[Very rough first attempt for ELAC meeting, August 2017 in Oslo. No references. Much hand-waving, or at least inconclusive gesturing. So please do not quote, cite, or circulate further! Thanks, JG.]

Punishment v Self-Defence: Where does the difference lie?

I

A question of justice is a question of allocation. Who is to get how much of what, and how? The words 'allocate' and 'distribute' are almost synonymous in ordinary English. So, as Les Green says, '[i]f we could free ourselves from the familiar Aristotelian categories, we should say that justice is always a matter of distribution.' But we cannot so easily free ourselves. We are lumbered with the idea that distributive justice is one species of justice, not the whole genus. Which species is it? That is much harder to say. The Aristotelian taxonomy begins with a contrast between distributive and corrective justice. Corrective justice is the justice of repair, restitution, redress, restoration, recovery, rectification. It is the justice that is done when some change is justly *un*done. Is distributive justice perhaps simply the residual part of justice, whatever is left of justice after corrective justice has been hived off? Surely not. To mention only the most obvious difficulty, surely procedural justice needs its own place in the taxonomy? Procedural justice is the justice of the 'how' – the justice of the mechanism by which just distribution or correction is achieved. Rawls famously tried to defend his norms of distributive justice by invoking independent norms of procedural justice which he took to be in need of no (or not

much) defence. Nozick responded by, *inter alia*, problematising the distinction between procedural and distributive justice. He resisted the idea that, where justice is concerned, the norms regulating the process of distribution can be disentangled from the norms regulating its outcome, at any rate in the way that is needed for the latter to be defended by invocation of the former. He cautioned that on this point the heading 'distributive justice' is 'not a neutral one.' It predisposes one, he thought, to expect norms of justice that are conformed to in the post-allocative 'end-state', never mind the history of allocative actions, the just or unjust processes, by which that state came about.

If Nozick is right, should we conclude that, taxonomically, procedural justice is but a sub-compartment of distributive justice, the place where 'history' is added to 'end-state'? Again, surely not. Procedural justice surely bears on the distributive and the corrective alike. More likely we should conclude, then, that justice is not very tightly compartmentalised. The Aristotelian taxonomy (and any attempt to improve it by extending its list of species of justice) is afflicted by indeterminacy. It is a scheme of paradigms or 'forms' that that readily blend into each other. This fact does not strip the taxonomy of its philosophical usefulness. It may help with orientation and emphasis and focus. It may help us to divide our philosophical labour into manageable shares. But it is also prone to lead us astray: what begins as a convenient division of philosophical labour easily mutates into a cramping and sclerotic compartmentalisation of reality.

Consider the commonplace addition of 'retributive justice' to the Aristotelian scheme. Rawls says that retributive justice is 'completely different' from distributive justice. But is it? We don't need to know what exactly 'distributive' means in this context to know that whatever we can distribute, we can distribute justly or unjustly. Is punishment not simply one of the many things that we - any of us who are capable of punishing - can distribute justly or unjustly? And is retributive justice, then, not simply distributive justice, where the thing being distributed

happens to be punishment? If so 'retributive justice' belongs at the same level in our taxonomy as, say, housing justice or education justice or asylum justice. For housing, education and asylum are likewise things to be distributed. All fall under the heading of distributive justice.

Here is a reason to think that things are not so simple with punishment. Among writings on the morality of punishment, some but not all have been concerned with its justice. Some but not all of those concerned with the justice of punishment have been concerned to resist the Rawlsian claim that the justice of punishment is 'completely different'. Such writings have, however, been pulled in two contrasting directions. One view, associated most closely with H.L.A. Hart, assimilates retributive justice to distributive justice in the way that I just envisaged. Punishment, on this view, is but another thing that is up for just distribution. The Hartian argument goes like this. To punish one must inflict suffering or deprivation on (supposed) wrongdoers. That premise is conceptual. Then comes an empirical premise. If punishment were abolished, we would pay a price in inflictions of suffering and deprivation later. For punishment is a useful tool of deterrence, incapacitation, sometimes even rehabilitation. In one way or another, the practice of punishment reduces wrongdoing, and hence reduces suffering and deprivation that would, in the absence of that practice, be inflicted by wrongdoers. So there is an inevitable trade-off between the suffering and deprivation of those who are punished on the one hand and the suffering and deprivation that is avoided by punishing them on the other. Or that, at any rate, we are asked to grant for the sake of argument. The task, on this assumption, is simply to work out upon whom the suffering and deprivation is to be inflicted. Some of this work is done by considerations of efficacy. Only if the practice of punishment will reduce the suffering and deprivation more than it will add to it (or only if punishment wins in some similar cost-benefit analysis) do we

have a case for maintaining the practice. That gives us what Hart calls punishment's 'general justifying aim'.

