A Reconciliation Theory of State Punishment: An Alternative to Protection and Retribution

THADDEUS METZ

Abstract
I propose a theory of punishment that is unfamiliar in the West, according to which the state normally ought to have offenders reform their characters and compensate their victims in ways the offenders find burdensome, thereby disavowing the crime and tending to foster improved relationships between offenders, their victims, and the broader society. I begin by indicating how this theory draws on under-appreciated ideas about reconciliation from the Global South, and especially sub-Saharan Africa, and is distinct from the protection and retribution theories that have dominated the Western philosophy of punishment for about 250 years. Then I argue that it neatly avoids objections to them and is prima facie plausible in its own right. I conclude that this reconciliation theory of state punishment should be taken seriously by philosophers of law and policy makers.

1. Introducing Theories of Punishment

I propose a theory of punishment that is informed by under-appreciated ideas about reconciliation from the Global South, especially sub-Saharan Africa, and conclude that it should be taken seriously as an alternative to dominant Western theories. A theory of state punishment is a comprehensive answer to four major questions about the justice of burdening or depriving someone in response to a legal transgression that appears to have been committed. One question is when the state may rightly punish people, with there being debate about whether it may punish, e.g., those who have broken the law but did so without fault. A second question is why the state may punish anyone at all; given that kidnapping is unjust, why think that imprisonment – which looks an awful lot like it – is just? A third question is how severe a penalty ought to be for a given person, e.g., a slap on the wrist, the death penalty, or something in between? A fourth question (for some reason less frequently addressed by philosophers) is which kinds of penalties the state should mete out, and here we might ask whether fines and

doi:10.1017/S1358246121000400

imprisonment should be the default modes of punishment or whether some other kinds of punitive burdens would be more appropriate.

For about 250 years in the West, there have been two dominant ways of answering this cluster of questions, which are the protection and retribution theories. As discussed below, the former answers these questions by appealing to respects in which society would be protected from crimes in the future by using penalties principally to incapacitate and deter. In contrast, the latter invokes considerations about the past, contending that just penalties are those that fit the nature of the crime that was already committed, regardless of whether they are likely to bring about any good.

Drawing on some ideas about reconciliation that have been prominent particularly in African cultures and philosophies, I spell out a novel alternative and argue that it neatly avoids objections to the protection and retribution theories and is prima facie plausible in its own right. According to this reconciliation theory, the state normally ought to have offenders reform their characters and compensate their victims in ways the offenders find burdensome, thereby disavowing the crime and tending to foster improved relationships between offenders, their victims, and the broader society. Elsewhere I have addressed the many who have suggested that reconciliation is best understood as an alternative to punishment; I have argued that in fact a *punitive reconciliation* is coherent and also more attractive than forgiveness or restorative justice models of it (Metz, 2022). Although I do draw on some of that reasoning, what I mainly strive to do here is instead to show that a *reconciliatory punishment* is a strong rival to, if not preferable to, much more familiar and influential theories of the justification of state punishment in respect of at least the English-speaking world.

In the following I begin by spelling out the protection and retribution theories, to remind readers of their basics and note some objections to them that have been common to make in the literature (section 2). Then, I spell out the essentials of the reconciliation theory, along the way indicating how it is grounded on ideas about criminal and compensatory justice salient especially in the African tradition but also present in some others in the Global South (section 3). Next, I show that the reconciliation theory avoids the problems facing the protection and retribution theories (section 4). I conclude by noting the need to address some prima facie problems with and gaps in the reconciliation theory that critics are likely to raise, suggesting that this new approach warrants further reflection (section 5).

2. Protection and Retribution Theories and Their Problems

In this section I recount the essentials of the two broad approaches to state punishment that have dominated Western thought for at least two centuries as well as point to well-known problems with them. I do this not merely to highlight the distinctness of the reconciliation theory (in section 3), but also to show that it straightforwardly avoids their problems and so merits consideration as a replacement of them (section 4).

2.1. Protection Theories

The protection approach to state punishment harks back at least to the work of Jeremy Bentham (1830), who argued on utilitarian grounds that penalties ought to be used to prevent crime. Utilitarianism is the view that an action or policy is right insofar as it is expected to produce happiness and reduce happiness in the long run, taking the interests of everyone into account. Bentham argued with care that, while punishment always causes some unhappiness, e.g., for harming the one punished and costing society some resources, it is often justified on balance by virtue of the greater unhappiness it precludes, particularly in the form of preventing crime, but potentially also by virtue of pleasing victims and their families.

How severe the punishment should be is whatever would maximize happiness and minimize unhappiness, with the right types of penalties being whatever would do the same. On this score, imprisonment is often recommended as what would effectively both incapacitate and deter would-be offenders. Ideally a penalty such as prison also ought to reform those disposed to commit crime. However, Bentham thought the most good that a penalty could normally do for society would be to deter the general population from committing crime (1830, chap. 3, bk 1), which means that prison would often be justified even if it would not rehabilitate, as it tends not to do (at least as it has typically been employed in the West). Utilitarians might well hold that current forms of imprisonment are unjust because of conditions such as overcrowding and gang life. However, in principle for them there is probably a way to imprison that would routinely best promote the general welfare.

