better than nothing. In contrast to the normative prediction, two robust findings have emerged in the literature (Camerer & Thaler, 1995; Guüh, 1995). First, proposers typically offer about 30–40% of the total amount, with a 50–50 split often the mode. Second, responders typically reject offers that represent less than 25% of the total amount. These findings suggest that individuals' behaviour is not entirely driven by self-interest. … The finding that responders are more likely to accept small offers when they come from a random device than from a human agent suggests that individuals punish unfairness and are not merely rejecting inequality (Blount, 1995). The willingness to sacrifice one's own interests (i.e., at a cost to one self) to punish those who are being

[16] The literature makes a reference to our feeling of fairness rather than our sense of justice. In this context, however, we take this distinction to be irrelevant.

> unfair suggests that emotions may underlie responders' rejection decisions.[17]

Note also how the nature of the game is such as to fall naturally in line with the idea that punitive attitudes are retributive. Players are aware that the traditional Ultimatum Game is a one-off interaction. Any impulse to punish at a cost to oneself, then, cannot be justified in terms of the effect that it may have on the proposer in future interaction. What is more, in their study, Srivastava *et al.* examine the extent to which in the Ultimatum Game, the cognitive appraisal of unfairness leads to the emotion of anger, which in turn, drives punitive behaviour (i.e., the rejection of offers).[18] The evidence gathered by the authors indeed suggests that anger mediates the influence of offer size on rejection rates as well as the influence of unfairness appraisals on rejection rates, evidence once again in line with Theses 4 and 5.[19]

[17] J. Srivastava, F. Espinoza, A. Fedorikhin, "Coupling and Decoupling of Unfairness and Anger in Ultimatum Bargaining," *Journal of Behavioral Decision Making* 22 (2009), pp. 475–489, on p. 476.

[18] *Ibid.*

[19] *Ibid.*, p. 481. The mediating role of anger was further confirmed in two ingenious ways. First, the authors decoupled the cognitive appraisal of unfairness from the anger reaction. This method is relevant against the background of research showing that behavioral response driven by emotions, and anger in particular, can be altered by leading people to believe that the emotion being experienced is caused by an external, unrelated source. Strivasana *et al.* hence induced their subjects to believe that their anger was to be attributed to something other than the unfair offer. They then recorded that rejection rates in this group fell to 60% as compared to 93% in the control group (*Ibid.*, p. 483). Secondly, in their final study (*Ibid.*, pp. 484-486), the authors confirm that the effect is explained by the specific emotion of anger as opposed to negative valenced emotions in general. Note finally that studies using other bargaining tasks (G. Ben-Shakhar, G. Bornstein, A. Hopfensitz, F. van Winden "Reciprocity and emotions in bargaining using physiological and

As Wiegman notes,[20] studies such as Lerner *et al.*[21] and Strivastava *et al.*[22] deal purely with incidental anger, i.e., how the anger caused by one person has spill-over effects on punitive attitudes directed to other persons. The evidence gathered by Fabiansson and Denson remedies this shortcoming thereby providing further evidence in favour of Theses 4 and 5.[23] Participants gave a brief speech about their life goals to a fictitious participant and were subsequently either insulted or not by the fictitious participant. Next, participants played two ultimatum games against the fictitious participant and non-provoking control counterparts. Across two economic bargaining tasks the authors found that provoked participants punished the speech task counterpart more than unprovoked participants. Angered participants were more likely to give money to a novel participant than the person who provoked them. Angered participants also proposed less fair offers to the speech task counterpart than participants who were not provoked. Finally, they were less willing to accept offers from the speech task counterpart regardless of how fair the offer was.

Beside evidence to the effect that anger incidentally and directly modulates punitive responses in situations in which one is

self-report measures," *Journal of Economic Psychology* 28 (2007), pp. 314–323) have also shown that physiological arousal and self-reported anger are associated with punishment decisions.

20 I.T. Wiegman, *Anger and Puishment: Natural History and Normative Significance* (Ph.D. Dissertation, Washington University in St. Louis 2014), p.14.