On some views this general justifying aim already implicates norms of justice, for its pursuit already has allocative effects. Others reply: the allocative effects are side-effects of an aggregative aim, so justice is not yet at issue. That debate reveals some quite radical problems in specifying what counts as a question of justice. But whatever the solution to those problems, on any view there are distinct questions of justice that may be asked about the selection – on *non*-efficacy grounds – of (supposed) wrongdoers for punishment, and the selection – on *non*-efficacy grounds – of their punishments. For example: Should actual wrongdoers be selected for punishment over merely supposed wrongdoers, even when doing so makes no difference to punitive efficacy, perhaps even when punitive efficacy points the other way? And should those punished be punished only in proportion to the wrongs that they are being punished for even when that would make the punishments less effective in preventing future wrongs? And so on.

Here we are asking about the norms of distributive justice that bear on the distribution of punishment, or what Hart calls norms of 'retribution in distribution'. He means, not that we can distinguish the retributive questions from the distributive ones and tackle them sequentially, but that we can subsume the retributive questions into the distributive ones. What appears to be a distinct 'form' of justice is exactly the same one that applies to housing, education, asylum, etc. True, the norms of justice governing the distribution of punishment are not the same *norms* as those governing the just distribution of housing, education, or immigration. There are no norms of justice that govern the distribution of all goods and ills alike. Different goods and ills are to be distributed according to different norms of justice. But all are norms of distributive justice all the same.

Or are some of them norms of corrective justice? A contrasting reconstruction of retributive justice, associated with

Jeffrie Murphy, has it that (to stick with Hart's terminology) the general justifying aim of punishment is the annulment of the ill-gotten gains of the person punished. Those who commit wrongs help themselves to extra liberties beyond those to which they are entitled, leaving the rest of us with fewer. Doing so is admittedly perpetrating a distributive injustice. The practice of punishment may reduce the incidence or the grievousness of such injustices, inasmuch as they still lie ahead. Hart is right about that much. But the case for the practice – its general justifying aim – is not to prevent those future injustices, or any other future wrongs. Such prevention is just a fortunate bonus. The general justifying aim is to correct those wrongs that have already been committed. The point, for Murphy, is to rectify the wrong that the punished person committed by annulling the liberty-gains that he unjustly, and hence wrongfully, made at the expense of the rest of us.

It could be that these two views – the distributive view and the corrective view – are reconcilable. Possibly we should read Victor Tadros's book *The Ends of Harm* as attempting a reconciliation, by arguing – very roughly – that the way to annul the ill-gotten gains of the wrongdoer (the extra liberties that she took) is to select her to bear some of the suffering or deprivation that has to be inflicted to reduce future wrongdoing by others. One way to correct past wrongdoing is to redistribute future suffering and deprivation to past wrongdoers, never mind that those past wrongdoers would not themselves be the inflictors of that future suffering and deprivation that is thereby redistributed. I find that the Tadros argument to that position requires too much convolution. Its ingenuity outstrips its insight. But that hardly matters for present purposes. What matters here is that, if he is right, Tadros in a way completes the philosophical programme of Hart and Murphy. They sought to deny the autonomy of retributive justice, to assimilate retributive justice to distributive justice in Hart's case and to corrective justice in Murphy's. That both views have their attractions may lead us to think, as Nozick already encouraged us to think, that distributive

justice blends into corrective justice. We may come to think, reading Hart and then Murphy, that so-called retributive justice is best thought of as a 50-50 blend of the two, a paradigm of the non-paradigm, a dual-aspect *tertium quid*. And we may think that Tadros integrates the two aspects, so that the attractions of the Hart and Murphy views are not really competing.

