To Appear in *Public Affairs Quarterly* 31.4, October 2107

FORTIFYING THE SELF-DEFENSE JUSTIFICATION OF PUNISHMENT

Zac Cogley
zaccogley@gmail.com

Abstract: David Boonin has recently advanced several challenges to the self-defense justification of punishment. Boonin argues that the self-defense justification of punishment justifies punishing the innocent, justifies disproportionate punishment, cannot account for mitigating excuses, and does not justify intentionally harming offenders as we do when we punish them. In this paper, I argue that the self-defense justification, suitably understood, can avoid all of these problems. To help demonstrate the self-defense theory's attraction, I also develop some contrasts between the self-defense justification, Warren Quinn's better known 'auto-retaliator' argument, and desert-based justifications of punishment. In sum, I show that the self-defense justification of punishment is more resilient than commonly supposed and deserves to be taken seriously as a justification of punishment.

Keywords: punishment; self-defense; desert; retributivism; David Boonin; Phillip Montague; Daniel Farrell

1 Introduction

My specific aim in what follows is to defend the self-defense justification of punishment against some recent criticisms by David Boonin.[1] But more generally, I want to use my discussion of Boonin to develop a more compelling self-defense-based account of the justification of punishment than has yet appeared.[2] Boonin's discussion of the self-defense solution is careful and clear; he thus provides an excellent articulation of some putative problems for the self-defense justification of punishment. Despite his concerns, Boonin notes that, "the self-defense solution is perhaps the most innovative and promising of the various nonstandard solutions that have been offered to the problem of punishment".[3] I believe he is correct. I aim to show the self-defense solution is even more promising than is widely supposed.

I'll begin by defining punishment and will also briefly discuss why and how it must be justified. I'll then give a brief outline of the self-defense theory of punishment's justification. In the bulk of the paper, I'll respond to four critiques Boonin makes of the self-defense theory: that it will justify punishing the innocent, that it justifies disproportionate punishment, that it can't account for mitigating excuses (especially in cases of provocation), and that even if inflicting harms on offenders can be justified via the self-defense theory, those harms won't qualify as punishment. Boonin's discussion thus provides a series of helpful challenges that, if met successfully, will help me to further develop the self-defense justification of punishment and defend it.

2 What is Punishment and What Kind of Justification Does it Require?

Punishment, as Boonin and I both define it, is the intentional, authorized infliction of reprobative, retributive harm on someone. These are conceptual, not normative claims. In defining punishment, neither Boonin nor I commit to claims about the appropriate justification of punishment (that the appropriate justification is a retributive one, for example).

In order to be punishment, the act in question must harm the person in question; if the act benefitted instead of harmed, we would categorize the act as rewarding, not punishing. The harm must also be intentionally inflicted—though it may be intentionally inflicted as a means to some other end, like deterrence.[4] Our focus is on legal punishment, so for the relevant purposes if the relevant act was not performed by an authorized state actor, that act would not count as punishment. Next, the harm the state actor imposes must be retributive: the harm must be inflicted in response to the person's having done a legally prohibited act. If the state preemptively puts someone in jail in order to prevent her from speaking at a rally for marijuana legalization, the person's rights may be violated, but she is not punished. Finally, what the state actor does must be reprobative: it must express condemnation of the person for doing the legally prohibited act.[5]

This last criterion is required to help mark off the difference between a fine and a fee. First, of course, a fine is imposed as the result of someone doing something that is legally prohibited, while a fee is a payment of money in exchange for a good or service. When the state requires people to pay a fee to renew their drivers licenses, marriage licenses, or marijuana user cards, the money paid helps to fund the state's regulation of a permitted activity. A fine, by contrast, is money that must be paid because someone acted in a way that is legally prohibited. An additional aspect of the difference between a fine and a fee is that the fine expresses opprobrium toward the person for doing the act. For example, if someone is fined for parking in a space reserved for people with disabilities, the fine expresses condemnation of her for doing so. That does not occur when someone pays the fee to renew her license plates.

The main philosophical problem presented by legal punishment is determining why the fact that a person broke the law justifies an authorized agent in intentionally, retributively, reprobatively harming her. Of course, it might not—which is Boonin's conclusion. Boonin claims that there is no satisfactory justification for punishment and that punishment should be abolished in favor of a system of mandatory restitution.[6] Theorists like myself who think that punishment can, in some circumstances, be justified need to explain when and why it is permissible to for a state actor to reprobatively harm someone guilty of breaking the law. Normally, the things we do to people in punishing them—like intentionally harming them by locking them away from the rest of society—would be wrong. Why would we be justified in doing those things to someone who broke a law? In the next section, I will briefly sketch how an analogy with self-defense helps us answer this question.

3 The Foundation of the Self-Defense Justification of Punishment

Consider what I will call an archetypal case of self-defense: a wrongful, fully-culpable attacker poses a threat to an innocent victim. Suppose that Kari travels to Colorado for vacation and then legally purchases some quantity of marijuana. Because Kari will not give him any of her marijuana stash, Chris attacks Kari. Most would agree that in such a situation Kari may act to ward off Chris's attack and that—if harming Chris is the only way to prevent herself from being harmed by his attack—Kari is permitted to harm Chris. Why is Kari permitted to harm Chris, or even kill him if he poses a lethal threat to her, if doing so is necessary to saving her own life? After all, normally we are not permitted to harm or kill others!

