The Justifications of Punishment

Leo Zaibert

Abstract 1

This chapter contains an overview of (a) the classical debate regarding the justification of punishment, (b) the main competing rationales that are opposed in that traditional debate, (c) the efforts to combine these different rationales. This overview allows us to see both what is attractive (and unattractive) about each of these rationales and why it is so hard to combine them. Additionally, the chapter offers novel reasons that stand in the way of successfully combining different rationales. Above all, the chapter explains how what different participants to this debate understand by "justification" is in fact different. Agreeing to a common sense of "justification" should contribute to making progress in the remarkably difficult enterprise of justifying punishment.

I discuss here the thorny problem of the justification of punishment from 2
a very general perspective. What I mean by this is that I will approach the problem very fundamentally – indeed *philosophically*. I am moved by the conviction that, at bottom, the problems involved in this justification transcend the confines of criminal law theory. But I will also emphasize here one particular tradition – the Anglo-American. There are two main reasons for this emphasis. First, that there exists, I think, a tendency amongst non-Anglo-American criminal law theorists according to which Anglo-American contributions to penal theory are deficient. For example, in his monumental *Tratado de Derecho Penal,* Luis Jiménez de Asúa devotes scant attention to Anglo-American sources and developments. But the little attention he does devote is very dismissive. "It is difficult – if not impossible – to speak about a *scientific* approach to penal theory in England or the United States, given that those authors focus on case law or local customs", he tells us.[1] Lamenting that the Puerto Rican penal code of 1902 was based

1 Luis Jiménez de Asúa, *Tratado de Derecho Penal*, Buenos Aires: Losada, Vol. 5, (1964), 59. Translations are mine.

on the famous 1872 Field Penal Code of California, Jiménez de Asúa claims that while this latter code at least displays "traces of the Hispanic civilization", it also (predominantly) displays the "primitive naivete of Anglo-Saxon law".[2]

3 A few decades have passed since Jiménez de Asúa's *Tratado*, but I am of the impression that similar overly dismissive attitudes about Anglo-American penal theory remain, despite valuable contributions in that tradition. Perhaps surprisingly, these two peculiarities of my approach are related: for if I am right in thinking that at bottom the justification of punishment is a philosophical problem, then it is much harder to argue that any given tradition has a monopoly on philosophy. Of course, different philosophical traditions may have their own peculiarities. For example, and as we will see below, the Anglo-American philosophical tradition is particularly influenced by utilitarianism, whereas the European and Latin American tradition may have been more influenced by deontological approaches. I am not sure I fully subscribe to this description, but, in any event, there clearly are utilitarians outside the Anglo-American tradition, and there also are Anglo-American deontologists.

4 Methodological preliminaries aside, I then turn to the discussion of the justification of punishment in earnest.

5 It is generally a good idea to distinguish allegedly purely conceptual (or descriptive) questions from allegedly purely normative (or, in our particular case, justificatory) questions. It seems plain that is one thing to ask what a promise, or an apology, or anything else, *is*, and quite another to ask when and why we are justified in keeping or breaking a promise, or in apologizing, or anything else. In the case of punishment, however, this is not easy to do. And thus, while I am concerned here with the *justification* of punishment, it is important to highlight from the start how the justificatory discussion of punishment is intimately linked to the very definition of punishment. The main reason for this intimate connection, and sidestepping here the many thorny debates concerning that fraught definition, is that whatever else it may seek to do, punishment seeks to make wrongdoers suffer (by somehow diminishing their well-being or by visiting upon them something they do not want, etc.). This essential definitional aspect of punishment highlights the urgency of a justification. After all, and to say the least, to make people suffer is, normally, not a nice thing to do.

2 Luis Jiménez de Asúa, *Tratado… op. cit.*, Vol. 1, 1369.

Of course, other than nakedly inflicting suffering, punishment – and 6
particularly state punishment: the subject matter of the criminal law (and of much political philosophy) – often pursues a variety of other goals. And these other goals are often very worthwhile. For, assuming that crimes are, in principle, bad things, insofar as punishment seeks to prevent crime, punishment is conducive to having a better society. (The "in principle" hedge is necessary in light of the tendency towards over-criminalization, whereby some "crimes" are not, strictly speaking, bad things.)[3] Punishment can contribute to these generally preventive goals in a number of ways: by deterring future criminals (or the one being punished), by rehabilitating or reeducating wrongdoers – or by incapacitating them. These are, in fact, the well-known classic consequentialist justifications of punishment, to which I will return.[4] But given that these goals can evidently be advanced in non-punitive ways – neither education nor rehabilitation needs to be punitive, and even incapacitation need not be punitive, as public health quarantines show – a question immediately suggests itself. Why is the punitive element – the infliction of suffering as such, the so-called "hard treatment" – necessary? This is a terribly complicated question, but one famous way of attempting to answer it points in the direction of another well-known classic justification of punishment: the retributivist justification. Retributivists believe that some actions are so bad that, as matter of justice, those who carry them out *deserve* to be punished – that they deserve to suffer.[5] And it is the opposition between this particular justification (centered around desert) and those mentioned earlier (centered around some good consequences of punishment) that fuels the classic debate around the justification of punishment.