But again this seems too simple. Quite a bit of the appeal of both the Hart and Murphy views, and of the Tadros integration, is that they lighten the dark moral cloud that otherwise hangs over the practice of punishment. Contrary to initial appearances, they tell us, punishment need not add new evils to the world, new sufferings or deprivations or such like. Punishment can serve simply to reallocate the evils that we are stuck with anyway, never mind what we do. In the Hart view, sufferings and deprivations inevitably lie ahead. All we can do is distribute them differently. If we are going to distribute them differently, we are going to be in the punishment business. In the Murphy view, our task is to eliminate the liberty-gains that the wrongdoer made at our (or someone's) expense. There is only a finite quantity of liberty to go round and some must now be taken from the wrongdoer - he must be deprived of liberty - to restore it (somehow) to the common stock. Both of these accounts allow us to preserve, if we are so inclined, a reassuringly close connection between justice and scarcity. There may be an infliction of suffering or deprivation involved in punishment, but it is suffering or deprivation that someone has to bear whatever we do. Punishment is a response to 'the circumstances of justice', as Rawls calls them, the conditions of life which constrain us to allocate. In the case of punishment, we are constrained to allocate the infliction of suffering or deprivation. If the person punished doesn't bear it, someone else bears it instead.

But this reassuring result is achieved at the expense of a kind of reductivism about punishment. For punishment is not merely the infliction of suffering or deprivation. It is the deliberate infliction of suffering or deprivation on the ground of (supposed)

wrongdoing. One is not punishing except to the extent that one takes the wrong to be the reason why one inflicts the harm or deprivation. Hart and Murphy and Tadros show that there may be reasons, perhaps reasons of justice, to inflict suffering or deprivation upon wrongdoers. They show that it may be unjust to inflict that same suffering or deprivation on non-wrongdoers. They explain why the just infliction of suffering or deprivation upon wrongdoers would not be disproportionate to their wrongs. But they do not show why such infliction needs to be by way of *punishment*, i.e. why it is that, in conforming to these standards the inflictor needs to act for the reason that a wrong was (supposedly) committed. Until they defend that feature of the practice, they have not defended the practice of punishment. The practice of punishment is more specialised; it is a practice with a built-in allocative rationale. The rationale is built in because the punisher cannot, conceptually, escape it. In being a punisher, she acts for a certain reason, viz. the reason that a wrong was committed. I do not claim that Hart and Murphy and Tadros do not explain how there could *be* such a reason. On that score, Tadros improves upon Murphy and Murphy improves upon Hart. The point is that none of them attends to the case for having a distinct practice of *acting upon* that reason.

Thinking closely about punishment helps us to see that the connection between justice and scarcity is contingent. As well as goods and evils that are scarce, and that therefore call for allocation, there are goods and evils that call for allocation by their very natures, even when they are not scarce. As well as punishment, we could include on this list reward, compensation, honour, blame, forgiveness, respect, love, and many others. They have built in connections, conceptually determined, to certain bases of allocation. Of course, they can all be forced back into the scarcity model by a move akin to the one Rawls makes with his 'primary goods' metric, viz. by boiling them down to just some of their ingredients. Punishment is distilled into suffering or deprivation, compensation into money, honour into relative

social esteem, and so forth. In the end, perhaps, everything is distilled to 'utility'. At some point, in any event, the distillate always turns out to be scarce, and then the reassuring connection between justice and scarcity is re-established. Thinking about the case of punishment should cause us to hesitate about such distillations. For in asking about the justice of punishment we are not typically interested only in how to allocate suffering and deprivation, or only in how to allocate its infliction, or for that matter only in how to allocate its deliberate infliction. We are interested in how to allocate *punishment*. And the answer to that is not fully supplied, but is certainly pushed in a certain direction, by the very concept of punishment. One cannot punish without acting for the reason that a wrong was (supposedly) committed, and that fact already forces certain distinctive justice thoughts upon us, beginning with 'where is the relevant wrong?'

Is this enough to warrant thinking of 'retributive justice' as a special form of justice? Perhaps. Partly because of Aristotle's own set up (which uses mathematical imagery suggestive of zero-sum conflicts) distributive justice and corrective justice are often taken to be forms of justice essentially bound up with scarcity. If so, retributive justice might best be regarded, as Rawls says, as something 'completely different'. For there are questions of justice about punishment that are not driven by the worry that, if there is no suffering and deprivation by way of punishment, then there will have to be suffering and deprivation elsewhere. There are questions about the justice of punishment that are not questions about how any scarce good is to be shared out ('geometrically') or reimbursed ('arithmetically'). They are questions about the relationship between the punishment and the wrong that remain even when there are no questions about the relationship between the punishee and others who might be candidates to suffer, or be deprived, in his stead.

