DOI: 10.31009/LEAP.2023.V10.01

Making Punishment Safe: Adding an Anti-Luck Condition to Retributivism and Rights Forfeiture

J. SPENCER ATKINS
Binghamton University

ABSTRACT

Retributive theories of punishment argue that punishing a criminal for a crime she committed is sufficient reason for a justified and morally permissible punishment. But what about when the state gets lucky in its decision to punish? I argue that retributive theories of punishment are subject to "Gettier"-style cases from epistemology. Such cases demonstrate that the state needs more than to just get lucky, and as these retributive theories of punishment stand, there is no anti-luck condition. I'll argue that Gettier-style cases demonstrate an impermissible instance of punishment, even though they meet the conditions of retributive theories of punishment. Retributive theories are therefore too weak. The safety condition from epistemology provides the anti-luck condition needed for permissible punishment. I argue that two forms of retributivism, rights forfeiture and what I call standard retributivism, are both subject to Gettier-style cases. Unlike the literature on standards of proof, this paper argues that safety is a condition on punishment itself, i.e. in all nearby possible worlds, the accuser must correctly accuse the convicted for the crime they actually committed.

Keywords: retributivism, safety, rights forfeiture, punishment, Gettier cases.

1. INTRODUCTION

Retributive theories of punishment argue that punishing a criminal for a crime she committed is sufficient reason for a justified and morally permissible punishment. Some retributivist theories—desert-based retributivists—even argue that punishing the guilty is intrinsically good. But what about when the state gets lucky in its decision to punish? In this

paper, I argue that retributive theories of punishment are subject to "Gettier"-style cases from epistemology. Such cases demonstrate that the state needs more than to just get lucky, and as these retributive theories of punishment stand, there is no anti-luck condition. I'll argue that Gettier-style cases demonstrate an impermissible instance of punishment, even though they meet the conditions of retributive theories of punishment. I'll then suggest that the safety condition from epistemology provides the anti-luck condition needed for permissible punishment. I argue that two forms of retributivism, rights forfeiture and what I call standard retributivism, are both subject to Gettier-style cases.

In the first section, I lay out the rights forfeiture theory of punishment and standard retributivism—an account of retributivism that all sorts of retributivism share. I then, in the next section, pivot to arguments for a relatedness condition on permissible punishment. This is a condition which states that the guilty must be punished for the crimes they committed and not some other crime. I also argue for what I call the "fine-grained claim", which states that accusations in criminal courts must be fine-grained enough to capture the relevant details of the crime. Next, in the third section, I present *Lexus*, a Gettier-style case where, by sheer luck, the state correctly accuses me of theft. The fourth and fifth sections demonstrate why *Lexus* is a problem for both rights forfeiture and standard retributivism, respectively: *Lexus* shows that these theories are too weak as they stand. I then turn to a solution: the safety condition from epistemology provides the needed anti-luck condition for theories of punishment. I argue that the anti-luck condition can take two forms: first, the accusation could not have easily been false (or the accusation is true in nearby possible worlds where the state accuses) *or*, second, the state's accusation is safely related to the crime that actually transpired. Thus the safety condition can either be a standalone condition for permissible punishment or it can modify the relatedness condition I argued for the second section. The last section entertains objections and responses.

2. WHAT IS THE RIGHT FORFEITURE THEORY? RETRIBUTIVISM?

This section outlines two plausible retributive theories of punishment: theories that offer the necessary and sufficient conditions for what I call "permissible punishment". I'll, first, address rights forfeiture and then I'll lay out what I call "standard retributivism". The goal here is to lay out these positions so we can see how they are in need of an anti-luck condition. But first, what is permissible punishment?

Various theories of punishment purport to show what conditions need to be met in order for some instance of punishment to be morally permissible. David Boonin (2008), for instance, thinks roughly that punishment is the deliberate infliction of harm on someone. This would normally be prohibited morally. The theories of punishment demonstrate what conditions need to be met in order for punishment to be morally justified. That is, they tell different stories about what constitutes "permissible punishment". With this clarification in mind, I now turn to rights forfeiture.