The most plausible explanation of Kari's right of self-defense invokes a principle of distributive justice:

> DJ: if it is inevitable that either A or B will be harmed, and if A's wrongful conduct is the reason it is inevitable that someone will be harmed, then justice permits acting so that A is harmed rather than B (so long as the harm that B suffers is not disproportionate to the harm that A would have suffered).[7]

Note that DJ does not justify Kari harming Chris unless she must do so to defend herself. Suppose Kari can either harm Chris or she can blow marijuana smoke in his face to confuse him and then escape. Then it is not inevitable that someone will be harmed, so the principle does not justify her harming him—call this the 'necessity' restriction. Importantly, DJ also applies to the actions of third parties. Suppose that Chris attacks Kari and she has no way to defend herself. However, Tony can harm Chris and thereby prevent harm to Kari. The principle permits Tony to intervene on Kari's behalf and harm Chris to prevent him from harming Kari. Finally, the principle requires that the harm imposed on Chris to defend Kari not be disproportionate to the harm Kari would have suffered—call this the 'proportionality restriction.' Thus, Chris may not be killed to prevent him from flicking Kari's ear.

While the principle justifies Kari harming Chris in self-defense, the reason is *not* that Chris *deserves* to be harmed or that it is good for him to be harmed (even granting these claims might be true).[8] The moral basis of DJ is justice: the proper distribution of benefits and burdens between individuals. One way to make out this claim might be in a Rawlsian manner.[9] People in an fair bargaining position choosing principles of justice to govern each other's behavior would choose a principle of justice that distributes harm to A rather than B. In other words, this is simply the distribution of benefits and burdens rational people would prefer when choosing from behind the veil of ignorance.[10]

Space does not permit me to further explore the moral basis of DJ or to discuss arguments in defense of it. (There is a rich literature on the justifiability of self-defense and on self-defensive killing in war.)[11] Therefore, I will simply assume, in what follows, that DJ offers the correct explanation of why and when an innocent like Kari would be justified in harming someone like Chris to prevent his wrongful attack.[12] Our question of interest, now, is how to extend this principle from the case of self-defense—where a harm will imminently occur unless defensive action is taken—to the case of punishment, where it would seem the harm done by the law-breaker has already occurred.

Multiple ways of applying DJ to punishment have been suggested. One begins from the observation that, as noted above, DJ allows for other-defense; it permits another party, C, to act so that A is harmed rather than B. And, as Phillip Montague suggests, the state can be a general instance of C: either the state enacts a system of punishment, or it doesn't.[13] If the state doesn't create and enforce a system of punishment, innocents will be harmed. If the state does create a system of punishment that has some deterrent value, some innocents will be spared from harm and instead some of those who persist in engaging in wrongful conduct will be harmed. No matter what the state does, then, some people will be harmed; the question is whether it will be innocent or wrongful parties. Justice permits the state using punishment to distribute harm away from innocent parties and onto parties who are responsible for wrongful harms.[14]

Another way to apply DJ to the case of state punishment emphasizes that even in individual self-defense cases there may be future harms made more likely by a past attack [15]. For example, suppose that Chris has successfully attacked Kari and stolen some of her marijuana. That fact might embolden some other unsavory vacationers to try to attack Kari in order to get some of the rest of her marijuana. "If Chris got away with it," they might reason, "no reason to think I won't!" Chris's attack thus might increase the probability of other harms to Kari in the future. Increased risk of harm is, itself, plausibly construed as a kind of harm.[16][17] Some of these future harms to Kari may be conditional on whether or not she harms Chris in retaliation for his attack, or not. Some reprobate vacationers may be deterred from attacking Kari if they know that she successfully retaliated against Chris. So it may also be true that Kari's harming Chris *after* he has already attacked will prevent future harm to her. Again, DJ also allows for third parties to act to defend innocents against suffering harm at the hands of culpable attackers. So Tony might justifiably harm Chris, if doing so will eliminate risks to Kari that have been increased by Chris's attack (and the harm Tony imposes on Chris is proportionate to the risk of harm Kari would otherwise suffer).

These observations also support the application of DJ to the case of criminal punishment. Besides the direct harm to victims already caused by the actions of those who break the law, other future harms are caused or made more likely by the actions of lawbreakers. For example, victims and those connected to them might appropriately fear a future attack by the offender, conditional on whether the offender is punished. The offender's attack may also reduce the overall level of security in a community, again conditional on whether or not the offender is punished.[18] In punishing an offender, the state certainly cannot prevent *all* harms which the offender's actions have already caused. However, some future harms are plausibly conditional on whether the offender is punished; the state may then be able to eliminate or mitigate these harms by punishing the offender. If so, the self-defense principle (DJ) will justify the state punishing the offender if the harms imposed on the offender are not disproportionate to the harms avoided.

Before continuing to Boonin's challenges, let me briefly distinguish the self-defense justification of punishment from Warren Quinn's 'auto-retaliator' justification.[19] At a great enough level of abstraction, Quinn's strategy shares a common structure with the self-defense justification. Quinn also attempts to justify punishment via analogy to an otherwise permissible activity. Quinn argues that we have the right to threaten to punish people for breaking the law by arguing that we would have the right to use 'auto-retaliator' devices that would mechanically identify and punish those who break the law. Quinn argues that the use of such devices is morally equivalent to setting up a system of punishment and that the use of such devices would be justified. If so, setting up a system of punishment would also be justified.

While Quinn's proposal has received much attention, his auto-retaliator devices are not analogous to punishment. Consider that once the device is 'programmed' and someone breaks the law, there is no further choice about whether or not to inflict harm. The devices *automatically* impose harm in such cases. As Quinn admits[20] and Boonin notes[21], this means that the harm imposed by the devices is merely foreseen—not intended—by those who deploy them. Since punishment is intentionally imposed on offenders, Quinn's devices are not directly analogous to punishment. In self-defense cases, by contrast, someone must choose whether or not to impose the relevant defensive harms, just as when we punish we must choose whether or not to harm someone who has already broken the law.[22]

4 Boonin's Challenges

4.1 Punishing the Innocent

Boonin's first objection is that the self-defense theory justifies punishing innocent people. If he is correct, and punishing the innocent is morally unacceptable, then the self-defense theory has implications that mean it must be rejected. Boonin suggests that the choice the state faces regarding whether and how to punish may be more complicated than initially suggested.