21 J.S. Lerner, J.H. Goldberg, P.E. Tetlock, "Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of responsibility,"

22 J. Srivastava, F. Espinoza, A. Fedorikhin, "Coupling and Decoupling of Unfairness and Anger in Ultimatum Bargaining."

23 E.C. Fabiansson, T.F. Denson, "The Effects of Intrapersonal Anger and Its Regulation in Economic Bargaining," PLoS ONE 7 (2012), https://doi.org/10.1371/journal.pone.0051595

personally responding to unfair treatment, there is also evidence that anger modulates our responses in cases of so called altruistic or third-party punishment. The latter is the type of behaviour displayed by those that incur costs in order to punish someone who has not directly harmed the subject or those one cares about. Third-party punishment is quite important to our present concerns, as it may be taken to model criminal punishment. Nelissen and Zeelenberg manipulated anger and guilt in a Dictator Game, a bargaining game similar to the Ultimatum Game in which, however, the proposer dictates the division of the money and the responder or receiver has no say.[24] In the particular setup of their study, the opportunity to punish the allocator was given to a third-party as opposed to the receiver (who had this opportunity in the Ultimatum Game). According to the authors, the evidence gathered suggests that anger and guilt independently constitute sufficient but not necessary causes of punishment. Low levels of punishment are observed only when neither emotion is elicited. As Nelissen and Zeelenberg note in their general discussion,[25] the impact of anger demonstrated in their studies is in line with views that hold punishment primarily to serve retributive purposes.[26] Theses 4 and 5 seem to receive support from these studies.

Importantly, however, this evidence may seem in contrast with Batson *et al.*, whose results indicate that anger is reported by subjects only when either subjects were directly harmed by unfair

[24] R.M.A. Nelissen and M. Zeelenberg, "Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions," *Judgment and Decision Making* 4 (2009), pp. 543-553.
[25] *Ibid.*, pp.548-549.
[26] K.M. Carlsmith, J.M. Darley, P.H. Robinson, "Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment"; J.M. Darley & T.S. Pittman "The Psychology of Compensatory and Distributive Justice," *Personality and Social Psychology Review* 7 (2003), pp. 324-336.

treatment (personal anger) or someone whom the subjects care about or identify with was harmed by unfair treatment (empathic anger).[27] However, no anger was reported by subjects who observed unfair treatment being imparted to someone other than themselves or someone whom they cared about, i.e., anger at the sheer violation of a moral norm of equity or fairness (moral outrage). Batson *et al.*, however, are mute with regard to punishing behaviour. In particular, it does not probe whether those who observe moral violations in the absence of personal or empathic anger, would still typically tend to want to punish the perpetrator. If they did, then we would have some evidence that clearly goes against some of the evidence discussed above. If they did not, however, we may have to consider the possibility that the scope of personal and empathic anger is wider than hitherto thought, so as to actually cover violations that experimenters have intuitively considered as impersonal (as the ones in the dictator game).

All in all, the evidence available in psychology indicates not only that anger and punitive attitudes co-vary but that the former plays a causal role in bringing about and modulating the latter. What is more, most of the studies do focus on bargaining games that seem particularly fitting in showing the retributive nature of punitive attitudes. As argued above, this is evidence in favour of Theses 4 and 5.

The type of studies discussed so far focuses exclusively on the connection between *occurrences* of anger and *occurrences* of punishment judgements. This methodology excludes by fiat a number of potential connections between anger or other justice-

[27] C.D. Batson, C.L. Kennedy, L-A Nord, E.L. Stocks, D.A. Fleming, C.M. Marzette, D.A. Lishner, R.E. Hayes, L.M. Kolchinsky and T. Zerger, "Anger at Unfairness: Is it Moral Outrage?," *European Journal of Social Psychology* 37 (2007), pp. 1272-1285.

related emotions and punishment. For one, it is blind to the possibility that anger may play a causal role in shaping our punitive judgements in other ways, as for example, the way suggested by Thesis 3. Fortunately, however, Milburn *et al.* present evidence to the effect that children who underwent harsh punishment are more likely to endorse harsher forms of punishment as adults such as capital punishment, and, in line with Thesis 3, this effect is mediated by the tendency of these individuals to experience anger ("trait anger").[28]

The Emotive Account of punishment would of course be strengthened by evidence showing, for example, that individuals whose capacity for justice-related emotions was damaged earlier on in their development do not display punitive attitudes or that cultures or genders that are less prone to justice-related emotions tend to display comparatively less punitive attitudes. Even in the absence of such evidence, however, we shall take the above to be sufficient evidence to hold on to the Emotive Account. As argued above, however, while our practice of punishment is shaped by our justice-related emotions, this is not to say that we are inclined to punish and consider punishment deserved exclusively when *in their grip*. These emotions shape individuals in their development and social interactions and have shaped human institutions throughout the ages. We may well have grown to understand when punishment is deserved even in the absence of occurrences of justice-related emotions just as we may recognize while in the grip of anger that punishment would be undeserved. These normative issues will occupy us next.