Utilitarianism is just one instance of a broader protection model of state punishment.[1] According to the latter, a necessary condition for

[1] Additional consequentialist, even if not invariably utilitarian, theories include: Braithwaite and Pettit (1990); Smart (1991); Husak (1992); and Shafer-Landau (1996).

state punishment to be justified is that it would have the desirable consequence of preventing crime, centrally by means of deterrence, incapacitation, and reform (again, in practice Western states tend to neglect the latter). By this approach, the right penalty on a given occasion is whichever amount and kind would prevent the most crime with the least degree of harm imposed.

It is important to see that, unlike the utilitarian, one could believe in basic moral rights and also hold a version of protection theory. For example, one might think that punishment is justified by the principles that make sense of using force in self-defence or defence of innocent others (as per Farrell, 1990; Murphy, 1992; Montague, 1995). When a criminal aggresses against others, perhaps he forfeits his rights not to be harmed, at least when inflicting harm on him, such as by putting him in prison, would do the long-term good of protecting innocent parties from becoming victims of crime. Prison would instil fear in would-be offenders, would prevent the actual offender from re-offending, and could (even if in practice it does not) rehabilitate his character. The maximum penalty that would be justified is whatever is no greater than the crime committed, analogous to the way one may not in self-defence shoot someone trying to steal one's toaster oven, while involving the least harm necessary to serve a protective function.

Another way to think of a protection theory of state punishment is in terms of it being a 'forward-looking' account of which penalties are justified and why. In order to know whether to punish a given individual, one must consider what the effects of doing so would be. Specifically, the penalty must be expected to render the one punished unable to commit crime that he would have been inclined to do or to scare off him and other potential offenders from committing crime. In addition, to know precisely how to punish him, one must again look into the future, to ascertain which quantity and quality of penalty would be the least amount required to perform those functions to a maximal extent (perhaps without being any greater than the crime already committed).

Protection theories face at least the following three major objections in the philosophical literature. First, they are known for entailing that it can be just to punish persons who have not culpably committed any crime when it appears unjust to do so. Utilitarianism in principle justifies punishing an innocent person if the long-term results of doing so would be for the greater good. Sometimes that is thought to take an extreme form in which an innocent is framed for a crime so as to deter even worse crimes. Other times it is claimed that utilitarianism justifies what is often called 'strict liability', punishing those who break the law even though it

was not at all their fault, e.g., serving alcohol to a minor who produced convincing fake identification, the thought being that such stringency would be likely to deter law-breaking.

The point probably applies to the defensive force version of protectionism, too (even if that is not as frequently recognized). After all, the logic of defensive force is usually taken to allow it to be used against 'innocent threats', those who pose harm to other innocent people for no fault of their own. If someone temporarily loses his mind and is attacking you, for most ethicists you may respond with the least force necessary to protect yourself, despite his innocence. However, so the objection goes, when it comes to punishment, it is nearly always unjust to inflict it on those who were not at all responsible for their actions.

A second objection to protection theories is that they tend to entail that certain overly harsh penalties can be just. That is again a stock problem with utilitarianism, which in principle could approve of, say, torture or the death penalty for those who commit traffic offences such as failing to indicate or speeding, if that would deter people from committing them and thereby save lives in the long run or even if that would promote much less significant benefit, such as convenience, for a much larger number of people.

The logic of defensive force also suggests that certain severe penalties are justified when they in fact seem not to be. If torturing a torturer would prevent more torture, then it would on grounds of defensive force be permissible, supposing that the logic of defensive force indeed justifies punishment. However, the state simply should not be in the business of torturing anyone, at least not as a penalty. Some defensive force theorists are willing to 'bite the bullet' when it comes to the death penalty, contending that, if it would indeed prevent more deaths, it is justified when inflicted on murderers (e.g., Montague, 1995, pp. 135–36, 155; Farrell, 2004). However, it is difficult to accept the natural extension of the point to torturing torturers or raping rapists.

A third objection commonly made to protection theories is the converse of the second, viz., that sometimes they prescribe penalties that are intuitively too light. Both utilitarianism and defensive force theory require minimizing the harm inflicted with punishment, whenever one can prevent no less crime that way than with a greater penalty that is more comparable to the gravity of the crime. For example, if five years in prison were all that it took to prevent a first-degree murderer from killing again as well as to deter others from killing, then no greater penalty would be justified on grounds of protection. Indeed, if no penalty at all were necessary to prevent

murder to the degree that some penalty would, then no penalty would be justified. The deep reason that the second and third objections both apply is that a protection theory ties the justification of punishment to the prevention of crime through deterrence, incapacitation, and (ideally) reform, where the results of penalties vary depending on the circumstances.

2.2. Retribution Theories

Since it seems easily able to avoid these three problems with protectionism, many have opted for the other major theory of state punishment, i.e., retributivism or a 'pay back' account. Broadly speaking, a retribution theory maintains that penalties are justified on 'backward-looking' grounds. One is to look into the past to see whether a crime was committed by a responsible agent and how grave it was. If there was a crime culpably done, then there is moral reason for the state to punish the criminal, and the right penalty is whatever is proportionate to the nature of the crime, where that includes the degree of responsibility for it. Hence, if there was no crime or no one responsible for it, no penalty is justified on retributive grounds, and if penalties do not fit the crime (including level of responsibility), either for being disproportionately severe or light, they are also unjustified.