Suppose the state has these options:

(1) Do nothing in response to violations of the law; if so, many innocent people will be harmed.

(2) Threaten to harm and harm only people who violate the law; in which case some people will be deterred from breaking the law and some innocents will be spared from harm.

(3) Threaten to harm and harm people who violate the law as well as their children; if so, even more people will be deterred from breaking the law and so even more innocents will be spared from harm.

Boonin supposes that option (3) has the following implication: "for every one innocent child who will end up being harmed as a result [of punishing the children of offenders], five other innocent people (maybe even innocent children) will be saved from being victimized by offenders in the first place."[23] Plausibly, if punishment has some deterrent value, threatening to harm and harming both the children of offenders and offenders will have more deterrent effect than only threatening to harm and harming offenders. If this is so, Boonin urges, the self-defense theory will imply that the state should do everything it can to shift harm onto the guilty so the innocent avoid harm. Thus, the self-defense theory implies that the state should punish innocent people.

In responding to Boonin, it may be helpful to first briefly return to the moral basis of the principle of self-defense (DJ), above. DJ allows that someone can permissibly be harmed if that person's culpable action results in a situation where either an innocent person will be harmed or, alternatively, the culpable person is harmed. The DJ principle does not make claims about the overall goodness of states of affairs where the guilty suffer rather than the innocent—it is not a principle of desert. Instead, DJ makes the more claim that, if harm is inevitable, then it *fair* for harm to be shifted onto those who are guilty *for making that harm inevitable* in order to avoid harm to those who are innocent *with respect to that very harm*.

The DJ principle's clearest import is thus for situations where the specific harms that would result for innocents through the culpable actions of a person or persons can be avoided by harming those very culpable actors in response. Deirdre Golash makes this point with the following observation: "If Charles Manson's cellmate threatens my life, I may not kill Manson to distract him; and if I, with my near-perfect moral character, threaten Manson at gunpoint, he is justified in defending himself, regardless of his past crimes."[24] Golash's perspicuous example helps to show that—contra the overstatement of some self-defense theorists[25]—DJ does not embody an overall ideal of 'cosmic justice' where the world is more just if people with good characters get good things and less just if people with poor characters do well. Perhaps such claims are true, but DJ is a more limited principle that takes no stand on them.[26]

An advocate of DJ's extension to punishment, like myself, should thus insist that the move from situation (2), where only the guilty are harmed, to (3), where some innocents are harmed in order to avoid additional harms to other innocents, does not involve an application of the DJ principle at all. This can be seen by a more careful examination of the cases Boonin presents. Suppose, first, that the state does exactly, and only, what (2) suggests. Note that it is possible that the total aggregate harm to people in situation (1) might be exactly the same as that in situation (2). But DJ implies that situation (2) may permissibly be selected, because it is fair that harm falls on those culpable for the harm, rather than those not culpable for the harm. So the state is allowed to choose (2) over (1) because of considerations of justice, not because of considerations concerning reducing the aggregate level of harm in society. It might be even better if aggregate harm were reduced, but DJ only justifies acting to reduce aggregate harm if harm is imposed on those culpable for its inevitability and is not disproportionate.

So imagine that the state has enacted situation (2). Now an enterprising bureaucrat points out that the state can use the same mechanism set up for approximating fairness for a very different end: decreasing aggregate harm to innocents. This is the situation Boonin describes in (3). While the state would be using the system of punishment previously employed only in the service of justice, the state is now considering employing that system for a different end: roughly speaking, keeping more of those citizens who are undeserving of harm from suffering it. The state would thus no longer be only doing something justified by DJ: treating offenders harshly to prevent harms that would otherwise have resulted from the offenders' actions. Instead the state would additionally begin doing something DJ does not address: using harsh treatment of offenders to make a greater number of innocent people better off than they otherwise would have been.

Step away from this discussion for a moment to consider the charge that because the self-defense theory implies that Kari's use of a gun in self-defense against Chris is justified, the self-defense theory also justifies Kari's use of a gun to enact a Robin Hood-like scheme. (Kari's Robin Hood-like use of the gun involves threatening callous and unpleasant, but rightful, owners of marijuana in order to take their weed and give it to more agreeable pot-smokers.) The fact that DJ implies using a gun in self-defense can be justified does not have clear implications for whether using a gun for other very different (and perhaps even morally laudatory) aims is also justified. Similarly, the fact that punishment can be justified when employed for self-defense in a way that more justly distributes harm does not imply that using punishment for other aims is also justified—even if how punishment is deployed in those cases affects the distribution of harms.[27] Therefore, I conclude the self-defense theory does not justify punishing the innocent. It avoids Boonin's first problem.

4.2 Disproportionate Punishment

Boonin's second challenge for the self-defense theory is that it will justify punishments that are disproportionate to the severity of the offense. Boonin advances the challenge in two ways. He first develops the claim that in some cases the self-defense theory justifies severe punishment for minor offenses; in others, the self-defense theory does not justify punishing major offenses severely enough.