[28] M.A. Milburn, N.M. Niwa and M.D. Patterson, "Authoritarianism, Anger, and Hostile Attribution Bias: A Test of Affect Displacement," *Political Psychology* 35 (2014).

V

The Internal Logic of Punishment

Proving a tight connection between punitive attitudes and justice-related emotions has an important advantage: just as our emotional reactions are susceptible of normative assessment, so can the punitive reactions that are a part of it. In fact, the latter inherit the normativity of the former, or so we will argue in this section.

Emotions, and anger as one of them, have *formal objects*.[29] An emotion's formal object plays a double role. First, it is essential in making each emotion intelligible as the type of emotion it is. Thus, for example, *danger* is thought to be the formal object of fear: fear is the apprehension of a particular object as dangerous or frightful.[30] The second role of formal objects has to do with the normativity of emotions. Emotions are not simply taken to be brute affective reactions but are assessable as more or less appropriate, fitting, or rational. Formal objects afford the norm against which this assessment is conducted, or again, the correctness conditions for emotional occurrences. Thus, for example, a particular object that is not extremely dangerous makes extreme fright a disproportionate and hence inappropriate (irrational, unfitting, or incorrect) emotional response.

[29] See Teroni (2008) for a comprehensive and up to date discussion of emotions and their formal objects.

[30] The apprehension at hand may happen at the personal or the sub-personal level and mastery of the concept of danger or frightfulness is not necessary to experience fear. Note also that emotions are not identified solely in terms of their formal objects. When characterizing an emotion, we also typically appeal to its phenomenology, the way it feels, and to its action tendencies, i.e., what it typically disposes one to do. Hence, for example, our experience while in the grip of fear will feel quite different from our experience while in the grip of guilt. Similarly, while the former emotion will typically dispose us to flee, the other will typically dispose us to make amends.

Importantly, the normative assessment of emotions can focus on distinct dimensions. In particular, we can distinguish normative assessments that are internal to the emotion from those that are external to it. On the one hand, internal normative assessments are judgments about the fittingness or appropriateness of an emotion to its object such the ones just described. They are internal in the sense that what regulates them, i.e., the formal object, is the very same thing that serves to make each emotion intelligible as the emotion it is. Hence, for example, judgments about the appropriateness of fear are regulated by danger, which in turn is what we use to understand the emotional occurrence at hand as an occurrence of fear. Hence, whether your fear of this spider is appropriate will, in this internal sense, be regulated by the danger that the spider poses.

External normative assessments, on the other hand, do not use the emotion's formal object as their norm. Hence, even though fear may be appropriate if a large bull were charging, it may on that occasion be prudentially best not to feel fear (if one could) and not to do whatever fear disposes one to do (typically, flee). Similarly, even if it were fitting to feel envy towards your rival because he has something good that you lack, it may well be that, from a moral point of view, it is never good to feel envy.[31]

Let us move on to the justice-related emotions in general and anger in particular. It appears that the formal object of anger is an injustice most often in the shape of an unjustified culpable wrongdoing. In anger, that is, we typically apprehend someone as the culprit of an unjustified wrong. On the basis of this, we typically experience feelings of injustice, cognitions to the effect that the wrongdoer deserves a disvalue for what he has done, and punitive action tendencies accompanied by feelings of entitlement

[31] See J. D'arms and D. Jacobson, "The Moralistic Fallacy," *Philosophy and Phenomenological Research* 61 (2000), pp. 65-90, for an elaboration of this point.

or righteousness. Anger is then internally regulated by unjustified culpable wrongdoing. If, for example, you incorrectly believe that Sam has stolen your bike, your cognition that he deserves punishment and any punitive action that you might initiate would be unjustified or inappropriate. Similarly, for punitive attitudes you may have towards someone who is incapable of responsible agency. Finally, extreme anger accompanied by harsh punitive attitudes for a minor unjustified culpable wrongdoing would be disproportionate, unfitting, or unreasonable. The wrongdoer would not deserve such harsh punishment.