Could we abbreviate my point by saying that I am standing up for the element of retribution in retributive justice? That is a misleading way of putting it. As Hart rightly pointed out in

resisting Anthony Quinton's famous 'definitional stop' argument, there is more to the ideology of retribution than the idea of punishment alone can explain, let alone vindicate. Believers in retribution come in various stripes. They are drawn together more by self-identification with their brand than by any shared proposition. Is there is any proposition that they believe in common? Perhaps all believe alike that the *guilt* of the wrongdoer is a standalone reason to punish her? If so, there is already ample work for the retributivist to do beyond merely pointing, Quinton-style, to the very nature of punishment. For nothing in the nature of punishment entails an allocative connection between guilt and punishment. Not all wrongdoing is guilty wrongdoing. Indeed, nothing in the nature of punishment entails that the wrongdoer, guilty or otherwise, is the only one who is to be punished. Vicarious and collective punishments may be abominations, archetypal retributive injustices, but they are not contradictions in terms. Even where vicarious and collective punishments are concerned, those punished have suffering or deprivation inflicted upon them for the reason that a wrong was committed (albeit not, or not necessarily, by them). So when I talk about the built-in allocative base of punishment, I don't have any specifically 'retributive' ideas in mind. Perhaps I should speak of 'punitive justice' in place of 'retributive justice'. For I mean to leave open, whether sound norms of distinctively punitive justice will conform to any characteristically retributivist expectations. All that I mean to propose, and even that rather tentatively, is that there is a distinctively punitive justice for there to be sound norms of. These are the norms that answer the following questions: Which wrongs by whom are reasons of how much force for punishing whom, and in what ways, and to what extent, and on whose authority, and subject to what conditions? And – not to be forgotten – *when and why are those the reasons for whom to act for*? Who, in other words, is to be the punisher?

II

Recent writings on self-defence, it seems to me, have sometimes played fast and loose with the contrast between self-defence and punishment. The most important way in which they have done so, I think, is in neglecting the role in the distinction of the 'punisher' question that I just italicised. This error has profound consequences. But it is not easy to explain how the error creeps in. To explain that properly, we need to work through some other false and misleading moves that afflict writings on the subject. They have a common theme. They exaggerate and inevitabilise punishment's retributive aspects in a way that increases the apparent moral distance between punishment and self-defence. They overpolarise self-defenders and punishers.

Trouble begins, I think, with Jeff McMahan's invocation of the concept of *desert* in the characterisation of punishment. True, people are often said to deserve punishment. But that is not all they are often said to deserve. They are also said to deserve their reputations, to deserve their place in the league table or on the scoreboard, or to deserve some peace. When we hear talk of deserts in such contexts we do not tend to get philosophically agitated. We hear the word 'deserve' to mean something simple and untroubling. When some treatment or status is deserved it is appropriate or suitable or fitting in view of some fact or facts about whomever or whatever it is that deserves it. The fact or facts in question are reasons for the deserved thing to be incurred or borne. A deserved reputation, for example, is a reputation that reflects the true qualities, or the true record, of the person who (or thing that) deserves it. I add 'thing that' because in this ordinary sense it is not only people who have deserved reputations. A certain model of car may deserve, or not deserve, its reputation for reliability. A holiday destination may deserve, or not deserve, its place in the Tripadvisor Top 10. Now it is true, of course, that no model of car, and no holiday destination, deserves to be punished. But that is not because of something

about deserts. That is because of something about punishments. So isn't it peculiar that, whenever it is punishments that are said to be deserved, philosophical agitation tends to be focused on the nature of desert, not on the nature of punishment? The concept of desert is then weighed down with extra philosophical baggage that seems to me entirely gratuitous.

Consider the extra baggage with which McMahan weighs down the concept of desert in *Killing in War*:

> Desert is noninstrumental. If a person deserves to be harmed, there is a moral reason for harming him that is independent of the further consequences of harming him. Giving him what he deserves is an end in itself: Although a deserved harm is bad for the person who suffers it, it is, from an impersonal point of view, intrinsically *good*.