There have been three prominent forms of retributivism. The most influential version is the desert theory, according to which state punishment should serve to give offenders what they deserve for having culpably done wrong. In the way one can positively deserve a reward for having been heroic or a well-paying job for having obtained qualifications, so one can negatively deserve to suffer harm for having mistreated other people. What one deserves is whatever is proportionate to what one did. Desert theory goes back some millennia, with 'an eye for an eye' appearing explicitly in the Hebrew Bible and, amongst classic philosophers, advocated at times in the work of Immanuel Kant.[2]

More recently, philosophers of punishment have articulated and supported backward-looking theories that are not grounded on desert. One is the fairness or fair play theory, according to which criminals gain an unfair advantage relative to law-abiding citizens such that state punishment must be imposed to remove it (e.g., Murphy, 1979; Sadurski, 1985; Davis, 1992). The idea is not that

[2] See Kant (1797/1996, pp. 472–77). More recent works include: von Hirsch (1986); Moore (1997); and Kershnar (2001).

criminals gain financially or materially from their crime, but rather that, in the act of committing a crime, they take a liberty that others have restrained themselves from taking upon obeying the law. Punishment is justified insofar as it removes the extra liberty the criminal took, with the greater the liberty taken, the greater the justified penalty.

A third backward-looking theory that contemporary philosophers have advanced is expressivism or censure theory, which is the view that the state ought to punish offenders so as to convey certain disapproving attitudes or judgements. For instance, upon punishing one who has broken just laws, the state thereby stands up for the victim who should not have been wronged and treats the offender as a responsible agent who misused his moral capacities. For the state not to punish the guilty would constitute a failure to respect the agency of both parties, where the greater the wrong done, the greater the disapproval that must be expressed and hence the greater the penalty should be (e.g., Feinberg, 1970, pp. 95–118; Hampton, 1988; Duff, 2001).

For all three of these retributive theories, punishment need not do any good in the future to be morally justified. Instead, for all three, it is sufficient that the penalty is proportionate to a crime that was (culpably) committed in the past, in stark contrast to the protection theories. However, like the protection theories, the retribution theories are natural allies of imprisonment as a mode of punishment. Even if jail would do nothing in terms of incapacitating the offender (say, because he would commit crimes against other inmates) or deterring others, it could be an appropriate penalty because the harm or restriction of liberty involved would fit the nature of the crime that the offender was responsible for having committed.

Although retributivism appears attractive in virtue of avoiding problems facing protectionism, it is not clear it can avoid all of them and it also faces some problems of its own. One concern common to both broad classes of theories is that they end up justifying penalties that are too harsh. Of course, for the retributivist the punishment must fit the crime, such that it would be wrong to take two eyes for one. However, it still seems to license literally taking a person's single eye if he has wrongfully gouged out someone else's. It is true that this mode of punishment would not be required, so long as some other penalty, say, a length of prison time, were equal in amount of harm. The point, though, is that there is nothing in the logic of retributivism to *forbid* maiming offenders who have maimed others, for that would be one way of imposing a proportionate penalty. Similar remarks apply to torture, rape, whipping, and death; these, too, could well be proportionate to crimes involving those activities, but the state would intuitively be wrong to mete them out.

A second objection to retributive theories is that they seem insufficiently responsive to the character of the offender, in their standard versions focusing exclusively on an offender's actions as opposed to attitudes. On the one hand, it appears that retribution accounts are forced to prescribe the same penalty for a first time offence and for the same offence undertaken a second time after having undergone the initial penalty. If the same crime is performed a second time, the same, proportionate penalty is warranted. However, many have the intuition that a stronger penalty is often appropriate the second time, and not merely because, say, the level of responsibility is greater; for it could in theory have been just as high on the first occasion. On the other hand, sometimes the character of an offender is such as to ground, not an enhanced penalty, but a reduced one. Here, consider those who are remorseful for their misdeeds. There is nothing they can do to change what they wrongfully did, with the fact of having done wrong being all that matters for standard versions of the retribution theory. A person's present attitudes are not relevant to the justification of punishment, and only their past actions are. However, many have the intuition that a somewhat lighter penalty, i.e., mercy is appropriate for someone who accepts the error of his ways, feels bad for what he has done, and would not do it again.

A third objection to retributivism is that, even if it can entail plausible judgements of who should be punished, how much, and in what manner, it offers an incomplete explanation of why state punishment is justified. Surely one major point of a criminal justice system is to reduce crime, so goes the criticism, while retribution theories counterintuitively maintain that the aim of punishment has absolutely nothing to do with that. A criminal justice system is extraordinarily expensive in terms of money and also requires substantial labour from people, where it is difficult to believe that these costs would be worth the mere 'gain' of imposing suffering merely for its own sake and without any expected long-term good.

In the following sections I spell out an alternative theory of state punishment's justification. Although I expect many readers to find it somewhat intuitive in its own right, my principal defence of it will consist of showing that it neatly avoids all the problems mentioned above with the long-standing protection and retribution theories.