We have a number of rights, many (or all) of which we can forfeit. We forfeit rights all the time. If I promise you, for instance, to march with you in your protest against the Grammys, then I have forfeited the right to do as I please at the time of the protest. Had I not made the promise to you, I could have done a number of things—rock climbing, chopping wood, paper writing—during that time. But because I made you the promise, I have willingly forsaken my right to do those things at the time of the protest, such that I am blameworthy if I do them at the time of the protest.

Proponents of the rights forfeiture theory of punishment think that punishment works the same way. They argue that wrongdoers (or the guilty) forfeit rights—rights against hard treatment, say—in virtue of their wrongdoing. I have a right against being imprisoned. If I, according to this view, assault someone else, I forfeit my right against some level of imprisonment, rendering it permissible for the state to punish me with hard treatment. I need not intend to give up these rights, nor know that I am giving up these rights, according to the rights forfeiture theorist. Merely culpably committing a crime is sufficient for giving up rights against hard treatment.

Christopher Heath Wellman (2012) argues that punishment "would be permissible only if it violated no one's rights" (372). Wellman identifies a necessary condition on permissible punishment. As I said above, we have rights against hard treatment, such as the actions that constitute punishment. If punishment violates these basic rights, then it is impermissible. We have a basic right against hard treatment; it would be wrong, for instance, for the state to put me in prison, if I did not forfeit my right against hard treatment. The rights forfeiture theory, according to Wellman, stands in a privileged position among theories of punishment, since it is the only one that can make sense of this right against hard treatment. Other theories, Wellman argues, violate the basic right against hard treatment because they do not require wrongdoers to forfeit this right.

Retributivism, on the other hand, is the view that breaking the law is a

sufficient moral reason to punish.[1] Michael Moore (1993) writes that it is the view that we ought to "punish offenders because and only because they deserve to be punished" (15). The good effects of punishment—moral education, deterrence, incapacitation—are merely "icing on the cake" or a "happy surplus" of good effects, according to the retributivist (15). Breaking the law provides, according to the retributivist, a *pro tanto* justification for punishment. As a *pro tanto* justification, retributivism leaves room to deny punishment if there were greater, tragic consequences.

Moore goes on to clarify that breaking the law is a necessary and sufficient condition for punishment. But this is not all: breaking the law *obligates* the rest of us to punish you. It is not merely *permissible* to punish on this view. Retributivism substantiates a moral obligation to punish. David Dolinko (1991), however, disagrees with Moore's view here. He believes that retributivism only makes it permissible for us to punish, so, consequently, we do not have to punish. Dolinko sets up the following distinction between these different retributivisms: bold and weak retributivism. I draw out this distinction to help me clarify my own neutral view of retributivism shortly.

Moore notes that the tension between retributivism and consequentialism is needless. Retributivism is compatible with both deontic and consequentialist frameworks. For instance, a consequentialist may believe that punishing the guilty is intrinsically good. She, consequently, may want to optimize that value. Many deontic retributivists—by contrast—will think punishing the guilty is consistent with an agent-relative norm. This camp can still think that it is intrinsically good to punish, yet pursuit of this good is constrained by agent-relative norms. Moore's distinction is fascinating and insightful—yet I limit my discussion merely to what I call "standard retributivism".

I have drawn a few distinctions between different camps in the retributivism literature. For my coming argument to work, I need not take sides here. The term I use to refer to retributivism is "standard retributivism". Standard retributivism is the collection of central claims about retributivism, i.e. the intrinsic goodness of punishing the guilty and the fact that some moral state—an obligation to punish the guilty, or the permissibility of punishing the guilty—follows from criminal activity. It therefore stays neutral about the weak/bold retributivism debate.

There is another distinction I need to consider: desert-based

[1] Consider two noteworthy exceptions. First, under unjust or illegitimate law, retributivists might say that breaking the law does not produce a sufficient reason for punishment. Second, some reasons for punishment may be minuscule and thus not worth the effort, e.g. jaywalking on a deserted street. Though there is reason to punish, it would not be worth it.

retributivism and fairness-based retributivism. I have largely been addressing desert-based retributivism, which states that culpable criminal behavior generates a reason to punish, and this punishment is intrinsically good. According to fairness-based retributivists, such as Richard Dagger (1993), culpable criminal behaviors generate reasons for punishment, not because punishment is intrinsically good, but because fairness demands that we sometimes punish. Punishment, according to Dagger, is justified "because it is necessary to the maintenance of the social order" (475). I draw this distinction to once again establish that my target is standard retributivism; fairness-based retributivism, too, is subject to luck-based worries.