Before delving into that classical debate, a preliminary matter is in or- 7
der. Theoretically speaking, the classical debate (classical at least in the sense that it has attracted the most attention) is not the most fundamental (fundamental at least in the sense of being conceptually prior). For there are thinkers – often called abolitionists – who believe that the badness of making people suffer is so bad that punishment simply cannot be justified.

3 Douglas Husak, *Overcriminalization: The Limits of the Criminal Law*, Oxford: Oxford University Press (2009).

4 The classical defense of consequentialism is found in Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation* in John Bowring (ed.) *The Works of Jeremy Bentham* (Vol. 1), New York: Russell & Russell (1962)

5 The classical defense of retributivism is found in Immanuel Kant, *The Metaphysical Elements of Justice*, New York:Macmillan (1965).

So, more foundationally, we can oppose those thinkers who believe that punishment *can* be justified to those thinkers who believe that punishment *cannot* be justified. The conceptual priority of this opposition should thus be obvious: the participants in the classical debate all agree that punishment can be justified – the classical debate, then, zeroes in on a subgroup among those who agree that punishment can be justified, dismissing those who ab initio deny that punishment can be justified.

8 Although I grant that this other debate is conceptually prior and therefore more fundamental, I am skeptical of abolitionism understood as a general thesis. To the extent that punishment is a pervasively common phenomenon both within and without the state, the idea that we can do without it tout-court strikes me as utterly unrealistic.[6] To the extent that friends, relatives, and even perfect strangers do punish each other (just as they sometimes reward each other) independently of the state, any position that may suggest "abolishing" these aspects of human existence strikes me as an over-reach, and as doomed to ultimately fail. And yet, highlighting the conceptual primacy of the debate between abolitionism and non-abolitionism is important for at least two additional, different reasons. First, because within the specific context of state punishment – which is after all the context that monopolizes scholarly attention – the invitation to rethink punitive practices implied by (a properly limited version of) abolitionism may be particularly interesting, and perhaps particularly urgent. Second, because, as I shall show below, some participants to the classical debate do (sometimes unwittingly) preserve or mobilize abolitionist rationales. In other words, there exists an interesting and not always sufficiently recognized affinity between abolitionism and one of the most famous justifications of punishment (consequentialism): although the official line of these consequentialists is that they agree that punishment can be justified, they are profitably seen (or at least I shall attempt to expose them) as crypto-abolitionists. Moreover, this affinity between consequentialism and abolitionism packs significant explanatory potential in terms of better understanding the justification of punishment.

9 The classical division concerning the justifications of punishment presents us, then, with two distinct groups of views: consequentialist justifications and retributivist justifications. The first of these groups is itself

6 The suggestion that there can be punishment without the state is surprisingly contentious: see Leo Zaibert, *Punishment and Retribution,* Aldershot: Ashgate (2006); see also Leo Zaibert

composed of a wide variety of justifications, whereas the second group is much more unitary. Amongst the consequentialist justifications we regularly encounter approaches that justify punishment based on prevention, rehabilitation, incapacitation, education – the sending of messages, and we can also encounter rationales such as appeasement and reconciliation. But this wide variety within consequentialism is, in a way, a mirage, since, in the end, all consequentialist thinkers assert that what justifies punishment is some good consequence that punishment is supposed to bring about. As we shall see immediately, this means that all consequentialists must see punishment itself as inherently bad – a badness that can be canceled out, or countered, or compensated, by the goodness of this or that consequence. Again, the inherent badness of punishment that they see is the result of its inherent connection to suffering: suffering is bad, bad, bad. All this goes back to famous Benthamite credo: "punishment is in itself evil"; and the only way in which it could be admitted (i.e., *justified*) is "in as far as it promises to exclude some greater evil".[7]

Retributivist thinkers, in contrast, do not invoke any of the consequences 10
that consequentialists suggest justify punishment, but instead turn to the notion of desert (or merit). The retributivist believes that punishment is justified when it is deserved. In other words, the retributivist believes that the badness of suffering – which, in a sense, she need not deny – can be modified by the fact that this suffering is deserved. This does not mean that the retributivist opposes or that she does not value some of those consequences which consequentialists care so much about – it simply means that she does not think that those consequences, good as they may be, are what justifies the infliction of punishment. To put it in the pithy words of a famous retributivist, these admittedly good consequences of punishment are best seen as merely a "happy surplus".[8]

This exposes the direct, though not always sufficiently acknowledged, 11
relationship between the question of punishment's justification and the general philosophical discussion of axiology (or value theory). For our purposes here, a brief definition of axiology should suffice: axiology is the study of which entities and phenomena have value (and why, and to what extent, etc.). The crucial contrast to keep in mind here is that between axio-

7 Jeremy Bentham, "An Introduction to the Principles of Morals and Legislation", in *The Works of Jeremy Bentham* (John Bowring, ed.), Vol. 1, New York, NY: Russell & Russell (1962), 83.

8 Michael S. Moore, *Placing Blame*, Oxford: Oxford University Press (1997), 89, 153.

logical considerations and deontic considerations. Deontic considerations relate to question of what to do (or not to do), they are related to human *action*. Obviously, there must be some connection between axiological matters and deontic matters: in principle, the fact that something is good gives us reasons to do it or to bring it about; the fact that something is bad gives us reasons to refrain from doing it, etc. Almost as obviously, I think, the connection is loose: some things may be good and yet they should not be done, some things may be bad and yet should be done, etc. And some things may be simultaneously good and bad – an axiological tension that generates an extraordinarily difficult decision at the deontic level. As I shall suggest, this axiological tension does capture much of the terribly complicated nature of punishment – and its justification.