3. The Reconciliation Theory

My favoured approach to punishment is grounded on ideas about how people should resolve conflict that have been salient amongst

sub-Saharan African peoples and have informed their moral philosophical thought. Their watchword has been 'reconciliation', which characteristically involves hearing out those involved in conflict and then offenders apologizing, making compensation, committing not to do wrong again, and afterwards rejoining society. Reconciliatory approaches of various kinds have been used to respond to large-scale social conflict in African countries such as Sierra Leone, Rwanda, and South Africa. The latter's Truth and Reconciliation Commission (TRC) has been particularly influential, famously having listened to victims' stories, awarded amnesty from criminal and civil prosecution to offenders if they fully disclosed their apartheid-era political crimes, and directed the government to compensate victims. Although the TRC advanced reconciliation as an alternative to punishment, and it is often associated with forgiveness, in this section I suggest that the best sort of reconciliation would be punitive (while in the next section I provide reason to think that the best sort of punishment would be reconciliatory).

To begin to see the plausibility of my approach, consider that many of South Africa's victims of human rights abuses during apartheid were not satisfied by the TRC and wanted perpetrators to face something that was routinely labeled 'justice' (Hamber, Nageng, and O'Malley, 2000, pp. 30–32, 37–39; Hamber and Wilson, 2002, p. 48). After all, torturers and murderers went scot free if they confessed their wrongdoing (they did not even have to express remorse), a bitter pill to swallow even for those cultures that prize a reconciliatory approach to resolving conflict. What was missing, I suggest, was not obviously retribution, but rather a sort of reconciliation that included a proper disavowal of the injustice that had been done. A desirable kind of reconciliation would be one not merely aiming to repair relationships by hearing out victims, healing their wounds, and providing reason to think they would not be revictimized. It would also include offenders feeling bad and placing burdens on themselves to express their guilt, or at least a public organization expressing the judgement that what they did was wrong by holding them accountable (beyond the discomfort of recounting their crimes).

Consider that in cases of conflict between family, friends, and lovers, that is what many of us want to see – we would like those who have wronged us to experience some guilt and to atone (cf. Metz, 2022). And we also typically want those who care about us to express the judgement that we were mistreated and to distance themselves from those who mistreated us, at least until the wrongdoers have fully expressed their remorse. Forgiveness should often ideally

come, but only after wrongdoers have undergone some burdens. Something analogous is apt in the sphere of criminal justice. There, too, offenders should accept burdens as a way to show regret for how they behaved and the state should impose burdens as a way to disavow the way they treated their victims.

In the light of these reflections, I submit that the sort of reconciliation the state should promote in the sphere of criminal justice is one that expresses disapproval of the offender's crime by imposing burdens, albeit burdens that are productive in the sense of improving relationships (first advanced in Metz, 2019). Punishment is justified if it serves these dual functions. On the one hand, penalties should be ways for the state to stand up for victims, if not also for offenders to express their misgivings, while, on the other hand, penalties should be of a kind that have offenders compensate their victims and reform their characters so that they will not revictimize anyone, consequent to which it would be reasonable for them to rejoin society.

Something like this approach to sentencing has evidently been adopted at times by the Yoruba people in what is today Nigeria. One philosopher whose ideas are grounded on Yoruba beliefs and practices remarks that 'the reconciliatory factor is lacking in Western theories of law and penology where the offender is punished without making restitutions; and emerging from prison, he is reconciled neither to himself, his victim, nor to society….(W)hen a culprit is punished, such is done with the view to fine-tuning the character of the said offender in line with the communalistic ethos of the Yoruba culture' (Balogun, 2018, pp. 246, 311). Another philosopher doing similar work also reports, apparently approvingly, that those who had committed crimes that 'do not threaten the existence of society' were often punished in Yoruba communities by 'being forced to labor on community projects or those of their victims in reparation/restitution for the loss caused' (Bewaji, 2016, p. 164).

Rwanda is another African country in which a broadly reconciliatory and punitive approach has been adopted, specifically in response to the 1994 genocide. The country had flirted with the idea of a *Fonds d'indemnisation* (FIND), a compensation fund into which those convicted of genocide were supposed to pay (Bornkamm, 2012). Although that did not materialize, the famous Gagaca courts did sentence more than 100,000 offenders to perform community service (Penal Reform International, 2010).

Colombia is a third society in the Global South that has considered a reconciliatory approach to sentencing. Like Rwanda it has not done so in the context of everyday crimes but instead as a measure of transitional justice, specifically in response to the long-standing conflict

between the government and the FARC guerrillas. As part of what was titled the 'Final Agreement' of 24 November 2016 (Government of the Republic of Colombia, 2016), which was meant to provide a definitive framework for peace between the Colombian state and the rebels, victim compensation is central. As to who is to do the compensating, the agreement proposes that it should be offenders in the first instance, with use of the compelling phrase 'restorative sanctions' (Government of the Republic of Colombia, 2016, p. 175): 'In the context of these (reparation) plans, stress will be laid on acknowledging the responsibility of the state, the FARC-EP, paramilitaries and any other group, organisation or institution that caused harm or injury during the conflict' (Government of the Republic of Colombia, 2016, p. 191; see also pp. 137, 145, 189). Restorative sanctions are to advance 'the overall aim of realising the rights of victims and consolidating peace. They will need to have the greatest restorative and reparative function in relation to the harm caused' (Government of the Republic of Colombia, 2016, p. 174). Included amongst a list of such penalties are repairing infrastructure, building houses and schools, engaging in waste disposal, growing crops, fixing roads, and improving access to water/electricity (Government of the Republic of Colombia, 2016, pp. 183–84).