Let's first focus on the claim that the self-defense theory justifies too much punishment for some offenses. The initial argument for this claim is due to Larry Alexander,[28] who believes that the distributive justice principle (DJ) should not incorporate a proportionality restriction: the stipulation that defensive harm must not be out of proportion with the harm threatened by a culpable attacker. His argument against the inclusion of the proportionality restriction begins with the observation that in order to prevent the theft of his rose bushes he would be justified in moving them to a private island surrounded by sharks, so long as the threat is made clear to potential thieves. If this is so, Alexander urges, he would also be justified in constructing a shark-filled moat to protect the rose bushes. And thus, Boonin concludes,

> if it is permissible to build a lethal moat to protect one's roses, then, according to the argument that attempts to justify the self-defense solution, it must be permissible to threaten lethal consequences for sealing roses and, finally, to inflict those consequences on those who steal them.[29]

Thus, Boonin claims that the self-defense theory justifies the death penalty for minor infractions like rose theft.

The self-defense theorist should point out that these cases are not close analogies. There are a number of differences that may be morally relevant. Moving the roses to a private island takes advantage of a naturally and already occurring hazard, while building a lethal moat brings into existence a new and artificial one. The building of the moat also occurs on private property and, in order to be subject to the threat, a person must trespass on that property. The mere fact that death is being threatened for the same act in all the cases is not enough to show that building a shark-filled moat to prevent rose-stealing is relevantly analogous to the state's painfully executing someone for the same act. If the threat is made clear, it is easier for people to avoid being subject to an unpleasant shark death in the moat than if the state decides to institute shark death for stealing roses. Most important for our purposes is that a person who builds a shark-filled moat does not face the choice of whether to impose harm after they have made the decision to threaten harm—as is the case with both self-defense and punishment. Alexander's case is most directly analogous to

Quinn's auto-retaliator machine, and is thus not a case of self-defense. The more direct analogy to self-defense would be that Alexander builds the moat to protect his roses, sees a would-be rose stealer crossing the bridge, and *then*—realizing there is no other way to protect the roses—must consider whether to push the miscreant into the shark-filled moat. While self-defense theorists may concur with Alexander that he is justified in initially building his moat, they should deny that he is justified in pushing the would-be thief to a sharky doom because it violates the proportionality restriction. The argument against DJ's proportionality restriction based on the putative similarity of these cases should thus be rejected; the self-defense theory does not justify too much punishment.

Boonin's second proportionality challenge concerns the opposite claim: that in some situations the self-defense theory justifies too little punishment. Boonin develops this critique by considering the crime of arson. Suppose that a $500 file would deter almost everyone from committing arson. However, a small number of people are not deterred by this fine; the only punishment that would deter them is death or intense torture. Since the harm of intense torture is disproportionate to the harm of arson, the self-defense theory will not allow the state to impose it. Further, as Boonin correctly observes, the self-defense theory also implies that increasing the fine to $1000 is unjustified, for the $1000 fine provides no protection that is not already secured by the $500 fine. If this is so, Boonin charges the self-defense view with justifying only what many people will find to be "an unacceptably light sentence."[30] For the following reasons, self-defense theorists should be comfortable with this result.

The first relevant point is that any theory of punishment will be revisionary in the sense that it will be committed to results that some number of people will find intuitively jarring. While Boonin himself rejects the permissibility of punishment, his preferred alternative—the theory of pure restitution—has intuitive implications that many people will find implausible, at least on first look.[31] So the mere fact that in one instance many people find the implications of the self-defense theory to be intuitively suspect is not a compelling reason to reject the theory unless the intuitions can be justified in a way that casts doubt on the theory, itself.

Second, the clearest way to ground the intuition that people who commit arson should be punished with a $1000 fine even if a $500 fine provides equivalent deterrent effect is to claim that people who commit arson *deserve* to suffer more harm than simply a $500 fine. Again, however, the core principle of the self-defense view concerns fairness and the just distribution of harm; it is not a desert-based (or retributive) view.[32] The self-defense theory rejects appeals to desert to justify punishment and applauds the limits on punishment her account imposes. The self-defense view allows punishment of offenders only when the harm imposed on offenders is necessary to avoid harms that would have been suffered by innocents had the punishment not occurred.[33]

To further see the distinction between desert and justice justifications, consider how to understand the normative force of a desert claim versus a claim based on principles articulating just self-defense. Part of the attraction of the self-defense view is the idea that, considered from the standpoint of justice, the arsonist's interests still count in the sense that it wouldn't be better to harm her unless doing so allowed us to avoid some other harms for which she is responsible. By contrast, we can understand the claim that the arsonist deserves to be punished by saying it would be intrinsically good if she were punished[34] or, perhaps, that it is better that she be punished than not be punished.[35] Note that in either case, the desert theorist is making a value claim that holds independently of any effects that punishing the arsonist might have. Minimally, the desert theorist thus says that the world would be a better place—other thing equal—if the arsonist is punished even if no other goods come of it. The self-defense theory, by contrast, does not imply that punishment can be justified if it only serves to harm the offender.

Finally, consider that sometimes the empirical facts do not allow us to do what we would be morally permitted to do, were the facts different (as Boonin notes in another context). For example, if there were a willing donor we would be able to do a life-saving transplant. But since there isn't a willing donor we have no morally permissible way to save the patient's life.[36] Similarly, if a $1000 fine did provide more deterrent effect for arson, we would be permitted to levy that punishment on arsonists. But that is not the case here since, by hypothesis, the $1000 fine is equivalent in deterrent effect to the $500 fine. This case simply helps to demonstrate that "when offenders violate the law, it is not always permissible for the state to do everything that it would be morally permitted to do in response [were the facts different]."[37] For all these reasons, the self-defense theorist should be comfortable that her theory only justifies what some think is too light a sentence. Therefore, the self-defense theory can adequately respond to Boonin's second critique.