Standard retributivism, then, is the view that culpable wrongdoing generates a reason to punish (to a large extent) regardless of the consequences that follow from the punishment. This view, as well as rights forfeiture, are the targets of my critique. I now turn to additional concepts to help me establish the need for a safety condition: the relatedness condition and the fine-grained claim.

3. THE RELATEDNESS CONDITION AND THE FINE-GRAINED CLAIM

I now turn to a widely accepted claim that is central for my argument: if the guilty are to be punished, they can only be punished for the crimes they committed. A punishment is permissible only if it is a response to what the wrongdoer has actually done. Call this the "relatedness condition on permissible punishment". This condition implies that it is not permissible for the state to punish the guilty for a crime they did not commit. Any plausible theory of punishment must accept this claim. I argue that standard retributivism and rights forfeiture are committed to the relatedness condition.

Consider, for example, that authors criticize (pure) consequentialist theories of punishment because they imply that it is permissible to punish the innocent—in scenarios where punishing the innocent will, say, increase deterrence. Call these objections "innocent-punishment objections".[2] One reason that consequentialists are subject to innocent-punishment objections is that they do not accept the relatedness condition. It may be permissible, according to pure consequentialist theories, to punish some guilty person for a crime that she did not commit, so long as doing so would have beneficial consequences. The retributivist can accept

[2] For innocent-punishment objections, see Tadros (2011) and Boonin (2014).

the relatedness condition in order to avoid innocent-punishment objections. But she need not, because she has other options. A retributivist could say we cannot punish the innocent, yet at the same time say we can punish the guilty for any reason. While this may be one reason to think that standard retributivists must accept a relatedness condition, let me offer a couple of other reasons.

To further my case that retributivists must accept a relatedness condition, I turn again to Wellman. Wellman thinks that rights forfeiture theory runs into a problem: if I forfeit a right against hard treatment in virtue of culpable lawbreaking, then, an objector may argue, I can be punished for any crime that is proportional to the crime I actually committed (2012: 380). Suppose that prosecutors frame me for a crime that I did not commit, yet, unbeknownst to them, I committed a crime that would warrant a proportional punishment. That strongly seems problematic. Suspicious prosecutors could merely "cook up" fresh accusations—by framing defendants—after cases where defendants are let off. But if there is no relatedness condition on permissible punishment, then the framing cases are instances of permissible punishment according to the forfeiture theorist—and that's a problem.

Now, we may think that a similar case can be posed against the standard retributivist. If I have culpably committed some crime, then I should be punished, according to the retributivist. But it strongly seems as though I ought to be punished for the crime I have committed. That is, there needs to be a correct connection between the reason that we punish a wrongdoer and what it is that the wrongdoer has done. Otherwise, we might think that retributivists fall prey to cases like the one above.[3]

The literature on the aggregate probabilities principle highlights how dominant (some principle like) the relatedness condition is in legal practice.[4] According to the aggregate probabilities principle, when multiple charges are brought against a defendant we can meet standards of proof by aggregating the probabilities of all the charges and charging him for "an unspecified offence" (Harel and Porat 2009: 263). Alon Harel and Ariel Porat write that "the court[s] could use [this principle] to examine all

[3] Expressivist theories of punishment—a family of retributivist theories— seem especially committed to the relatedness condition. This is because, according to this view, punishment is a communicative condemnation of certain behaviors. Robert Nozick writes, "retributive punishment is an act of communicative behavior", which has two goals: "connect the wrongdoer to value qua value" and to connect the wrongdoer such that "the value qua value has a significant effect in [the wrongdoer's] life, as significant as his own flouting of correct values" (1981: 320, 376-7).