In short, then, the normative logic internal to punishment is similar to that of anger and other justice-related emotions. Being essentially connected to these emotions, internally, punishment is regulated by considerations of culpability and seriousness of wrongdoing. The internal logic of punishment is purely retributive. Just as any other emotion, however, anger, for example, and its punitive action tendencies may also be normatively assessed externally via considerations such as expediency, deterrence, rehabilitation, education, rights, and morality. Hence, while internally, questions about punishment are purely retributive, externally, such questions are regulated by a potentially open and diverse set of considerations distinct from considerations of deservingness. Hence, for example, while in the grip of an appropriate occurrence of anger, you may be correct in thinking that someone deserves punishment and in wanting to inflict it upon this person. Yet there may be external considerations such as prudence to the effect that punishment on this occasion (or on all occasions relevantly similar to this one) is not what you have most reason to do (it may be too costly or even dangerous to punish the person in question).

Let us sum up our line of argument so far. We started off with the idea of *legal* punishment and then transitioned to that of *generic*

punishment in order to shed some light on the former. In the process, we have learnt that punishment has a retributive core, which it inherits from the justice-related emotions on which it is based. More in detail, we seem to have reached four significant conclusions:

I. Something counts as punishment only if it is a disvalue inflicted as an intentional response to a perceived wrongdoing.

II. Punishment is always either assessed as deserved to some degree or assessed as undeserved.

III. Judgements about the deservingness of punishment are internally regulated by the formal object of justice-related emotions in terms of culpability and seriousness of wrongdoing, that is, purely in retributive terms.

IV. Normative judgements about punishment are also regulated externally by an open and potentially diverse set of considerations (e.g., education, deterrence, etc.) that are distinct from deservingness.

In line with what claimed in section 1, we take the four theses above to afford a psychologically based, interpretative account of our punitive practices. We have called this account the *Emotive Account* of punishment. With these conclusions in hand, it is time to retrace our steps back to legal punishment and determine what normative consequences follow for it.

VI

Emotions, Punishment, and the Criminal Law

The first important consequence of the *Emotive Account* is that any account of the criminal law that purports to be built around the notion of punishment will carry with it the retributive logic that is internal to this practice and its emotive basis. As claimed above, in the absence of these, our ideas and practices of punishment would not be intelligible.

This is a momentous conclusion for the criminalization question and that from the perspective of both instrumentalists and non-instrumentalists. The latter will now have at least one clear substantive account of the distinctiveness of the criminal law. While this may sound as good news for non-instrumentalists, as we will presently argue, it may not be as good news for them as it at first appears to be. As for instrumentalists, they will now face the task of explaining how from their point of view we should be interested in a practice whose retributive core is at least on the face of it insensitive to considerations of an instrumentalist kind. Let's elaborate this last challenge first.

Consider, for example, an account such as Tadros' where an instrumentalist, non-consequentialist view for the justification of punishment is offered.[32] On this view, what distinguishes punishment from non-punitive penalties is the idea that the former primarily involves making the offender suffer, while the latter is supposed to ensure fairness in the distribution of resources. This distinction provides a basis to determine the scope of the criminal law, which, he claims, is and ought to be about punishment rather than penalties. An upshot of this instrumentalist view is that punishment is imposed on people as a

[32] V. Tadros, "Criminalisation and Regulation.," in R.A. Duff, L. Farmer, S.E. Marshall, M. Renzo (eds.), *The Boundaries of the Criminal Law*.

means to prevent further wrongdoing by others provided that the constraint on inflicting pain on people is lifted because of their wrongdoing.

If the view of punishment defended above is correct, an instrumentalist view of this kind is untenable. Punishment is not only the infliction of suffering, for that is also what assault amounts to. Neither will it do to mention that punishment can be inflicted provided that the offender's wrongdoing lifts the constraint not to inflict pain on him. This still fails to identify punishment. Punishment is rather a *response* to a perceived wrongdoing. To inflict pain, hard treatment, or undesirable experience as a deterrent for others, however, is not to inflict pain, hard treatment, or undesirable experience as a *response* to an offender for what he has done. Deterrence is not necessarily a response. The fact that punishment deters, if that were indeed a fact, is incidental to punishment.

What is more, an account such as this ignores the idea of deservingness. As argued, the concept of punishment comes with that of deservingness. The instrumentalist may reply that, insofar as her account is purely justificatory, the conceptual link between punishment and deservingness may be blissfully ignored. Whatever is true of the concept of punishment, only deterrence counts at the justificatory or normative level. The problem with this stance, however, is that it deeply violates the normative elements inherent to the concept and practice of punishment. If the account were correct, legal discourse about punishment deservingness would be either unjustified or only justified to the extent to which it served the aim of deterrence. This, however, is not the way in which we understand and use this concept. It would distort its meaning. Punishment deservingness follows a retributive logic that is often impervious to non-retributive

considerations. Its peculiar type of normativity imbues our discourse.