12 Despite surface etymological affinities, there is an important difference between "deontic" and "deontological". "Deontic", as we have seen, picks out an *area* of moral philosophy, just like axiology picks another such area. "Deontology", on the other hand, is not an area of moral philosophy, but a type of comprehensive moral theory. Deontological ethical theories are typically – at least since John Rawls's towering work – contrasted against teleological ethical theories.[9] For our purposes here we can focus on the way these comprehensive ethical theories differ regarding their respective treatments of the concepts of the right and the good. The "right" is shorthand for what is to be done (what, indeed, is the right thing to do), and the "good" is shorthand for what is or would be valuable (whether or not it ought to ultimately be done). Typically, deontological ethical theories define the right independently of the maximization of the good, whereas teleological theories define the right in terms of the maximization of the good.

13 A classical exemplar of a deontological theorist is Immanuel Kant, for whom the rightness of token actions is not directly linked to the goodness of their token outcomes.[10] A classical exemplar of a teleological theorist is Jeremy Bentham, for whom the rightness of every action is directly linked to the way in which it maximizes the good.[11] And thus the usual linkages between retributivism and deontological ethical theories, on the one hand, and consequentialism and utilitarian ethical theories, on the

9 John Rawls, *A Theory of Justice* (revised edition), Cambridge MA: Harvard University Press (1999), 21–30 and *passim*.

10 Immanuel Kant, *The Metaphysical Element of Justice*, (John Ladd, trans.), New York: Macmillan (1965).

11 Jeremy Bentham, *op. cit*.

other, take center stage. (Utilitarianism is not the only teleological moral theory: Aristotle's virtue ethics is another conspicuous example, but I shall here focus on utilitarianism, given its intimate connection to punishment theory.) Typically, those attracted to deontological ethical theories tend to be attracted to retributivism and those attracted to teleological theories (like utilitarianism) tend to be attracted to consequentialism. And, of course, both Kant and Bentham fit this (typical) bill perfectly well.

Peculiar views are always possible: say, someone could define the good, 14
not as utility, or pleasure, or desire-satisfaction, but instead as "deserved punishment" and then she would see the right thing to do as always a matter of maximizing deserved punishment. But we can leave those peculiarities aside and stress three main clarificatory points. First, the distinction between the deontic and the axiological is different from the distinction between the deontological and the teleological. Second, one could be concerned with either deontic or axiological matters whether one adheres to a deontological or a teleological comprehensive moral theory. Finally, for a variety of reasons but above all given its influence and its remarkable maximizing ethos, the type of teleological ethical of more relevance in our context is utilitarianism.[12]

The debate between consequentialists and retributivists, in the final ana- 15
lysis, goes back to their fundamental axiological disagreement. The utilitarian finds, in essence, only one thing good: utility. As I have mentioned, while utility can be defined in a variety of ways: as contentment, or happiness, or as desire-satisfaction, or in other peculiar ways, these all boil down to one single thing. That is why, historically speaking, utilitarians typically have cared about only one thing: pleasure (however broadly construed). Of course, since there are occasions in which the opposite of pleasure – suffering, construed similarly broadly – is unavoidable, the utilitarian will sometimes talk about minimizing suffering, rather than about maximizing pleasure. But the maximization of pleasure and the minimization of suffering are simply two different manifestations of one and the same axiological ethos. The deontologist, in contrast to the utilitarian, has theoretical space for all sorts of things to be considered good, and, moreover, she has much more theoretical space for certain actions being right – or, in our context, *justified* – without them maximizing the good. In the case of punishment, retributivists (who, again, tend, for the reasons sketched above, to gravitate

12 Unless otherwise noted, I use "utilitarianism" and "classical utilitarianism" interchangeably.

toward deontology) see the goodness of deserved punishment as related to justice, independently of any consideration of punishment's expediency or usefulness in attaining this or that goal.

16 These axiological considerations reveal a remarkable difference between what each of the opposing camps in the classical debate over the justification of punishment understand by "justification". The bluntness of Bentham's position whereby "punishment is in itself evil",[13] of course, is the result of the clash between the definitional point with which we started – whatever else punishment may seek to do, it seeks to inflict suffering on a perceived wrongdoer – and utilitarianism's obsession with diminishing suffering in the world. The suffering punishment inherently causes is supposed to prevent greater suffering: nothing else. This greater suffering may be be prevented because the wrongdoer (and/or others) will learn the lesson, because the wrongdoers will be afraid of having this suffering inflicted on them (the etymological root of "deterrence" is after all *terror*, as in *terrorism*), or because the type of suffering inflicted on the wrongdoer incapacitates her, rendering unable to do this wrong again, and so on. It does not really matter why or how prevention is achieved, provided it is achieved. Crucially, however, if there is a way of preventing evils without the evil of punishment, the consequentialist's axiology requires her to prefer that way.