[4] For more about the aggregate probabilities principle, see Schauer and Zeckhauser (1996), Levmore (2001), Harel and Porat (2009), and Posner and Porat (2012). For more about the formulation, see Nau (2001). For objections, see Menashe (2014).

charges in aggregate and decided whether the standard [of proof] is met with respect to *at least one* charge" (262). Suppose that a defendant is charged with two crimes—rape and pickpocketing—and that the probability that he committed each crime is .9. If we use the aggregate probabilities principle, the probability he committed at least one of these crimes is .99.[5]

The relatedness condition, as I've described it, seems to come into tension with the aggregate probabilities principle. If the defendant is charged with an unspecified crime, this seems to undermine the sufficient relatedness required for the relatedness condition. Harel and Porat (2009) argue, however, that the aggregate probabilities principle sometimes satisfies the relatedness condition, viz. in situations where the defendant is charged with two counts of the same crime.[6]

It is worth noting that Eric Posner and Ariel Porat (2012) critique courts' use of the aggregate probabilities principle. Posner and Porat note that there are prohibitions against aggregation, but the noteworthy number of exceptions to this rule renders the courts blameworthy. With respect to criminal law, aggregation seems permissible only if the relatedness condition is met.

Consider another argument for a relatedness condition. A wrongdoer is a wrongdoer in virtue of some culpable behavior. The predicate "is a wrongdoer" is made true of someone in virtue of some specific actions they have culpably done in the past. That is, the culpable behaviors explain the fact that someone is a wrongdoer. Now, if a court punishes a person, they identify that person as a wrongdoer and respond to them as such. I think that the court must be able to identify what *makes* this person a wrongdoer, i.e. the events and actions that ground or explain the convicted's status as a wrongdoer. When I, for example, respond to another person as, say, a philosopher, I have some notion about what it is that *makes* them a philosopher, e.g. the number of articles and books she has published, her position at a university, her off-putting and condescending way of arguing. In a similar way, I think the courts must respond to wrongdoers on the true grounds of what makes the wrongdoer a wrongdoer. The courts punish only if they respond to the accused for the culpable crimes that they have

[5] Consider Harel and Porat's explanation: "The probability that the defendant committed each one of the offenses is .9 and therefore the probability, for each one, that he did not commit the offence is 1 - .9 = .1. Consequently, the probability that he did not commit any of the offences is (.1)2 = .01, and the probability that he committed at least one of the offences is 1 - .01 = .99" (2009: 262 n. 3)

[6] Menashe (2014) argues that, even in such circumstances, the aggregate probabilities principle does not satisfy the relatedness condition.

actually committed, i.e. the facts that make them a wrongdoer. Let's call this the "grounds argument".

Why think that the state must recognize a wrongdoer's grounds for being a wrongdoer? I think this is a matter of respect. We all have a qualitative identity. There are elements of our lives that make us who we are. We desire to be seen in the ways we see ourselves. I want to be seen as, say, a philosopher. Those who recognize me in that way affirm my identity and thereby respect me. But this does not go far enough. I want to share—and be seen for—the reasons that make me a philosopher. I think this is a matter of respect. There is an analogous sort of respect that must play out in the courtroom: the courts must not only see the wrongdoer as a wrongdoer, but also for the reasons that make her a wrongdoer.

My explanation is especially prevalent when we consider the responsibility to hear our friends' reasons why they have done something wrong. If my friend cheats on his spouse, I owe it to him to hear his reasons for becoming a "cheater". That is, I owe it to him to understand the reasons that make him a culpable wrongdoer. I can assess those reasons as I see fit, but I must at least understand them. I think there is some analogous responsibility for the courts to understand what grounds the wrongdoer's status as a wrongdoer, i.e. the courts must understand the grounds before punishment is permissible.

It is reasonable to think that the courts express disapprobation on behalf of the community. That is, one of the functions of the court is to condemn anti-social behaviors and express prevailing community values. However, the court must also discern how those values have been undermined by the suspect's behavior. To do this, the court must understand the reasons—the suspect's actions in particular—that make the suspect a wrongdoer. Once those reasons have been (accurately) identified, the court can then condemn the wrongdoer's behavior.[7]

The grounds argument goes some distance to establish a relatedness condition on permissible punishment. According to this argument, the court must understand what it is that makes a wrongdoer a wrongdoer and then respond accordingly. This is, I think, closely related to the relatedness condition. The court can only punish the wrongdoer for the reasons that make them a wrongdoer. If, consequently, I were framed for a crime that I

[7] We may worry that this picture isn't necessary for retributivist theories. According to retributivism, we would punish regardless of whether or not prevailing community values are communicated. This communication is not central to the justification of punishment. This is all true. My goal, however, is to show that identifying a wrongdoer for the reasons that make them a wrongdoer is a plausible concept generally, not necessarily tied to any particular theory of punishment.

did not commit, then the court would bypass the real reasons that make me a wrongdoer, and consequently fail to respond to me in an appropriate way. If this argument holds, then it strikes me that there is plausibly a relatedness condition on permissible punishment. This is my (and perhaps the courts') worry above with the aggregate probabilities principle.