The instrumentalist may at this point want to dig her heels. Legal punishment, if justifiable, must come in line with deterrence or whatever other considerations instrumentalists take to provide the ultimate justification of punishment. Deservingness and its retributive logic are acceptable only to the extent to which they fall in line with the instrumentalist logic. Purely teleological views of the justification of punishment, for example, are common place in the relevant Italian and post-1966 (post Alternativ-Entwurf) German criminal jurisprudence.

The question that our analysis helps us pose, however, is the following. How far would these jurists continue to support their justificatory views if the latter became more and more at odds with the retributive logic of punishment, i.e., if there was an ever widening gap between what people took to be deserving and undeserving of punishment, on the one hand, and what the jurists took to be justifiable legal punishment, on the other? By making punishment just another tool for the regulation of behaviour, the risk that instrumentalists incur is to deprive punishment of its very nature and make its legal practice unrecognizable and alienating. In this respect, instrumentalists seem to lack the correct understanding of the human practice of punishment.

The *Emotive Account* is therefore inimical to instrumentalist positions about criminalization and the justification of punishment. This may be thought to be good news for non-instrumentalism, at least insofar as it is in line with the retributivist features of the account. The news, however, is not so good for non-instrumentalism either, though for different reasons.

Let us consider here a non-instrumentalist view such as Duff, which defends a partly retributivist partly teleological view of the justification of punishment.[33] On this view, the retributive logic of punishment is given heed. Such non-instrumentalist view will also be able to show that legal punishment is more than just another instrument for the regulation of behaviour in the hands of the state in that its nature, or at least its justification, is quite unlike that of non-punitive sanctions. However, in deciding what and how to punish, non-instrumentalism is bound to accept that the state has aims other than retribution. These are the aims and constraints mentioned above, whose justificatory force is external to the retributive logic of punishment. Liberal democracies typically appeal to deterrence and rehabilitation as aims that criminal punishment should secure. Criminal policies and decisions, then, must be in line with these aims and constraints in order to be justified.

But in the context of liberal democracies, these aims are only the tip of the justificatory iceberg. Deterrence, for example, has any weight only insofar as a state aims at ensuring the safety (or security) of its citizens, where safety is in turn valuable insofar as it contributes to citizens' well-being or, perhaps, as it is necessary to respecting their rights and autonomy. Many in fact believe that liberal democracies derive their ultimate justification from the ideal of their citizens' equality and autonomy, where the latter is minimally understood as involving a kernel of liberal rights including the right of citizens to participate in government and the freedom to decide how to lead one's life compatible with the equal freedom of all other citizens.

Importantly, however, if this justificatory story is correct, the ideals of equality and autonomy will regulate not only the external

[33] R.A. Duff, *Punishment, Communication, and Community*.

logic of punishment but also its internal one. Or more precisely, given that the internal logic of punishment cannot be internally regulated by anything other than itself, given their ultimate justificatory status, these ideals will externally regulate punishment. This means that retributive considerations of deservingness must either fall in line with such ideals or be systematically overridden or excluded by them. The question, then, is whether retributive punishment can fall in line with such ideals and, if so, to what extent.

In the recent debate, Dubber is one of the few authors that attempts a reconciliation of autonomy as the ultimate source of legitimacy for the liberal state, on the one hand, with the criminal law understood as centrally involving retributive punishment, on the other.[34] According to Dubber and the democratic republican tradition to which he appeals, to say that autonomy or self-government is the ultimate source of state legitimacy is to say that the governed must, directly or indirectly, consent to its actions. "This means that, to put it bluntly, punishment in a democratic republic can be legitimate only as *self-punishment*."[35] Dubber then goes on to examine how the ideal of autonomy understood in this way can be made to square with, respectively, the definition of criminal laws (the realm of criminal law), their application to a particular case (the realm of criminal procedure law), and the infliction of sanctions (the realm of prison law).

For our purposes, it is most important to focus on the definition of criminal laws or the realm of criminal law. In short, on the view proposed by Dubber citizens have, *qua* persons, a right to their autonomy; the function of the law is to protect

[34] M.D. Dubber, "A Political Theory of Criminal Law: Autonomy and the Legitimacy of State Punishment" (March 15, 2004). Available at SSRN: http://ssrn.com/abstract=529522 or http://dx.doi.org/10.2139/ssrn.529522.
[35] *Ibid.*, p. 5.

autonomy;[36] and *criminal* law helps the state discharge this function through punishment. Crime is defined as an autonomous attempt on behalf of a person to compromise or destroy another person's autonomy. When crime has occurred, a person's right to her autonomy has been violated. At this point, punishment becomes the "vindication" or "the dramatic reaffirmation" of the victim's autonomy, as it communicates to the world that the offender's attempt to deny the victim's personhood was unsuccessful.[37] Finally, not only does the victim have the right to have the offender punished, but the offender himself has the right to be punished, insofar as treating him as an ahuman source of danger denies him the "dignity and respect" he "deserves" as a person.