17 For the consequentialist, then, to justify punishment is to render it *tolerable*. In a recent book I unearthed a way of referring to consequentialist justifications that has fallen into disuse but that captures its nature quite well. I resurrected Leibniz's way of referring to this family of justifications – as "medicinal" – in order to highlight the extent of their unbridled instrumentality (and its feeble connection to morality).[14] On the consequentialist view punishment is akin to a medicine which may be bad-tasting, or which may be painful or have some other unwanted side-effects, but with which we put up because the medicine is effective in preventing (or curing) a disease whose effects are even worse. Evidently – and on pain of masochism – if we can avoid (or cure) the disease without the medicine (or any of its unwanted side-effects) we would prefer that. Of course, when that is *ex hypothesi* not possible we (well, most of us) acquiesce with the medicine

13 Bentham, *op. cit.*, 83.

14 Gottfried Wilhelm von Leibniz, *Theodicy: Essays on the Goodness of God, the Freedom of Man, and the Origin of Evil* (Austin Farrer, ed.), New Haven, CT: Yale University Press (1952). Zaibert, (2018), *op. cit.*

(and its side-effects), in order to avoid the much more serious consequences of getting the disease. From this perspective, it would be as *absurd* (or pathological) not to see the little bit of suffering needed to prevent the serious disease as justified as it would be not to recognize the preferability of the situation in which the serious disease is prevented without even having to put up with the little bit of suffering.

Matters are importantly different when it comes to retributivism. Since 18
retributivists believe that what justifies punishment is the fact that it is deserved (and not any of its potential consequences, and no matter how beneficial or otherwise defensible these may be), they do not merely *tolerate* punishment. For them, punishment, when deserved, is no longer a bad thing. Justifying punishment, for retributivists, is more than merely *tolerating* it (which is all consequentialist can mean by "justification"): when punishment is justified its infliction is no longer bad – it is no longer true that we would be better off not inflicting it. No one needs to deny that the world would be (axiologically) better had the Nazi holocaust, say, not occurred at all; but once it occurred, the world *is* (axiologically) better if the perpetrators of these horrible crimes are punished than if they enjoy impunity – even though that may mean adding suffering to the world. Again, the retributivist need not at all shun the usual goals dear to the consequentialist: it would be very good indeed if something like the holocaust never happened again and all that – but, independently of those potential consequences, the retributivist would see value in the infliction of deserved punishment in and of itself. Why? Because justified punishment realizes *justice* – and justice is precisely what the medicinal model necessarily leaves out. Leibniz's use of the term "medicinal" in this context is a put-down: "[T]rue retributive justice", he tells us, goes – because it ought to go – "beyond the medicinal". The "harmony of things" that is essential to Leibniz's rationalistic theodicy demands "evil in the form of suffering".[15]

The contrast then should be crisp. For the consequentialist to say that 19
punishment is justified is to say that it somehow makes sense in that, although in itself a bad thing (an infliction of suffering), it prevents worse things. For the retributivist to say that punishment is justified is to say that it is *just* – and justice is valuable in and of itself (regardless of what it may prevent or bring about down the road): not a bad thing. The differences between merely tolerating on the one hand, and rendering just, on the other – between these different accounts of what the very enterprise of

15 Leibniz, *op. cit.*, 425

justifying punishment even *means* or *entails* – help highlight two important aspects concerning the debate over punishment's justification. First, it highlights the ideological proximity between consequentialist justifications and abolitionist (and even nihilistic) positions I flagged earlier. Second, it highlights the profundity and magnitude of the difficulties inherent to the effort to combine retributive and consequentialist rationales. I shall address these in order (devoting much more attention to the second issue).

20 Insofar as the consequentialist is committed to the view that punishment is itself irremediably bad, and that it is, at best, only to be tolerated when useful to avoid more bad things, she is revealed to be in much closer company to the straight-up abolitionist than may appear at first glance. This is so because if it were possible to achieve the commendable goals of preventing crime (modulo the eradication of overcriminalization, etc.) without having to inflict suffering in any way, then the consequentialist would be as much in favor of abolishing punishment altogether as the abolitionist properly so called. The only difference between the abolitionist and the consequentialist in this regard is that the latter is more of a realist, and she thus thinks that since it is not possible to completely eradicate crime (or much less wrongdoing in general), she is not in favor of getting rid of punishment entirely either. Punishment is as much an evil for the consequentialist as it is for the abolitionist – although the former grudgingly concedes that it is a necessary evil.

21 This sort of crypto-abolitionism contributes to explaining the considerable literature devoted to the replacement of punishment with compensatory schemes, public health quarantines, and the like.[16] Often these efforts amount to little more than euphemistic games in which putative "restitution" or "treatment", for example, do look rather punitive: these things are forced upon "wrongdoers", these things are not something that the wrongdoers want done unto them, and in fact these are things wrongdoers find unpleasant, and that are inflicted upon them against their will (and as a response to a "wrong"), etc. Even in cases in which these approaches are not fully reducible to euphemistic exercises, they reveal a remarkable lack of concern with the notion of justice. In fact, the notion of justice is sometimes explicitly jettisoned by these types of consequentialists and by

16 The examples are many. But to list just two, consider David Boonin, *The Problem of Punishment,* Cambridge: Cambridge University Press (2008), and Gregg Caruso, *Rejecting Retributivism: Free Will, Punishment, and Criminal Justice*, Cambridge: Cambridge University Press (2021).

abolitionists – even if this is done inconsistently, sometimes scandalously so.[17] To talk about "justification" without caring about justice strikes some of us as a terribly uninspiring and misguided exercise. In fact, the connection between these approaches and morality – a morality that is richer than a monomaniacal obsession with the diminution of suffering – is itself terribly uninspiring and misguided.