It is hard to see how any of this specialised axiology belongs to the very concept of desert. Surely the case, such as it is, for publishing school rankings or holding beauty contests or reviewing purchases on Amazon.com or similar practices is exclusively instrumental? Doing so is, for example, incentivising or diverting or cautionary. The verdicts passed may be deserved or undeserved. Whether the verdicts are deserved or undeserved depends on which facts are the ones upon which those evaluated are to be evaluated according to the norms of the practice, and whether those facts are accurately conveyed in the verdicts. The verdicts may also do deserved or undeserved harm to the reputation or the self-esteem or the business of those on the receiving end of them. But why would we think that the mere fact that such harms are deserved, when they are deserved, makes it intrinsically good that they are incurred? Still less that it would make the incurring of them *an end in itself*, which means both intrinsically and unconditionally good? Surely one might instead think that the harms to reputation and business and so on are necessary evils. They are among the sad but inevitable effects of the evaluative practice. They have no silver linings except for

those that they inherit from the value of that practice as an instrument for, say, incentivising, amusing, or cautioning.

Those who think, when writing negative reports about employees or negative reviews of hotels, that the fact that the review is deserved is *itself* a silver lining are surely extremely punitive people. They interpret every invitation to deliver a negative verdict as an invitation to punish. Even if that leads them to avoid delivering negative verdicts, they are still extremely punitive people in a sense: for they cannot imagine that someone might deserve to lose the title, or deserve to lose the business, or deserve to lose the trust of others, other than as a punishment. They project certain special features of deserved punishment back onto the very concept of desert.

They are also extremely punitive in another sense. They have an extreme view about punishment. It is an understatement to call it retributive. I suggested before that, for a retributivist, the guilt of the wrongdoer is a standalone reason to punish her. A lot of trouble is concealed in the word 'standalone'. Minimally, it means just this: it means that the reason is *complete*. It is not a mere fragment of a reason. On this reading of 'standalone' a retributivist is one who thinks at least this: that a punisher who offers the wrongdoer's guilt as her only reason to have punished that wrongdoer makes her punitive action rationally intelligible, even if not yet rationally defensible. There is no missing link, no extra premise that we need, to connect what she thought ('guilty') with what she did (punish). But a complete reason to punish is not necessarily a 'standalone' reason to punish in any of the following three stronger senses. First, it is not necessarily a *sufficient* reason to punish. Possibly there must be other reasons to punish that add to the force of the minimally retributive one before punishment is defensible, on some occasions or perhaps even on all. Secondly, it is not necessarily a *mandatory* reason to punish. Even if it is a sufficient reason to punish, it may leave punishment as a merely eligible, not required. In the language of justice, the guilt of the wrongdoer could make punishment just

without making lack of punishment unjust. Thirdly, and for present purposes most importantly, a complete reason need not reflect the intrinsic value, let alone the value in itself, of doing what it is a reason to do. It may be that the value of doing what one has a standalone reason to do lies only in its consequences. And it may be that the value of the existence of the standalone reason lies only in the good consequences of its existence. The guilt of the wrongdoer may be a complete reason to punish him because of the instrumental advantages of its being so. And that, to return to McMahan's terminology, may be all that it takes to yield a minimally retributive view of what makes punishment deserved. The fact that the wrongdoer is guilty is a complete reason to punish him – he deserves it – even if there is no intrinsic value, let alone 'end in itself', in punishing him.

Consider this exchange on the subject of friendship between George and Lennie in Steinbeck's *Of Mice and Men*:

> 'We got somebody to talk to that gives a damn about us [said George]. We don't have to sit in no bar room blowin' in our jack jus' because we got no place else to go. If them other guys gets in jail they can rot for anybody gives a damn. But not us.'
>
> Lennie broke in. '*But not us! An' why? Because … because I got you to look after me, and you got me to look after you, and that's why*.'