Now, so far as I am aware, no contemporary society has been implementing the reconciliation theory systematically in the context of criminal justice. The 'traditional' Yoruba had also adopted deterrence and incapacitation measures (Bewaji, 2016, pp. 44, 175; Balogun, 2018, pp. 311–12), while the more 'modern' Nigerian state has not adopted anything reconciliatory. Rwanda used community service systematically only as a transitional justice measure, not for day to day criminal justice. And then Colombia also has proposed restorative sanctions only in the context of transitional justice, and has not even implemented them; so far as I have been told,[3] they remain at the level of policy, not practice. Furthermore, there has been no thorough philosophical exposition and defence of a reconciliatory approach that would prescribe it as the central, if not sole, justification of state punishment in a 21st century society. That is my aim.

To continue the articulation of my favoured reconciliation theory of state punishment, note that it includes a 'backward-looking'

[3] By Colombian participants in the Conference on Transitional Justice and Distributive Justice: Comparative Lessons from Colombia and South Africa; see note 6 below.

condition, and specifically appeals to elements of the expressive theory sketched above. However, reconciliation differs from that form of retribution in two important ways.

First, where it is possible for the burdens placed on an offender to do some good in the form of repairing relationships, they should. Recall that the retributive model does not 'look forward' in any respect; all that matters is that the penalties match the crimes that were committed, which tends to justify having people waste away in prison. In contrast, for the reconciliation theory, if penalties could compensate and reform, they must take that form.[4]

For examples of burdensome compensation, perhaps someone who has cheated on his taxes should be made to perform some dull tasks for the state revenue service. Maybe a person who has robbed a household should wear a uniform and serve as a neighbourhood-watch guard for a time. Possibly someone who has unjustifiably taken the life of a breadwinner should farm with his hands, providing sustenance to the victim's family. For examples of burdensome reform, a court should often prescribe mandatory therapy to get to the root of what caused the mistreatment of others, something that would be time-consuming and psychologically difficult. Consider as well penalties meant to instil empathy and an awareness of the consequences of actions, such as a judge sentencing drunk drivers to work in a morgue. Finally, there are the points that sometimes the hardship of punishment can itself be a way for offenders to appreciate how they have mistreated their victims, as well as that the guilt consequent to moral reform would also be a foreseeable burden that offenders should undergo.

There are admittedly some situations in which compensation would be impossible to effect, say, where an offender has killed his victim; no burden placed on him would be sufficient to make up for the harm done. There are also cases where reform would happen on its own, with the offender having had a proverbial 'come to Jesus' moment; no burden placed on him would be necessary to change the offender's character and thereby prevent him from doing any further harm. Even so, by the present account, the need of the state – and ideally of the offender – to disavow the crime would remain, continuing to justify punishment. That being said, where punishment would do some good in the forms of compensation or reform, it should, thus making a reconciliatory approach different from a retributive model.

[4] The following examples have been cribbed from Metz (2019, pp. 126–27).

A second respect in which reconciliation differs from retribution concerns which degree of burden constitutes an appropriate disavowal.[5] For standard versions of retribution, whether the desert theory, fairness theory, or the expressive theory, punishment must be proportionate to the crime that was committed. For those theories, there is some amount of harm or wrong done to victims, perhaps discounted by the degree of the offender's responsibility for it, and punishment should exactly match that amount. In contrast, my approach to reconciliation prescribes penalties that track the crime, in the sense that the worse the crime, the greater the penalty, but without involving the same degree of harm/wrong as the crime. Instead, it recommends a range of penalties somewhat below that amount. For example, suppose we assigned cardinal numbers to the gravity of a first-degree murder with 1000 (in the case of full responsibility), a theft with 100, and jaywalking with 1. Then, instead of punishing such a murder with a penalty weighted 1000, the appropriate penalty would be in a range of, say, 750-500, and theft would similarly be punished with 75-50 instead of 100. Offenders would invariably receive a sentence that is less than retributivists think is deserved or fair, for instance.

Although my aim in this section is to spell out the reconciliation theory, and not so much to defend it, I do note here one major motivation for favouring backward-looking penalties that track the crime but are not proportionate to it. It is that almost no societies in fact impose proportionate penalties. In jurisdictions that base punishment at least in part on the nature of the crime, none so far as I know seeks out a penalty equal in severity to, say, torture.

Having discussed similarities and differences between the reconciliation and retribution models, I now do the same for the reconciliation and protection models. Like the forward-looking theories, the reconciliation model has us look into the future to consider which penalties are justified. Like them, it holds that one reason to impose punishment is normally to prevent crime.

However, it differs from the protection theories in some crucial ways. As noted above, the reconciliatory approach does not take crime prevention to be a necessary condition for a penalty to be justified. Where no sort of punishment is expected to mend broken relationships, the reconciliatory approach deems punishment to be justified nonetheless because of the need to disavow the fact that relationships were broken in the first place.

[5] Here I do a bit more to flesh out a brief suggestion from Hampton (1988, p. 137).

Another salient difference between reconciliation and protection is the sort of good that should come from the imposition of penalties. While they both seek to prevent crime, the protection model makes deterrence and incapacitation central, whereas the reconciliation model does not. Instead, insofar as the latter prescribes doing what is likely to prevent further wrongdoing, it would have a court do what is expected to reform the offender's character, which is only one (and often secondary) element of the protection theory.