Independently of the problems with the consequentialist justifications of 22
punishment, retributivism owes a story too: how is it that desert renders punishment *just*? While the retributivist is perfectly amenable to admitting that suffering is in principle (i.e., by *default*) bad, she insists that this badness is not *invariant*. Retributivists believe he fact that punishment is deserved changes its moral valence. But, how so, exactly? In general retributivists differ as to how to explain this, but they all (even if tacitly) recognize that *context* matters: what is bad in one context need not be bad in other contexts. This is a view only superficially simple. The retributivist is revealed as open to the possible nuances and modifications that changes in *contexts* may bring about. In contrast, the consequentialist is revealed as utterly indifferent to contexts: all they care about is to reduce suffering, regardless of context. That is why some thinkers – including thinkers who do not consider themselves consequentialists or utilitarians – believe that the very supposition that desert can change the moral valence of suffering is, at best, to engage in "magical thinking".[18] (At worse it is to be a bloodthirsty barbarian.) Some of these thinkers are desert skeptics *tout court*: they think no one can deserve anything, and any talk of desert is magical thinking.[19] Others reveal their skepticism in a more moderate way: they seem to accept that we can deserve all sorts of things, but never suffering.[20]

Of course, that retributivists believe that that fact that it is deserved can 23
change the valence of a certain infliction of suffering from bad to good does not commit them to believing that this particular type of good is greater or more important than any other type of good. Retributivists can recognize that there are other goods, and that these other goods can defeat the good

17 See Caruso, *op. cit.*, see also Karl Menninger, *The Crime of Punishment*, New York: Viking (1968).

18 Martha Nussbaum, *Anger and Forgiveness*, Oxford: Oxford University Press (2016), 24, 25, 29, 33, 36, 38, 54, 127, 136, 161, 184, 185, and 187.

19 See Nussbaum, *op. cit.* See also Derk Pereboom, *Free Will, Agency, and Meaning in Life*, Oxford: Oxford University Press, 2014.

20 Derek Parfit, *On What Matters*, Oxford: Oxford University Press (2012). For my response, see the appendix to Zaibert (2018), *op. cit.*

of deserved suffering. Retributivists, that is, can be, moral particularists. More importantly, our discussion has shown that retributivists can be *pluralists* in ways that consequentialists cannot.[21] Consequentialists have but one only goal: to diminish suffering at all costs, whereas retributivists could have many competing goals – including the diminution of suffering.

24 Because of their pluralism, retributivists do face complicated issues related to axiological conflicts; their purported solutions to these conflicts are not always convincing. The very admission that suffering can be good – even if the retributivist's position is not that suffering *simpliciter* can be good, but only suffering within a highly qualified context – has led many to accuse retributivists of being sadists or worse. Those accusations seem unwarranted, independently of the eventual success of retributivism as a justification of punishment. And it is true that retributivism's richness and pluralistic complexity makes it susceptible to a number of problems that consequentialists need not face. But to avoid having to face these problems, as consequentialists do, by flattening the moral landscape is to reveal the shallowness of one's position.

25 The at best emaciated sense of justification at play in typical consequentialist approaches thus contrasts with the robust sense of justification (tied to a robust sense of justice) of retributivism. The main difficulty facing attempts to combine retributivism and consequentialism, then, is that what each of these camps understands by "justification" is remarkably different. And yet, the attractiveness of trying to find a way of reconciling consequentialist and retributivist justifications is not for that any less real. Most people – both inside and outside academic circles – find some force in ideas of justice. It is good that some Nazis got punished for their crimes; and this goodness could not have been easily replaced by any consequentialist consideration: even if whatever *ex hypothesi* good consequences of punishing, say, a sadistic Nazi guard at the crematoria in Auschwitz, could have achieved without actually punishing him, most people would hold on to the idea that he should have been punished nonetheless. But, of course, most people also think that preventing future crime, rehabilitating offenders, and other paradigmatic consequentialist rationales are also very forceful reasons. So, it has seemed to many that it would be great if we could combine the attractive aspects of retributivism and the attractive aspects of consequentialism.

21 See Zaibert (2018) *op. cit.*

During the last fifty or so years a number of interesting attempts to accomplish this feat have been offered – the so-called mixed justifications of punishment. The locus classicus for the beginning of this push is the simultaneous but independent proposals by Anthony Quinton and by John Rawls, and the also very similar proposal by H.L.A. Hart.[22] The idee mere of this push, most conspicuously found in Quinton, is that retributivism and consequentialism are, despite appearances, actually answers to different questions. (I have dubbed this whole exercise "the two-question strategy".)[23] One question is eminently logical, and its answer "does not provide a moral justification of the infliction of punishment but an elucidation of the use of the word". This answer is retributivist, which, by the lights of defenders of mixed justifications a la Quinton, is "not a moral but a logical doctrine".[24] The other question wonders about the consequences of punishment, thus (somehow) thereby rendering it a properly "moral" question – and the answer to it is consequentialist. A remarkable weakness of this approach is that, of course, no retributivist has ever understood her position as merely semantic thesis. This way of putting things not only wildly misrepresents retributivism (rendering, in passing, the accusation of barbarism frequently leveled against it virtually unintelligible), but it unceremoniously assumes that properly ethical matters are necessarily consequentialist. But there is more that is problematic with these mixed justifications. 26