4.3 Mitigating Excuses

The third problem that Boonin presents for the self-defense theory also concerns the proportionality of punishment. It is common for the law to recognize provocation as a mitigating excuse, so that a person who attacks another after being provoked will typically not be subject to as much punishment as a person who attacks another without being provoked. Boonin charges that the self-defense theory will have very counter-intuitive results: he claims it will imply punishing provoked people *more* than unprovoked people![38]

Boonin's argument here rest on the plausible empirical claim that a greater level of harm will have to be threatened in order to prevent a provoked attacker than an unprovoked attacker. That idea is that people who are provoked 'lose their cool' in a way that makes them less likely to dispassionately consider potential costs and benefits of potential courses of action. Suppose, then, that given what punishments we have a right to threaten and impose, fining people $1000 for *unprovoked* assault is maximally permissibly deterrent. (Punishing assault with painful death would give us more deterrent value, but imposing a painful death is disproportionate to the harm suffered by potential innocent assault victims.) Suppose, further, that the maximally deterrent permissible punishment for *provoked* assault we have a right to threaten and impose is a $2000 fine. Then we are justified in imposing—and justice allows imposing—a greater fine for provoked assault than unprovoked assault. To many, this will intuitively seem like it gets the proportionality wrong. Provoked assault should be punished *less* severely than unprovoked assault.[39]

Now, one response for the self-defense theorist would be to emphasize the point that ended our discussion of the 2nd objection. Namely, the facts may conspire to prevent (or allow) us to do what we would not be justified in doing, were the facts different. For example, suppose we could permissibly impose a $10,000 fine for either provoked or unprovoked assault because that amount not disproportionate to the harm that would be otherwise be suffered by assault victims. If a $10,000 fine has significant deterrent value then equal fines of $10,000 are justified, even if it seems intuitively surprising to punish provoked and unprovoked assault equally.

This is a less satisfying response than it was in the discussion of arson, above. This is because the unintuitive nature of punishing provoked attacks more severely than unprovoked attacks can be developed in another way that puts significantly more pressure on the self-defense theory. That is to note that punishing provoked assaults more severely than unprovoked assaults is *unfair*, given the commitments of the self-defense theory. The self-defense theory emphasizes culpability to explain why innocent victims may be permissibly defended by imposing harms on their attackers. Someone who attacks without provocation is *more culpable* for the attack than someone who is provoked into doing so. Therefore, the problem presented by punishing provoked attacks more severely is more difficult than the above problem illustrated by the discussion of arson.

While this presentation of the problem is more serious, the self-defense theory has a solution. The self-defense theory can emphasize the very same points that Boonin does when discussing the presumptive difficulty presented by mitigating excuses for his own theory of pure restitution.[40] First, the self-defense theory can emphasize that the total amount of harm produced by unprovoked attacks is greater than that produced by provoked attacks. Physical and financial harms are likely identical for victims of assault regardless whether they provoked the assault or not, but the psychological suffering caused by an unprovoked attack is probably greater. More significantly, the secondary effects on other victims are greater in the case of unprovoked attacks. A person who engages in unprovoked attacks poses a greater threat to the objective security of other people in the community than does a person who only attacks when provoked.[41] Given that fact, the self-defense theory will imply that the state is entitled to punish unprovoked attackers more severely, since more harm can be avoided by punishing them than can be avoided by punishing provoked attackers.

The second—and related—way the self-defense theory can respond to Boonin's final challenge is to emphasize the moral salience of being fully, as opposed to partially, responsible for a harm. (This is a natural extension of the self-defense theory's emphasis on culpability.) As Boonin notes, if there is merit to the judgment that the provoked attacker should be punished less than the unprovoked attacker, that is presumably because the provoked attacker is less responsible for his actions than is the unprovoked attacker. And, if the provoked attacker is less responsible for his actions he is less responsible for the results of them, including the harms caused to innocent victims and the rest of society.

To see the relevance of this point, imagine a case where someone intentionally and maliciously throws two rocks toward you.[42] You have a shield which you can use to deflect both rocks back onto your attacker, thereby harming him instead of you. You can also deflect just one rock, or allow yourself to be hit by both. DJ justifies deflecting both rocks so your attacker is harmed by both rocks. Now imagine someone who is simply careless in throwing the two rocks. Just like before, you can either deflect both rocks, only one, or allow yourself to be hit by both. Additionally, suppose that the careless person is only 50% responsible for the potential harm to you. Then, DJ's proportionality restriction, which says that justice permits acting so that the culpable party suffers harm so long as the harm that the culpable party suffers isn't disproportionate to the harm the innocent party would otherwise suffer, permits you only to deflect one rock. Just because you are threatened by two rocks doesn't mean you can deflect both. Since the careless person is only responsible for the amount of harm that would be caused by one rock, you are only justified in deflecting back that amount of harm.[43]

Now return to provoked attackers. Suppose that a representative provoked attacker is only 50% responsible for the harm he threatens, while a representative unprovoked attacker is 100% responsible for the harm threatened by his attack. Then, the self-defense theory will allow punishing the provoked attacker only 50% as severely. The smaller amount of harm for which provoked attackers are culpable means that it would only be proportionate to impose a lesser amount of harm in response to what they do. Thus, the self-defense theory has two broad explanations of why we would be justified in punishing provoked attacks less severely than unprovoked attacks, both of which flow from the theory's emphasis on offenders' culpability for harms.