George and Lennie are not defending the idea that, when they are looking after each other, each should act for the reason that he will thereby receive, or be entitled to receive, reciprocal looking-after when the tables are turned. Nor are they suggesting that that each should look after the other for the reason that, by doing so, he can look forward to less loneliness, less aimlessness, better conversations, or quicker releases from jail. All of these are possible reasons for one person to look after another. It is not that George and Lennie reject them as reasons. On the contrary they rely on the fact that they are reasons for one person to look after another, and to do so reciprocally. It is just that, in this passage, they are not defending acting *for* any of these reasons. They are

pointing to reasons for being friends, complete with the (conceptually determined) feature that, in looking after each other, they will act for the reason 'he's my friend.' It is the existence of that reason, they think, that gives them the value of conformity with all the other reasons that they list. If they were not friends, they would have less of this further reason-conformity to look forward to. And if there were no such reason as 'he's my friend' to act for, then they could not, conceptually, be friends. That is a decent stab at a defence of the existence of the reason 'he's my friend.' It makes it a complete reason for each of two friends to look after the other. But it points to no intrinsic value in friendship or (unpacked) in doing for each other, for the reason of friendship, whatever friends do for each other.

You may think that a defence of friendship which does not depend on there being any intrinsic value in acts of friendship, nor (therefore) in the reasons for which friends act, is a depressing defence. Maybe. That is neither here nor there. It does not affect the point that the example has for us here. The example shows that even when we agree that F is a complete reason to φ, we still need to ask *why* F is a complete reason to φ. And while the existence of some intrinsic value in φing, or in acting for reason F, gives us a possible answer, the answer might equally be in terms of the instrumental value in φing, or in acting for reason F. And one particular kind of answer might be this: that acting for reason F is a good instrument of conformity with various other reasons to φ. And if acting for reason F has that benefit, then without further ado there is such a reason as F. It doesn't follow that we always have reason F whenever having it would be a good instrument of conformity with those other reasons to φ. There might obviously be other conditions. Not every pair of people who would do well in looking after each other by looking after each other for the reason 'he is my friend' automatically have that reason. There is more to being friends than just being well-placed to enjoy the instrumental advantages of friendship. But the instrumental advantages of friendship are

nevertheless reasons to become, and to stay, friends, and they remain, even in friendship, reasons to do what friends do. It is merely that one does them better *as* friends, i.e. by acting for the reason 'he's my friend'.

How does all this connect back up to desert and punishment? Not too obviously. But here is a possible connection worth thinking some more about. I said earlier that Hart, in spite of his claim to be defending 'retribution in distribution', failed to mount a defence of the specific practice of *punishment*. He defended a punishment-like practice, a practice of suffering or deprivation inflicted only on guilty wrongdoers for the sake of future wrong-reduction. He identified some important reasons why non-wrongdoers, and wrongdoers who are non-guilty, should not be punished. But reasons not to punish non-wrongdoers are not reasons to punish wrongdoers. The reasons Hart listed *in favour of* punishing wrongdoers did not include the fact that they were wrongdoers. As I mentioned, Hart said rather too little about why such a reason might exist. But he said even less about why that very reason is so very central to the nature of punishment. Why is it that punishers, to qualify as punishers, need to inflict the suffering or deprivation *for the reason* that the wrong was committed? Now we can see a possible argument, or a possible kind of argument, that he might have made. Possibly, there are advantages, from the point of view of conformity with some or all of the other reasons that Hart *did* list, in having a practice, participation in which entails acting for a particular reason that Hart *did not* list. Perhaps a good way to draw people away from scapegoating, disproportionate measures *pour encourager les autres*, a destructive blending of peace with war, etc., is to have us treat the fact that a wrong was committed as a standalone reason for someone to suffer or be deprived. Perhaps the hiving off of a distinct practice of punishing, with its built-in emphasis on the fact that a wrong was committed, keeps certain excesses and distractions at bay, even just filters the noise of endless competing considerations. If so, there is a reason for us to

have a practice of punishment. Like the practice of compiling school league tables or leaving Amazon reviews it will inevitably be a practice with its own built-in norms of allocation, which might well be called norms of desert. But there is no sign, so far, of any intrinsic value in people getting what they deserve.

This is, of course, the same point that Rawls made about punishment in his early paper 'Two Concepts of Rules' (seemingly forgotten or abandoned in his remarks on punishment in *A Theory of Justice*). Rawls criticised those who draw

> no distinction between the justification of the general system of rules which constitutes penal institutions and the justification of particular applications of these rules to particular cases by the various officials whose job it is to administer them.