Rawls was particularly explicit in further linking these two questions to two branches of government. The first question, which allegedly has an unproblematically retributive answer, belongs in the judicial branch: this is the question that judges need to ask themselves. The second question, which allegedly has an unproblematically consequentialist answer, is the question that legislators need to ask themselves. (These two questions roughly correspond to what Hart calls the question of punishment's "distribution" and the question of punishment's "general justifying aims".) The two-question strategy, then, operates with a rather odd (and thin) sense of mixing. The mixing (such as it is) happens to be a mere emergent property of the entire system (of government, in this case), but within the 27

22 Anthony Quinton, "On Punishment", *Analysis* 14.6 (1954): 133–142; John Rawls "Two Concepts of Rules", *The Philosophical Review* 64.1 (1954): 3–32; H. L. A Hart, *Punishment and Responsibility: Essays in the Philosophy of Law*, (2nd ed.), Oxford: Oxford University Press (2008).

23 I discuss it a length in Zaibert (2006) op. cit., 10 ff., and 140 ff.

24 Quinton, *op. cit.,* 134

context in which each of the questions themselves operate, there is scarcely any mixing at all. In other words, and phrasing it in Rawls's ingenious elaboration of the strategy, if you are a judge you are not supposed to be mixing any rationales: you are supposed to be a straight-up retributivist, and that is that. Similarly, if you are a legislator, then you are not supposed to be alive to the pull of different rationales: you are supposed to be a straight-up consequentialist, and that is that. This is a pretty humble victory for a supposedly "mixed" justification of punishment.

28 More recently, a different type of mixed justification – although it is not always, or not conspicuously advertised or recognized as such – has gained in popularity: the communicative (or expressivist) justification of punishment. Consider the gist of this approach, as formulated by R. A. Duff, arguably its most influential contemporary champion: "criminal punishment, I argue, should communicate to offenders the censure they deserve for their crimes and should aim through that communicative process to persuade them to repent those crimes, to try to reform themselves, and thus to reconcile themselves with those whom they wronged".[25] In at least one sense, Duff's communicative justification is even more ambitious than the traditional mixed justifications, for he wishes punishment to accomplish reconciliation, something rarely mentioned in the classical mixed justifications. (Repentance as such is rarely mentioned, but perhaps one could see repentance as part of rehabilitation, and rehabilitation is more commonly found in traditional mixed justifications.) But in another sense, this type of communicative justification is humbler than its forerunners. For Duff's communicative justification is (often) explicitly circumscribed to criminal punishment. And despite Rawls's particular version of the two-question strategy, the standard formulation of the classical mixed justification of punishment is not explicitly restricted to criminal punishment, and in this sense, it is more ambitious than Duff's.

29 Perhaps the most obvious peculiarity of the emphasis on communication is that what is deserved, turns out to be "censure", and it is not clear that, as a conceptual matter, censure needs to involve the sort of hard treatment that has been seen as one essential (definitional) aspect of punishment. Of course, it is possible that the appropriate sort of censure that wrongdoers deserve is inseparable from suffering, and it is also possible that censure,

25 R. A. Duff, *Punishment, Communication, and Community*, Oxford: Oxford University Press (2003), xvii.

in and of itself, constitutes hard treatment. But it is also possible that censure is either not possible or not advisable (why should we engage in a communicative process with Amon Goeth, or with Otto Moll, for example), and that a wrongdoer nonetheless deserves hard treatment; or it could be that censure, in and of itself, is not all that is deserved. Bernard Williams expressed the problem with his usual eloquence: "the idea that traditional, painful, punishments are simply denunciations is incoherent because it does not explain, without begging the question, why denunciations have to take the form of what Nietzsche identified as the constant of punishment 'the ceremony of pain'".[26]

Even if deserved *censure* can be shown to capture enough of the traditional focus on deserved *suffering* that has mattered to retributivists (and that is *definitionally* connected to punishment), it remains unclear how to juggle these disparate goals. Imagine reconciliation is best achieved by not giving a wrongdoer what they deserve (whether censure or hard treatment or whatnot), or that giving what they deserve is not conducive to reform or to reconciliation, and so on. To the extent that reform and reconciliation can be seen as the types of goal that traditional consequentialist had typically had in mind, then the very foundational point as to the differential essence of justification as such from the retributivist and consequentialism that I have explained above would resurface here. If reform could be promoted (or assured) at the expense of deserved censure (and particularly if this deserved censure is in any way suffering-inducing) then, from that consequentialist perspective it is absurd (or perverse) to insist on censuring. Similarly, if the sense in which censure is really deserved, then this retributivist rationale (if that is what it is) will not cede the stage too easily. In either case, and despite how laudable these disparate goals are, the communicative approach does not help us understand how to resolve the conflicts between all these goals. In this regard, the communicative justification does not really advance matters: like classical mixed justifications of punishment, it seeks to combine several laudable goals, but it does not really tell us how to do so – and it remains particularly silent regarding the unavoidable conflicts amongst these values. 30

At a higher level of abstraction, the communicative approach inherits the instability with the very meaning of justification I described above. When 31

26 Bernard Williams "Moral Responsibility and Political Freedom", *The Cambridge Law Journal* 56.1 (1997): 96–102, at 100. For the unclear provenance of this passage see Zaibert 2018, p. 7, fn. 15.

we seek to justify punishment, are we trying to articulate what would make punishment *tolerable* or, instead, what would make punishment *just* (and therefore good-in-itself)? The instability runs through the communicative approach in that, presumably, deserved censure is good in itself and therefore inflicting it (that is, communicating it) would render punishment a good thing, not something merely to be tolerated. Many other goals of the communicative approach, however, are very similar to the traditional consequentialist rationales, and as such they merely render punishment tolerable – not just or good in itself.