For a third key difference, remember that the good the reconciliation model seeks to promote is not merely to prevent crime (specifically in the form of offender rehabilitation), but also to make up for crime that has already taken place. A central reason for imposing one penalty rather than another is that it is expected to help compensate victims. The best sort of burden to place on an offender is often labour that would improve his victim's quality of life. Of course, it would be time consuming to oversee that kind of punishment; it is much easier simply to put someone in jail. However, that does not make imprisonment just, let alone the ideal.

Summing up, note how considerations of reconciliation are relevant to answering the cluster of four questions pertaining to the justification of state punishment. First off, who is it that should be punished by the state? In the first instance, the answer is those who have failed to relate to other people or the state in the right sort of way, requiring them to make amends. Second, why should the state be in the business of inflicting penalties on people? Roughly speaking, the answer is to express disapproval of the crime by getting the offender to *clean up* his own mess, *contra* imposing punishment for reasons of retribution or incapacitation, and to clean up *his own mess*, in contrast to doing so for reasons of general deterrence. Third, how severe should a penalty be? The answer is that it must track the crime in the sense of be a function of how grave it was, but that it should not be proportionate to it. Fourth, which kinds of penalties should the state mete out? The answer is usually not prison and fines, if a state has the resources to avoid them, since these are unlikely to compensate victims and reform offenders; instead, the norm should be productive burdens such as labour that is expected to improve victims' quality of life and offenders' character.

Although I hope that the reader finds this theory to be prima facie attractive, it is in the following section where I really make the case for it. I now show that the objections to protection and retribution theories are plausibly avoided by the reconciliation theory, giving us reason to consider it as a replacement.

4. Advantages of the Reconciliation Theory

I have in other work argued that the reconciliation theory should be found attractive by those who think that we have a dignity at least in large part because of our relational nature (see Metz, 2019). If what gives us a superlative non-instrumental value is substantially our capacity to relate positively or cohesively (an idea salient in the African philosophical tradition, but having a greater resonance, in at least the Global South), then it is natural to hold that the aims of punishment should be both to express disapproval when that value is degraded and to mend broken relationships. In contrast, I here aim to show that even those without such foundational moral commitments should find much attractive about a reconciliatory approach to sentencing, insofar as it avoids widely recognized problems with the rival protection and retribution models. In this section I demonstrate that the reconciliation theory articulated in section 3 avoids the objections to the theories discussed in section 2.

Recall the problems facing protection theories such as utilitarianism and defensive force accounts. One was that they seem to justify punishment of the innocent, roughly since, for these views, responsibility for wrongful harm is not necessary for one to be liable for punishment. Utilitarianism naturally supports strict liability, while the logic of defensive force permits it to be used against innocent threats. In contrast, there is intuitively no need to reconcile with someone who has not culpably broken a just law. If someone is truly not responsible for having caused or threatened harm to another, then no disavowal and moral reform are warranted. While compensation could be called for, the state need not force an offender to make it in a manner that is intentionally burdensome for him, which would be apt only in the case of someone responsible for wrongdoing.

A second problem with the protection theories is that they tend to prescribe certain extremely severe penalties that seem impermissible. That is a standard problem for utilitarianism, which is well known for entailing that severe burdens should be placed on an individual if it would produce trivial benefits for many others. In addition, given that the logic of defensive force permits one to kill in order to save innocent life, say, because the aggressor has forfeited his right to life, it appears also to permit the state to torture a torturer in order to prevent the innocent from being tortured. However, the reconciliation theory does not permit penalties that are overly harsh. On the one hand, disavowing crimes requires penalties for them to be within a certain range that tracks, but is less than proportionate to, their gravity.

That would forbid the death penalty for traffic infractions, on the one hand, and also normally rule out killing killers and torturing torturers, on the other. Furthermore, death and torture are not instances of productive burdens; they are not forms of labour that would serve the functions of compensating victims or reforming offenders' character. Hence, the backward-looking and forward-looking elements of reconciliatory sentencing do not seem to permit intuitively extreme penalties.

The third problem with implementing punishment in order to deter or incapacitate (or even reform without consideration of the need to disavow the past crime) is that sometimes overly light penalties, or no penalties at all, for serious crimes would be sufficient to produce the relevant consequences. All protection theories require minimizing the amount of harm inflicted on an offender, if sufficient to protect society. However, the principle that the more serious the crime, the more severe the penalty should be looks compelling. That need not entail a system of proportionate sentencing, but instead is consistent with one that tracks the nature of the crime in the manner of the reconciliatory approach spelled out here. Disavowing unjust ways of treating victims with a penalty that falls within a certain range that is pegged to what would be proportionate but is less than that sidesteps the problem of insufficiently harsh sentences as it applies to protection theories.

Turning to retributivism, it is open to its adherents to maintain that anything less than a proportionate penalty is unjust, i.e., to argue that the reconciliation theory retains the problem facing protection theories regarding overly light penalties for serious crimes. However, I believe in fact that retributivism must be jettisoned if we want to avoid the implication that killing, torturing, raping, and maiming are just penalties when imposed on (certain) killers, torturers, rapists, and maimers. Recall the objection that retributivism permits a literal eye for an eye. Weighty disavowal of egregious behaviour seems possible with penalties that are somewhat less than proportionate, whereas it would also be apt for those guilty of such horrific crimes to labour in strenuous ways expected to make their victims better off. Furthermore, consider the point that imposing such sentences arguably expresses, not just disavowal of the offender's misdeeds, but also that he is without a dignity, which is an unjust kind of treatment (see Hampton, 1988, pp. 136–37).