4.4 Harm vs Punishment

The final objection Boonin levies against the self-defense theory is that it does not justify harming offenders intentionally and reprobatively.[44] If these are essential elements of punishment—as both he and I accept—then the self-defense theory might justify a practice that superficially appears to be punishment, but isn't. Boonin presents the following case to demonstrate the problem. Suppose that I wrongfully throw a rock at you. Luckily, you're holding a shield which you can use to deflect the rock back onto me (causing roughly the same amount of harm as you would have otherwise suffered). So either you allow yourself to be hit by the rock or you deflect it back onto me. The principle of justice, DJ, would permit you to deflect it. But if you deflect the rock, the harm to me will be merely foreseen, not intended. Additionally, deflecting the rock doesn't seem to express disapproval of my act in any way—it would not be reprobative like punishment. If so, the practice DJ justifies might be a practice more analogous to one where harms are only imposed foreseeably and no condemnation is expressed. The practice justified by the self-defense theory would be more akin to quarantine[45] than punishment.

First, let me show that the self-defense theory of punishment can justify intentionally harming offenders. We need to be clear that in claiming that we punish intentionally the claim is not that punishment necessarily harms offenders *for its own sake*. For a harm to be intentional, it's sufficient that the harm be intended as a means to some other end. For example, a parent punishes a child by spanking even if the parent's ultimate aim is to deter the child from doing the act again.[46] That is because the parent has "chosen pain as a means to achieve her end, even if…she would prefer not to."[47] By contrast, when you defend yourself by deflecting my rock with your shield you have not chosen pain as your means. The means of your defensive action is your use of the shield. But in many cases addressed by DJ, intentional harm occurs. For example, in what we might call 'direct self-defense cases,' you prevent your own death by intentionally killing your assailant. In other cases covered by DJ, you chose pain as a means to your ends. For example, suppose you have already been attacked and injured by a wrongdoer. You can tell that, unless you do something unpleasant to your attacker—thereby 'making an example' of him—other potential attackers will be emboldened and will attack you. Supposing that the pain you cause to your already successful attacker is not disproportionate to the pain you avoid by making an example of him, DJ justifies intentionally harming him.

The self-defense theory can avail itself of these resources in explaining why punishment justified by DJ counts as intentional. When the state decides to punish, it has chosen pain as a means to its ends. As I've noted, one end the state might have in mind is deterring future offenders. There are other ways to deter than using punishment—we could put up signs warning potential offenders that they are being watched, for example. That wouldn't involve intentionally inflicting harm. But if, for example, what we're going to do to offenders is lock them away so they have little social and human contact with the hope that the *badness* of that situation will lead them and others to make different choices, we've chosen a harmful route to our end. This contrasts with quarantine, where we might lock people away with the hope that a disease will be stopped or crime will decrease because we decrease the amount of social interaction between people. Quarantines work just as well if the accommodations aren't onerous. Thus, it need not be part of our aim that quarantine is bad for those confined. When the state chooses deterrence, by contrast, it chooses a harmful method of achieving its goals.

Let me now consider Boonin's second concern. How can the harms DJ justifies express condemnation, when you don't convey disapproval of my rock-throwing by deflecting the rock toward me? To see the worry, recall the comparison with quarantine. Locking people away to prevent disease doesn't denounce. Punishment does denunciate, so we need to understand how DJ can justify harms that also convey condemnation. Here the self-defense theorist should emphasize that a practice of institutional punishment justified by DJ will still have the elements that make punishment express censure. For example, the self-defense theorist should urge the state to continue the practice of publishing a list of acts the state considers wrong and that punishment only be imposed if, at the end of a trial, the offender is found guilty of having committed one of those acts. When, after a finding of guilt, the state intentionally harms someone who did something they had no right to do, the state's act will express condemnation—no matter the justification for the state's action. The self-defense theory can therefore justify the intentional, reprobative harms that are distinctive of punishment.[48]

5 Conclusion

I've here defended the self-defense theory of punishment's justification against four recent challenges presented by David Boonin. The self-defense theory does not imply that we should punish the innocent, does not endorse disproportionate punishment, does not fail to accommodate a mitigating excuse for provocation, and does not fail to justify *punishment.* I said at the outset that my main aim in the paper is to show that the self-defense theory does not have these problematic commitments. I have also tried to show how strong the self-defense theory is, quite apart from Boonin's criticisms. To help show the theory's attractiveness, I also distinguished the self-defense theory's commitments from alternative theories of punishment's justification that are sometimes conflated with it, like desert-based and Quinn-inspired accounts. I believe that the self-defense theory is the most plausible theory of punishment's justification on offer, while at the same time being perhaps one of the most underappreciated. My larger hope is that this paper leads to recognition of the merits of a suitably articulated self-defense justification of punishment.

Northern Michigan University

Acknowledgements: Thanks to Antony Aumann, Kristopher G. Phillips, Andrea Scarpino, and Kari Theurer for bearing the burden of first read. I benefitted enormously from the friendly challenges and comments from an audience at the Rocky Mountain Ethics Congress, especially the commentary from Molly Gardner. Steve Munzer helped better frame the paper. Throughout the project, Daniel Farrell offered excellent advice. Finally, David Boonin was a friendly—and trenchant—interlocutor. I am additionally indebted to the Peter White Scholar Award from Northern Michigan University.

[1] *The Problem of Punishment.*

[2] Surprisingly—to this author, at least—the self-defense solution is often overlooked in the literature on punishment. To take one example, in his recent book, *Punishment,* Thom Brooks

neither mentions the self-defense justification of punishment nor cites any of the theorists who defend it. I have in mind work by Quinn, "The Right to Threaten and the Right to Punish"; Farrell, "The Justification of General Deterrence"; Farrell, "Punishment Without The State"; Farrell, "The Justification of Deterrent Violence"; Farrell, "Deterrence and the Just Distribution of Harm"; Cederblom, "The Retributive Liability Theory of Punishment"; Montague, *Punishment as Societal Defense*; Montague, "Recent Approaches to Justifying Punishment"; Kelly, "Criminal Justice Without Retribution"; Kelly, "Desert and Fairness in Criminal Justice." Boonin's attentive discussion of the self-defense justification is thus welcome. My hope is that this paper bolsters the case for taking it seriously.