Hart is often remembered as an implementer of that two-level Rawlsian approach to punishment. But he is not. It seems to me that he draws an orthogonal distinction at the first level, a distinction between the general justifying aim of punishment and the rules for its distribution. The latter, as Hart portrays them, are also part of what Rawls calls 'the general system of rules which constitute penal institutions'. Hart has surprisingly little to say about what is supposed to hold at Rawls' second level, which is the level at which we discuss what reasons, and what reasoning, are to be those of the actual punishers. Until he adds more about that, and in particular a case for the fact that a wrong was committed to be the punisher's reason for inflicting suffering or deprivation, he has not (to repeat!) defended the practice of punishment. His is another defence of something like what Rawls calls 'telishment', which is punishment stripped of the defining feature that it is inflicted *for a wrong*. One may share Rawls' view that 'as one drops off the defining features of punishment, one ends up with an institution [viz. telishment] whose utilitarian justification is highly doubtful.' But one need not be a utilitarian to share those doubts. One may simply be someone who thinks, as I do, that there had better be

instrumental value in punishment that improves upon the instrumental value of telishment, even telishment that is distributed exactly as if it were punishment.

By reinstating the missing feature, do I make Hart more of a retributivist? Surely not. I suggested that what retributivists share is the belief that the guilt of the wrongdoer is a complete reason to punish her. I have only insisted on the need for Hart to explain how the fact that a wrong has been committed could be a complete reason to inflict suffering or deprivation. As I said before, not all wrongdoers are guilty and not all those punished for wrongs need, conceptually, be wrongdoers. It is open to Hart to stick with his original claim, which indeed I think is correct, that problems about vicarious and collective punishment, about scapegoating, and more generally about punishment of the innocent, are to be dealt with under the heading of reasons against punishment, not under the heading of reasons in favour. The guilt of the wrongdoer is not, on this view, a reason in favour of punishing. Rather the fact that a wrong was committed is a reason in favour of inflicting suffering or deprivation, which, when inflicted for that reason, becomes punishment. The innocence of the person whom it is proposed to punish is, as Hart rightly argues at length, a very strong reason against punishing her. It belongs to the world of *defences*. Guilty wrongdoing, in particular, is simply wrongdoing without justification or excuse. The fact that I have a justification or excuse tells against punishing me. But my lacking a justification or excuse, *pace* the retributivists, plays no part in the case for punishing me. Or at any rate, that is a plausible position for Hart to maintain even after we make the Rawlsian tweak to his views.

III

I have focused on how to make the Rawlsian tweak in Hart's views partly because his are the views about punishment to which I am otherwise most sympathetic. But focusing on his

views also helps us to see how we can close - or perhaps better, de-radicalise - the gap between cases of justified punishment and cases of justified self-defence. Let me explain.

Hart thinks, as I do, that the practice of punishment is unjustified unless the suffering or deprivation that it imposes upon those punished is made up for in the reduction or mitigation of later wrongs. If there is no scarcity - if the imposition on the person punished is not needed to prevent more and worse actions by that person or others later - we are not justified in having the practice. Hart is right about punishment's general justifying aim. Yet there is still a *reason* to punish without scarcity, i.e. when the general justifying aim is not satisfied. It is the fact that a wrong has been committed. This combination of theses may seem paradoxical. Isn't it an existence-condition for a reason for action, any reason for action, that it is capable, in principle, of justifying an action in conformity with it? So if the fact that a wrong has been committed is a complete reason to punish, don't there have to be at least some conceivable cases in which the that fact alone is justification enough for punishing someone? Not so. Here is one reason why not. Suppose - although this is a simplification - that it is a positive feature of every punishment that it is an action of conformity with, and not just an action performed for, the (distinctively punitive) reason that a wrong has been committed. Let's grant that this is a built-in positive feature of all punishments. We need to juxtapose that, however, with a built-in negative feature of all punishments. It is a negative feature of all punishments that they are inflictions of suffering or deprivation on the person punished. It is a mistake to assume that the built-in positive feature is *ever* capable, by itself, of compensating for the built-in negative feature. So it is a mistake to assume that punishment is ever justified without more reasons in favour of it on top of the reason that makes it punishment.