Turn, now, to the second objection to retributivism I had mentioned, that it is utterly unresponsive to an offender's character. Retributive theories have little leeway for prescribing differential penalties to a first-time offender and a second- or third-time offender;

supposing he has committed the same crime, the same proportionate penalty is required. The natural thing to say, however, is that the second- or third-time offender has not learned his lesson, and so merits a greater penalty, which the retributive theory cannot easily make sense of but which the reconciliation theory can. Working within the range of tracking penalties, it would be open to a judge adopting the reconciliatory model to prescribe ones more severe for those who have demonstrated recalcitrance.

The flipside of the retributive focus on a penalty proportionate to a past misdeed is the inability to impose a lighter penalty in the face of moral reform that has taken place after the crime but before the imposition of a sentence. For many, some kind of penalty remains appropriate, again in the ballpark of the severity of the crime, but one that is less than what would be apt for a shameless offender who lacks remorse. Again, working within the range of tracking penalties, it would be open to a judge to prescribe lighter ones for, and hence display mercy to, those who have atoned on their own.

Finally, the third problem with retributivism above is that it cannot account well for the intuition that one proper function of a criminal justice system is crime prevention. Some kind of expected benefit to society seems essential to justify the enormous expense and time of a criminal justice system, but retributivism is a strictly backward-looking theory; for it, the only reason to punish is to impose a penalty proportionate to the crime that was committed in the past (and hence is deserved, fair, etc). In contrast, the reconciliation theory includes forward-looking elements, prescribing the placement of burdens on offenders that, when possible, will do some good in the form of rehabilitating offenders so that they do not re-offend and of compensating victims so that in the ideal case the effects of the crime are nullified. Those benefits plausibly make it worth setting up a punishment institution, beyond the admitted importance of the state distancing itself from crimes that have occurred and expressing support for victims.

5. Conclusion: Disadvantages of the Reconciliation Theory to Be Considered

My aim is not, with a single essay, to convince anyone to change her mind about the justification of state punishment, but rather to articulate a plausible alternative to the two approaches that have dominated Western thinking for more than two centuries. Drawing on ideas about reconciliation that have been prominent in parts of the

Global South, I have advanced a novel theory of who should be punished by the state, why, how much, and in what ways that I submit should not be dismissed as a rival to the protection and retribution theories. In a nutshell, the right candidates for punishment are those who should atone for their misdeeds by undergoing burdens that disavow the way they treated their victims by tracking the crime and that in the best case have the effects of compensating the victims and rehabilitating the offenders.

Being a new approach, it naturally could use further development and consideration. For example, while I have provided reason to think that the logic of reconciliation rules out punishing someone for the sake of general deterrence, i.e., instilling fear amongst would-be offenders in society, does it permit punishing someone for the sake of special deterrence, that is, instilling fear in him so that he would not commit the crime again? Is the only justifiable mechanism to prevent crime the imposition of a burdensome sort of rehabilitation, or can scaring the offender also be a way to get him to 'clean up his own mess'?

For a second topic that deserves reflection, what should the state do if an offender refuses to engage in the requisite sort of punitive labour? Suppose he will not attend the prescribed therapy or do the work that would direct funds to his victim. What resources does the reconciliation theory have to address this problem? Here is one strategy worth considering. It might be that issuing threats and, if necessary, imposing hardships on this recalcitrant offender would be justified, not as a form of punishment, rather as a form of defensive force. In failing to submit to the appropriate penalties, the offender would be committing a new offence, where a threat of, say, indefinite detainment until he complies might be justified as a kind of non-punitive coercion. If the reader is inclined to hold that it would count as a kind of penalty, that need not mean that the reconciliation theory is incoherent for including two different sorts. After all, no one charged the TRC with incoherence when it offered human rights violators amnesty from any punishment if they fully confessed, but retributive penalties if they did not.

A third issue that arises is how well the reconciliation theory of how a judge should sentence offenders fits with our best understanding of which kinds of criminal laws a legislator should pass. The kind of thing that should be criminalized by a Parliament should naturally be the kind of thing that should be punished by a court of law. Now, which sort of behaviour merits a reconciliatory response? Presumably not purely self-regarding actions, i.e., those that do not directly harm or interfere with others and that instead harm or

degrade the person performing them alone. If there is no victim, then there is no need to inflict a penalty that would compensate a victim or would disavow the way she was treated. I, for one, find it a welcome implication of the reconciliation theory that actions such as drug use, which need not be discordant in respect of others, should be legal, even if they merit other responses from the state such as education and treatment centres. However, those who believe it can be justifiable to punish people merely because of the way they have treated themselves will disagree.