A notable exception to this pattern is Victor Tadros' *The Ends of Harm*. There Tadros attempts to ground state punishment via a series of duties that wrongdoers incur as the result of their conduct. For example, Tadros implicitly accepts that as the result of wrongful conduct a person incurs a duty make an agreement with others who also pose risks of wrongful harm in order find someone to avert threats of harm for which she is responsible. Additionally, Tadros accepts that people then have duties to act on such agreements. Finally, Tadros accepts that third parties have the right to force the self-sacrifice involved in such an agreement even if the agreement isn't made Tadros, "Answers," 74–79.

Tadros attempts to ground these duties in a discussion of cases of self-defense *The Ends of Harm*, 169–264. But, as will become clear below, appealing to such duties is a major departure from how I appeal to self-defense in justifying punishment. For additional discussion of Tadros' account see the symposium on his book in *Criminal Law and Philosophy* (2015), especially Farrell, "Using Wrongdoers Rightly." In this paper, I directly demonstrate the resilience of the self-defense theory against Boonin's critiques.

[3] *The Problem of Punishment*, 198.

[4] Ibid., 14.

[5] Feinberg, "The Expressive Function of Punishment."

[6] *The Problem of Punishment*, 213–75.

[7] Similar principles are invoked by Montague, "Self-Defense and Choosing between Lives"; Montague, *Punishment as Societal Defense*; Montague, "Recent Approaches to Justifying Punishment"; Farrell, "The Justification of General Deterrence"; Farrell, "Punishment Without The State"; Farrell, "The Justification of Deterrent Violence"; Farrell, "Deterrence and the Just Distribution of Harm"; Cederblom, "The Retributive Liability Theory of Punishment"; Kelly, "Criminal Justice Without Retribution"; Kelly, "Desert and Fairness in Criminal Justice." Boonin's statement of the principle omits the proportionality restriction invoked by several of the authors Montague, *Punishment as Societal Defense*, 45–46; Farrell, "The Justification of Deterrent Violence," 302–3. Boonin's version of the principle is "if there is a situation in which it is inevitable that either A or B will be harmed, and if this situation is A's fault, then it is just to distribute the harm to A rather than to B" *The Problem of Punishment*, 196. I discuss proportionality in section 4.2.

[8] I further discuss the distinction between the self-defense justification and a desert-based justification in Section 4.2.

[9] Rawls restricts his theory's application to an ideal society with no need of punishment: *A Theory of Justice*, 8. However, this is a theoretical starting point; there is no in principle reason why the theory cannot be further extended.

[10] For a related line of thought, see Kelly, "Criminal Justice Without Retribution"; Kelly, "Desert and Fairness in Criminal Justice."

[11] Interested readers would do well to consult Jeff McMahan's work, especially his book *Killing in War*.

[12] Boonin appears to accept this assumption as well: *The Problem of Punishment*, 196.

[13] "Recent Approaches to Justifying Punishment," 25–26.

[14] Montague, *Punishment as Societal Defense*, 62–64; Cederblom, "The Retributive Liability Theory of Punishment," 307; Boonin, *The Problem of Punishment*, 196–97.

[15] Farrell, "Punishment Without The State," 444.

[16] Boonin, *The Problem of Punishment*, 251–53.

[17] Note that in individual self-defense cases the wrongful attacker may only be culpable for having increased the likelihood that the innocent party will be harmed. Thus, in such cases the innocent part is justified in actually harming the other, even though the other has only increased the risk of harm to the innocent.

[18] Boonin, *The Problem of Punishment*, 241.

[19] "The Right to Threaten and the Right to Punish."

[20] Ibid., 339–41.

[21] *The Problem of Punishment*, 206–7.

[22] Space does not permit a more detailed investigation of the difficulties that beset Quinn's attempt to justify punishment. In Section 4.2, I allude to other contrasts between with the self-defense justification and Quinn's theoretical framework. For more discussion, see Farrell, "The Justification of Deterrent Violence," 307–11; Boonin, *The Problem of Punishment*, 194–207.

[23] *The Problem of Punishment*, 201.

[24] *The Case Against Punishment*, 98.

[25] Montague, *Punishment as Societal Defense*, 47; Cederblom, "The Retributive Liability Theory of Punishment," 307.

[26] Additionally, the cosmic justice claim is not intuitively plausible as stated. Perhaps some of those with good characters were lucky to have upbringings that make it easy for them to act morally and some of those with bad characters try far harder to do the right thing. Such intricacies may complicate attempts to link just treatment with desert, but are not a problem for the self-defense theory, properly articulated. See Kelly, "Criminal Justice Without Retribution"; Kelly, "Desert and Fairness in Criminal Justice."

[27] Essentially, I am urging that the self-defense theory has more restricted application than Boonin demands. In a footnote, he anticipates this response and in reply claims that a defender of the self-defense theory cannot avoid endorsing using punishment as suggested in situation (3) because there are cases where virtually everyone agrees that it is justified to act to shift harm onto smaller numbers of innocents to avoid harm to larger numbers, as in the trolley problem: *The Problem of Punishment*, 201. But it is too much to ask of the self-defense theory, which concerns the proper use of defensive force by innocents against potential harms for which some people are culpable, to be a theory of the justifiable use of *all* force. There may be other cases in which the use of force is justified. But if those cases involve shifting harms no one is culpable for from some innocents onto other innocents—as in the trolley problem or in my last hypothetical—they are not within the scope of the self-defense theory.

[28] "Self-Defense, Punishment, and Proportionality."