32 Admittedly, solving these conflicts is truly a very difficult enterprise, and I am as unable to offer a definitive solution as are those scholars I criticize. But I would like to offer a series of remarks that I think point in the right direction. I am keenly aware that these remarks may appear as further confounding matters (or as otherwise overly abstract and disconnected from "real world" problems), but I do believe that they are on the right track, and that they can help us understand the complicated problem of the justification of punishment better. Above all, I think that the beginning of wisdom is, as I suggested above, to get clear as to what it is that we are trying to do when we are trying to justify punishment. This simply cannot be a bad plan.

33 Despite the confusion in the specialized literature as to what exactly we wish a justification of punishment to accomplish, there exists a very odd presupposition that pervades virtually the entirety of the specialized literature. The widespread assumption is to treat the very process of justifying *anything* as much simpler than it is. Whatever it is that a punishment theorist thinks justifies punishment – whether it be the fact that it is deserved, or that it prevents greater evils, or that it combines these two rationales (and more), or whatever – once that justifying element (or collection of elements) obtains, then our problems are solved. In other, words, the idea seems to be that once justification is obtained, then normative problems disappear.

34 I have appealed to Bernard Williams famous mobilization of the notion of simplemindedness, which he defined as the condition of "having too few thoughts and feelings to match the world as it really is".[27] Williams further explained that simplemindedness entails neither "lack of intellectual

27 Bernard Williams "A Critique of Utilitarianism", in J. J. C. Smart and B. A. O. Williams *Utilitarianism: For and Against*, Cambridge: Cambridge University Press (1973), 149; See also (Zaibert 2018, 17).

sophistication" nor "simple-heartedness". I think both these (admittedly unflattering) characteristics tend to be on display in the specialized literature on the justification of punishment. To begin at the end: there really is no question that there is work of great sophistication in this literature. There is no reason, either, to doubt the good faith and the good intentions of the participants in our debate: they do want to reduce suffering, and a genuinely better society, and so on. But, and now moving away from what simplemindedness *is not* and focusing directly on what it *is*, the problem is punishment theorists have paid insufficient attention to the complexity of the moral world which we inhabit.

Imagine that it were incumbent upon you to punish an instance of 35
serious wrongdoing, of the sort that calls for significant suffering. And imagine, further, that the facts on the ground completely justify the infliction of this considerable amount of suffering. It strikes me as either a shallow or an otherwise utterly self-serving person who would not feel some distress about having to inflict this suffering, arguing that, after all, such suffering is (*ex hypothesi*) justified. To inflict suffering when one is not justified to do so is of course a downright awful thing to do. But to think that to inflict suffering when one *is* justified to do so is a morally *neutral* thing, regarding which a decent person can remain indifferent or indolent, strikes me as evincing a high degree of simplemindedness. Hans Welzel's famous remark as to the difference between killing a mosquito (for whatever reason), and killing a human being (say, in self-defense – or at any rate fully justified) captures this simplemindedness well.[28] With all due respect to animal rights activists, to truncate the life of a human being – even if we are justified in doing so – contains an irreducible and ineludible element of *tragedy* which is utterly absent in "truncating" the life of the mosquito.

It is safe to assume that any human being with even a modicum of moral 36
sensibility would be deeply affected if she had to kill someone, even if she were just justified in doing so. In other words, imagine that you are the victim of an unprovoked attack that risks ending your life, that you have no way of escaping or otherwise frustrating this attack, and that the only way you have to repel it is to respond to the attack with deadly force (imagine, that is, that all the conditions that are necessary for self-defense to be successful obtain), and that you indeed succeed in killing your attacker and saving your life. *Ex hypothesi*, this killing was justified. But, again, it would

28 Hans Welzel, *Das Neue Bild des Strafrechtsystems: Eine Einfuehrung in die Finale Handlungslehre*, Goettingen: O. Schwartz (1961): 24

be chilling if, after killing a human being (admittedly justifiably) you went on your merry way, light as the breeze, since after all you were justified.

37 These are variations of a general motif that has received penetrating attention ins some quarters of contemporary Anglo-American moral philosophy and that in fact goes beyond the discussion of justification as such or of punishment as such – but an attention whose relevance to our topic has been greatly underestimated. I have in mind two related sorts of cases, one made famous by Williams with his famous example: "the lorry driver who, through no fault of his, runs over a child, will feel differently from any spectator, even a spectator next to him in the cab, except perhaps to the extent that the spectator takes on the thought that he might have prevented it, an agent's thought. Doubtless, and rightly, people will try, in comforting him, to move the driver from this state of feeling, move him indeed from where he is to something more like the place of a spectator; but it is important that this is seen as something that should need to be done, and indeed some doubt would be felt about a driver who too blandly or readily moved to that position".[29]