This view contains many interesting claims, such as, for example, the claim that all crime should be conceived as a violation of autonomy[38] or that offenders have a right to be punished.[39] Given our purposes, however, the view's main difficulty consists in explaining how exactly inflicting pain, hard treatment, or undesirable experience on the offender as a response to her action will *vindicate* the victim's *autonomy*, in particular when the offender is himself understood as having rights to autonomy. Why would sending the offender on a luxurious cruise as opposed to inflicting pain, hard treatment, or undesirable experience on him not vindicate the victim's autonomy? What is it about infliction of pain, hard treatment, or

[36] *Ibid*., pp. 30-33.
[37] *Ibid*., p. 33.
[38] See J. Stanton-Ife, "Horrific Crime," In R.A. Duff, L. Farmer, S.E. Marshall, M. Renzo and V. Tadros (eds.), *The Boundaries of the Criminal Law*, for an argument to the opposite conclusion.
[39] Deigh doubts this view: J. Deigh, "On the Right to Be Punished: Some Doubts," *Ethics* 94 (1984), pp. 191-211.

undesirable experience that makes this a particularly appropriate way to vindicate violated autonomy?

Even if retributive infliction of pain, hard treatment, or undesirable experience could be shown to be particularly appropriate from the victim's point of view, how does it square with the view that offenders too are autonomous beings? While it is clear that non-punitive coercion can coherently be envisaged as facilitating compliance with a system of rules aimed at realizing liberal ideals, on a par with more positive incentives such as economic ones, it is far from clear that coercion of the kind involved in retributive punishment would be similarly suited. Retributive punishment does not as such aim at giving citizens-as-rule-followers reasons to comply with rules. Rather it focuses exclusively backwards, on citizens-as-rule-violators, on what they deserve in light of what they have done.

None of this is to say that we find punishing wrongdoers unfitting or, in fact, that we should do so. The point or points are rather these. First, whatever plausibility we find in the idea and practice of punishing wrongdoers is there prior to any story about victims' and offenders' autonomy. The "logic" inherent to liberal autonomy cannot explain, take over, or recast in its own terms the internal logic of punishment. Second, if anything, liberal autonomy, at least as discussed above, is inconsistent with such logic.

While we cannot exclude the possibility of non-instrumentalist views whose liberal credentials better conform with the internal logic of punishment, we hope the above to provide enough material for at least initial scepticism with regard to both instrumentalism and (liberal) non-instrumentalism.

VII

Conclusion

The argument above presents considerable difficulties for both instrumentalism and non-instrumentalism about the criminal law. While the former is misguided about the distinctive nature of the criminal law and runs the risk of developing a practice that is potentially detached from its human understanding, the latter correctly describes the criminal law as distinctive tool for the regulation of behaviour but seems, at least initially, at odds with the justificatory background of liberal states. This, of course, is not to say that normative theorists should give up developing views of this kind. Yet, it is a conclusion that does incline us to consider what alternative views of criminal punishment there are. In line with one part of our conclusion, one option is to accept that there is an unsurmountable tension between retributive punishment and liberal ideals. We would then have to understand whether it matters more to us to maintain such ideals while abandoning retributive practices or to maintain our retributive practices while abandoning or relaxing certain liberal ideals. Yet, in line with the other part of our conclusion, it may simply not be an option to change our retributive practices by simply altering their normative logic from the outside. These practices ultimately rely on deep engrained mechanisms and meanings. While not impossible, their modification or abandonment should be expected to be a long and difficult process.[40]

Aarhus University

[40] We are grateful to Johanna Seibt and Alessandro Spena for their incisive comments.

If you need to cite this article, please use the following format:

Katrine Krause-Jensen & Rodogno, Raffaele, "Criminal Law and the Internal Logic of Punishment," *Philosophy and Public Issues (New Series)*, Vol. 5, No. 1 (2015), 135-171, edited by S. Maffettone, G. Pellegrino and M. Bocchiola