Fourth, there are admittedly real concerns about how practical it would be to implement the reconciliation theory on a daily basis. Although, as I pointed out, something like it appears to have been accepted by some Yoruba clans in Nigeria and as a transitional justice measure in Rwanda, that is different from using it routinely in a mass society. However, perhaps those currently employed as corrections officials, parole officers, and counsellors could be repurposed to ensure that offenders do the work of changing their beliefs, desires, and emotions that led to the crime as well as the work of doing what would make their victims better off. Where that kind of shift is not feasible, and plea bargaining combined with prison time are unavoidable due to the enormous numbers of criminals in the system, at least the reconciliation theory would plausibly tell us that a certain measure of injustice is present, providing a picture of to what a state should aspire.[6]

University of Pretoria
th.metz@up.ac.za

[6] Thanks to Julian Baggini for having shared written comments on a previous draft of this essay. For their oral input on some of these ideas, I am also grateful to participants in three gatherings: the Conference on Transitional Justice and Distributive Justice: Comparative Lessons from Colombia and South Africa, organized by the South African Institute for Advanced Constitutional, Public, Human Rights and International Law in 2018; a KJuris: King's Legal Philosophy Workshop, organized by the King's College London Dickson Poon School of Law in 2020; and the London Lectures Series: A Philosophers' Manifesto, organized by the Royal Institute of Philosophy in 2020. I have not been able to answer in this draft all the important queries I received from these colleagues, but look forward to continuing the debate in future work.

References

Oladele Balogun, *African Philosophy: Reflections on Yoruba Metaphysics and Jurisprudence* (Lagos: Xcel Publishers, 2018).
Jeremy Bentham, 'The Rationale of Punishment', (1830), https://www.laits.utexas.edu/poltheory/bentham/rp/.
John Ayotunde Bewaji, *The Rule of Law and Governance in Indigenous Yoruba Society* (Lanham, MD: Lexington Books, 2016).
Paul Christoph Bornkamm, *Rwanda's Gacaca Courts: Between Retribution and Reparation* (Oxford: Oxford University Press, 2012).
John Braithwaite and Philip Pettit, *Not Just Deserts* (Oxford: Clarendon Press, 1990).
Michael Davis, *To Make the Punishment Fit the Crime* (Boulder, CO: Westview Press, 1992).
R. A. Duff, *Punishment, Communication and Community* (Oxford: Oxford University Press, 2001).
Daniel Farrell, 'The Justification of Deterrent Violence', *Ethics* (1990) 100, 301–17.
Daniel Farrell, 'Capital Punishment and Societal Self-Defense'. In William Aiken and John Haldane (eds), *Philosophy and Its Public Role* (Charlottesville, VA: Imprint Academic, 2004), 241–56.
Joel Feinberg, *Doing and Deserving* (Princeton: Princeton University Press, 1970).
Government of the Republic of Colombia, 'Final Agreement to End the Armed Conflict and Build a Stable and Lasting Peace', (2016), http://especiales.presidencia.gov.co/Documents/20170620-dejacion-armas/acuerdos/acuerdo-final-ingles.pdf.
Brandon Hamber, Dineo Nageng, and Gabriel O'Malley, '"Telling It Like It Is…."; Understanding the Truth and Reconciliation Commission from the Perspective of Survivors', *Psychology in Society* (2000) 26, 18–42.
Brandon Hamber and Richard Wilson, 'Symbolic Closure through Memory, Reparation and Revenge in Post-conflict Societies', *Journal of Human Rights* (2002) 1, 35–53.
Jean Hampton, 'The Retributive Idea'. In Jean Hampton and Jeffrie Murphy (eds), *Forgiveness and Mercy* (Cambridge: Cambridge University Press, 1988), 111–61.
Douglas Husak, 'Why Punish the Deserving?' *Nous* (1992) 26, 447–64.
Immanuel Kant, 'The Metaphysics of Morals', Mary Gregor (trans.). In Mary Gregor (ed.), *Immanuel Kant: Practical*

Philosophy (Cambridge: Cambridge University Press, 1996), 353–603 (originally published in 1797).
Stephen Kershnar, *Desert, Retribution, and Torture* (Lanham, MD: University Press of America, Inc, 2001).
Thaddeus Metz, 'Reconciliation as the Aim of a Criminal Trial: *Ubuntu*'s Implications for Sentencing', *Constitutional Court Review* (2019) 9, 113–34.
Thaddeus Metz, 'Why Reconciliation Requires Punishment but Not Forgiveness'. In Krisanna Scheiter and Paula Satne (eds), *Conflict and Resolution: The Ethics of Forgiveness, Revenge, and Punishment* (Cham: Springer, 2022).
Philip Montague, *Punishment as Societal-Defense* (Lanham, MD: Rowman & Littlefield, 1995).
Michael Moore, *Placing Blame* (Oxford: Oxford University Press, 1997).
Jeffrie Murphy, *Retribution, Justice, and Therapy* (Dordrecht: D. Reidel Publishing Company, 1979).
Jeffrie Murphy, *Retribution Reconsidered* (Dordrecht: Kluwer Academic Publishers, 1992).
Penal Reform International, *Eight Years on…A Record of Gacaca Monitoring in Rwanda*, (2010), https://cdn.penalreform.org/wp-content/uploads/2013/05/WEB-english-gacaca-rwanda-5.pdf.
Wojciech Sadurski, *Giving Desert Its Due* (Dordrecht: D. Reidel Publishing Company, 1985).
Russ Shafer-Landau, 'The Failure of Retributivism', *Philosophical Studies* (1996) 82, 289–316.
J. C. Smart, 'Utilitarianism and Punishment', *Israel Law Review* (1991) 25, 360–75.
Andrew von Hirsch, *Past or Future Crimes* (Manchester: Manchester University Press, 1986).