38 The other sort of case was illustrated by Michael Walzer's no less famous discussion of dirty hands problems. He asks us to imagine a particularly decent politician, who wants to win an election without getting his hands dirty (i.e., without having to do anything morally questionable). Soon he realizes that his scruples can stand in the way of getting anything done or even win the election, and that he may, despite his desire to remain perfectly clean, need to strike this or that deal with this or that other shady character: that is just the way the world is. Walzer believes that *our* attitude toward this politician is very complex: "because he has scruples of this sort, we know him to be a good man. But we view the campaign in a certain light, estimate its importance in a certain way, and hope that he will overcome his scruples and make the [shady] deal. It is important to stress that we don't want just anyone to make the deal; we want him to make it, precisely because he has scruples about it. We know he is doing right when he makes the deal because he knows he is doing wrong. I don't mean merely that he will feel badly or even very badly after he makes the deal. If

29 Bernard Williams, "Moral Luck" *Proceedings of the Aristotelian Society, Supplementary Volumes* 50 (1976): 115–135, at 124.

he is the good man I am imagining him to be, he will feel guilty, that is, he will believe himself to be guilty".[30]

A decent politician should feel *guilty* about getting his hands dirty, even 39
if getting his hands dirty is exactly what he is supposed to do – even if he is *justified* in getting his hands dirty. The lorry driver who faultlessly runs over the child is, technically speaking, not even justified: what happened to him was a mere accident. But like the justified agents, he has done nothing wrong. The killer in self-defense is, of course, paradigmatically justified. And yet, in all these cases, there is something to be regretted – and this regret is, as Walzer insightfully insists, a thick moral emotion, not just a matter of a generic "feeling bad". Decent agents will feel bad in cases like these, and decent people will try to help these agents feel better – but, paradoxically, if the agents are too quick to overcome their sorrow and feel fine, this would betray a disturbingly hollow moral personality.

The justification of punishment should not be different, particularly if we 40
take seriously the reality of suffering. Punishment is a matter of getting our hands dirty; as such, it typically calls for agent regret. (The sort of "dirt" Walzer describes and the sort of "agent regret" Williams describe may, in some cases, be quite negligible: think of Goetz and Moll, for example, and the magnitude and quality of evil they brought about).[31] In general, the suffering we inflict when we punish is not canceled out by the fact that we may be justified (whatever that justification may turn out to be) in inflicting it. To expect otherwise is to exhibit simplemindedness; it is to (naively) expect the moral universe to be much simpler and flatter than it is. Punishment is a deeply dilemmatic phenomenon – and not all of this dilemmatic baggage goes away the moment punishment is justified. We are justified in doing all sorts of things whose moral implications transcend such justification. In other words, the fact that we may have been justified in having done something does not eliminate all the moral and political effects of having had to do it: that is just not how the morality and the politics – the morality and politics of real human beings – is supposed to work.

I am aware that these reflections further complicate the already very 41
complicated matter of the justification of punishment. But I cannot help thinking that these complications are indeed necessary. Oddly, a rationale

30 Michael Walzer, "The Problem of Dirty Hands", *Philosophy and Public Affairs* 2.2 (1973): 160–180, at 166.

31 I discuss these issues in Zaibert (2018), *op. cit.*

for this sort of move can be seen in Hart's own motivation for a mixed justification of punishment, when expressed support for: "the view that there is just one supreme value or objective (e.g., Deterrence, Retribution or Reform) in terms of which *all* questions about the justification of punishment are to be answered, is somehow wrong".[32] I agree that that view is indeed wrong, but I lament that neither Hart nor thinkers in his wake have taken the sort of pluralism that we need nearly seriously enough.

Further Reading[*]

Alexander, Larry and Kim Kessler Ferzan (2009) Crime and Culpability: A Theory of the Criminal Law, Cambridge: Cambridge University Press.

Altman, Matthew C. (2021). A Theory of Legal Punishment: Deterrence, Retribution and the Aims of the State, London: Routledge.

Betegón, Jerónimo (1992). La justificación del castigo, Madrid: Centro de Estudios Constitucionales.

Fletcher, George P. (2000) Rethinking Criminal Law, Oxford: Oxford University Press.

Gardner, John (2009). "Introduction" in Punishment and Responsibility, (2nd ed.) Oxford: Oxford University Press: xiii-liii.

Hart, H.L.A. (2009). Punishment and Responsibility (2nd ed). Oxford: Oxford University Press.

Honderich, Ted (2006). Punishment: The Supposed Justifications Revisited, London: Pluto Press.

Kaufman, Whitley R. P. (2012). Honor and Revenge: A Theory of Punishment, Dordrecht: Springer.

Kleinig, John (1973). Punishment and Desert, The Hague: Martinus Nijhoff.

Rodríguez Horcajo, D. (2016). Comportamiento humano y pena estatal: Disuasión, cooperación y equidad, Madrid: Marcial Pons.

Tadros, Victor (2011). The Ends of Harm: The Moral Foundations of Criminal Law, Oxford: Oxford University Press.

Zaibert, Leo (forthcoming 2022), "Rethinking Mixed Justifications" in Matthew Altman (ed.), A Theory of Legal Punishment: Deterrence, Retribution and the Aims of the State, London: Palgrave.

32 Hart, *op. cit.*, 2.

* These are sources that, although not cited above, contain important discussions of our topic. As with the sources above, the Anglo-American tradition – in which I mostly work – predominates. But, as noted at the outset, the value of these contributions transcends traditions.