[29] *The Problem of Punishment*, 202.

[30] Ibid., 203.

[31] Ibid., 218–75.

[32] Boonin argues at length against the idea that desert-based retributivism provides adequate justification for punishment, so it is surprising to find him appealing to intuitions he doesn't find to have significant probative force. See Ibid., 87–103.

[33] The arson case is thus actually a putative threat to DJ's 'necessity' restriction rather than the 'proportionality' constraint.

[34] Davis, "They Deserve to Suffer"; Moore, "Justifying Retributivism."; Hurka, "The Common Structure of Virtue and Desert"; Berman, "Punishment and Justification."

[35] Berman, "Rehabilitating Retributivism."

[36] *The Problem of Punishment*, 235–36.

[37] Ibid., 236.

[38] Strictly speaking, Boonin's claim is that the self-defense theory has two options: treat all cases of assault equally, or treat provoked and unprovoked cases differently. He argues that if self-defense theorists go with the first option they will be committed to the unintuitive idea that we should punish provoked and unprovoked offenses equally. I agree with him; the only plausible response for the self-defense theory is to distinguish provoked and unprovoked defenses, so that is the response I develop.

[39] The intuition that a $2000 fine is deserved or undeserved is given no role in the self-defense theory. So much the worse for that intuition, from the perspective of the self-defense theorist.

[40] Boonin, *The Problem of Punishment*, 256–59.

[41] All three of these are plausible, but ultimately empirical, claims.

[42] This case is due to Boonin (personal communication).

[43] Compare a situation where you are threatened by two rocks: one thrown by a malicious person and another that simply falls through natural causes. DJ only permits you to deflect the thrown rock toward the malicious person.

[44] In *The Problem of Punishment*, Boonin focuses only on the first problem: harming offenders intentionally *The Problem of Punishment*, 205–7. He raises the second problem in personal communication.

[45] Ibid., 22.

[46] Ibid., 14.

[47] Ibid., 14, fn 15.

[48] In response to the same concern—that his own theory of pure restitution will not allow the state to express condemnation of offenders—Boonin notes that even if pure restitution is instituted, the state will still hold offenders legally responsible for their unlawful behavior. As he notes, when currently the state acts to require restitution of an offender, "victims typically see restitution as in part a symbolic statement about what happened to them"—the statement being that the person "did something he had no right to do" Ibid., 268. Therefore, there is no reason the system of pure restitution cannot express condemnation. The same is true for punishment justified via the self-defense theory.

References

Alexander, Larry. "Self-Defense, Punishment, and Proportionality." *Law and Philosophy* 10, no. 3 (August 1991): 323–28.

Berman, by Mitchell N. "Punishment and Justification." *Ethics* 118, no. 2 (January 1, 2008): 258–90.

Berman, Mitchell N. "Rehabilitating Retributivism." *Law and Philosophy* 32, no. 1 (January 1, 2013): 83–108.

Boonin, David. *The Problem of Punishment*. Cambridge, Mass.: Cambridge University Press, 2008.

Brooks, Thom. *Punishment*. New York: Routledge, 2012.

Cederblom, Jerry. "The Retributive Liability Theory of Punishment." *Public Affairs Quarterly* 9, no. 4 (October 1, 1995): 305–15.

Davis, Lawrence H. "They Deserve to Suffer." *Analysis* 32, no. 4 (1972): 136–40.

Farrell, Daniel M. "Deterrence and the Just Distribution of Harm." *Social Philosophy and Policy* 12, no. 02 (1995): 220–40.

———. "Punishment Without The State." *Noûs* 22, no. 3 (September 1988): 437–53. doi:10.2307/2215712.

———. "The Justification of Deterrent Violence." *Ethics* 100, no. 2 (January 1990): 301–17.

———. "The Justification of General Deterrence." *The Philosophical Review* 94, no. 3 (July 1985): 367–94.

———. "Using Wrongdoers Rightly: Tadros on the Justification of General Deterrence." *Criminal Law and Philosophy*, 2015, 1–20.

Feinberg, Joel. "The Expressive Function of Punishment." In *Doing and Deserving: Essays in the Theory of Responsibility*, 95–118. Princeton: Princeton University Press, 1970.

Golash, Deirdre. *The Case Against Punishment*. New York: New York University Press, 2005.

Hurka, Thomas. "The Common Structure of Virtue and Desert." *Ethics* 112, no. 1 (2001): 6–31.

Kelly, Erin I. "Criminal Justice Without Retribution." *The Journal of Philosophy* 106, no. 8 (2009): 440–462.

———. "Desert and Fairness in Criminal Justice." *Philosophical Topics* 40, no. 1 (2012): 63+.

McMahan, Jeff. *Killing in War*. New York: Oxford University Press, 2009.

Montague, Phillip. *Punishment as Societal Defense*. Lanham, MD: Rowman and Littlefield, 1995.

———. "Recent Approaches to Justifying Punishment." *Philosophia* 29, no. 1–4 (May 1, 2002): 1–34.

———. "Self-Defense and Choosing between Lives." *Philosophical Studies* 40, no. 2 (September 1981): 207–19.

Moore, Michael S. "Justifying Retributivism." *Israel Law Review* 27, no. 1–2 (January 1993): 15–49.

Quinn, Warren. "The Right to Threaten and the Right to Punish." *Philosophy & Public Affairs* 14, no. 4 (October 1, 1985): 327–73.

Rawls, John. *A Theory of Justice*. New York: Oxford University Press, 1999.

Tadros, Victor. "Answers." *Criminal Law and Philosophy* 9, no. 1 (2015): 73–102.

———. *The Ends of Harm*. New York: Oxford University Press, 2011.