That still leaves the fact that a wrong was committed as an *extra* positive on top of whatever positive effects punishment has

on the incidence or severity of future wrongs. Like Hart, I think that even that role as extra reason should be denied to it. Thinking back to the remarks on friendship by Lennie and George in *Of Mice and Men* helps us to see why. Lennie and George are defending action out of friendship. They are defending it by pointing to all the other reasons that may be conformed to when one acts out of friendship. But they are not defending the counting of the reason 'he's my friend' on top of the reason 'we'll have someone to talk to', 'we won't rot in jail', etc. They are defending the counting of it *instead* of those reasons. It is a substitute reason. It does not add to the weight of reasons in favour of maintaining a practice of friendship. It does not even add to the weight of reasons in favour of maintaining their friendship. It is simply the reason that must exist, and must be acted for, if they are to maintain their friendship. It does not tip any balances or add to any scales or anything like that. It has the weight in their reasoning that it should have if it is to play its role in helping to secure their conformity with the other reasons that apply to them. Inevitably, in the reasoning of George and Lennie, it will sometimes overshoot and sometimes undershoot relative to the case for its existence and use. Acting for the reason 'he's my friend' will sometimes have them holding onto a friendship for longer than it should be held onto, or giving one up too soon. These are simply costs of the practice. The question must always be asked like this: Is having the reason 'he's my friend' to act upon nevertheless more successful, as a way of securing conformity with the other reasons mentioned by George and Lennie (or similar), than would be their attending to conform with those reasons directly? The same question is applicable to punishment. It is a question about the reason-conformity efficiency of acting for various reasons, which are available as alternatives. They are not to be aggregated, for here aggregation effectively involves double-counting.

If all that is true then there is what McMahan would call an 'unavoidability' constraint in the practice of punishment. The

reason that punishers act for, and their acting for it, must itself be judged by the contribution that their acting for it makes to conformity with that unavoidability constraint. And if that is true, then the practice of punishment converges to a significant extent with the practice of self-defence, as understood by McMahan himself. In both there is an unavoidability constraint. Both are practices concerned with avoiding net deficits (net sufferings, net deprivations, net wrongs, etc.). In both there is also a wrongdoing condition. Both are practices concerned with proper responses to wrongdoing, whether guilty or otherwise and whether responses to the wrongdoer or not. The main difference between them is that the punisher acts for the wrongdoing reason and the self-defender acts for the unavoidability reason (or a rather simplified version of it, in which unavoidability as between himself and the person he is defending himself against is all that counts).

It seems to me that it impossible to see this main difference without thinking of punishment and self-defence as practices, institutionalised in law or otherwise, which exist side-by-side in order to divide up rational labour in optimal ways. At the highest level of abstraction, reasons to punish apply to the self-defender as much as they do to the punisher. It is merely that, qua self-defender, she is not to act for them. She must act only for reasons of self-defence, which means preventatively. The job of exacting punishment not well-suited to self-defensive situations. It will be done worse than if it were left to a different agent. That agent too will do his job worse, we may think, if he slips into too much of a preventative as opposed to punitive outlook. Each has a share of the rational labour. Other features of the two practices bear out this assessment. Punishment, as I already mentioned, is a *deliberate* infliction of suffering or deprivation. It differs from spontaneous retaliation. Why? Presumably because retaliation, without deliberation, is less likely to be justified. In particular it is less likely to be justified because it already invites a dangerous blurring of the boundary between punishment and self-defence:

dangerous because the clear separation of the two practices, punishment and self-defence, is the best way to optimise conformity with the assembled reasons that, at the highest level of abstraction, underlie both of them. We need to beware of practice-changes, then, in which self-defenders get more like punishers and punishers get more like self-defenders. Or so I am guessing. Maybe such changes will turn out for the best, given other things that are going on in our currently very fragile social fabric. There are many contingencies involved. And that is really my main point. There are many contingencies involved. Never mind whether, as I suspect, punishment and self-defence are purely instrumental practices. Be that as it may, the case for keeping the practices as distinct in our everyday thinking as they have traditionally been kept - the case for thinking, say, 'self-defence now, punishment later' - is itself, fundamentally, an instrumental one. Like the case for maintaining a sharp distinction between friends and lovers, it rests on the advantages that come of maintaining an inevitably simplified rational division of labour in a world in which we are otherwise likely to conform less well with all reasons by trying to cope with too many of them at once. It is less of a recipe for disaster if we leave some reasons to be coped with by other people, and on other occasions, and, in particular, through other practices.