I now turn to another claim the standard retributivist must accept. We may wonder about how coarse-grained or fine-grained an accusation of the guilty must be. For instance, I may have stolen a car. But suppose I am charged with stealing a different car than the one I actually stole. This seems like a problem. The state must, I think, correctly connect features of the crime to the accusation. Thus I am not merely charged for stealing *a car*; instead, I am charged with the fine-grained charge of stealing *this particular car* at some time. Call this the "fine-grained claim".

A retributivist may reject the fine-grained claim: if coarse-grained accusations are more effective at getting the guilty their due, then so much the worse for fine-grained accusations. So long as the accusation is "close enough" to the crime actually committed, then such an accusation will be permitted according to standard retributivism. This response is plausible yet misguided. Consider that the law itself is largely fined-grained. Charging the guilty with a coarse-grained accusation not only runs the risk of disproportionally harsh or light punishments, but it also undermines the relatedness condition, which I have been at pains to show is a plausible condition on permissible punishment.

The state, I have argued, must establish a connection between my actual culpable crime and the punishment it gives me: I cannot justly be convicted of a crime that I did not commit. The fine-grained claim ensures that the state has an accusation sufficiently detailed enough to identify the crime I committed. Thus the relatedness condition and the fine-grained claim go hand in hand.

Suppose I am charged with armed robbery. If I am merely charged with armed robbery, then other details are left out, e.g. whether I was successful, whether or not I stole money, how much money I stole, etc. These details are relevant for giving me my due, not only because they may affect the severity of my punishment, but also because they help establish the relatedness condition. Such details, that is, may be relevant for both determining the degree to which I am to be punished, and establishing my actual, culpable wrongdoing as the reason that I am punished.

Let me qualify the fine-grained claim a bit. So long as the state has access to some evidence relevant to refining the accusation, the state must

include that evidence in the accusation. Thus if there is some evidence—for instance, if I destroy the object that I stole—that the state cannot have access to, then the fine-grained claim is softened to some degree. Call this the discoverable qualification. We might also think that the degree of fine-grained-ness depends on whether the fineness is relevant to the case. Therefore, if the state does not get the exact second that I committed armed robbery correct, that's okay. This is because the time at which I committed the crime is not directly relevant to the degree of punishment that I will be subjected to.[8] Call this the relevance qualification.

The forfeiture theorist must also accept the relatedness condition and the fine-grained claim for similar reasons. The forfeiture theorist thinks that we forfeit rights in virtue of our wrongdoing. This means that the state must identify me correctly as the perpetrator of my own crime, namely, that the state must relate what makes me a wrongdoer (or what explains why I have forfeited a right against hard treatment) to its reason for my punishment. Thus the forfeiture theorist must accept something like the relatedness condition. Moreover, accusations must be fine-grained enough to capture the right that we have forfeited. Thus the forfeiture theorist must also accept the fine-grained claim.

Now that I have gone some way to establish the relatedness condition and the fine-grained claim, I now argue that these theories of punishment are subject to Gettier-like cases from epistemology. Some situations are problematically tainted by luck, such that retributivist conditions and right forfeiture conditions are insufficient for permissible punishment.

4. GETTIERIZING RIGHTS FORFEITURE AND STANDARD RETRIBUTIVISM

This section demonstrates the need for an anti-luck condition on permissible punishment. There is an extensive literature on Gettierizing *the burdens of legal proof* (see Pardo 2005, 2010, 2011; McBride 2011; Moss 2017). Unlike this literature, I suggest that we can Gettierize the accusation and reason for punishment. If my argument is successful, it shows that safety is a requirement not only for burdens of proof—as some authors say—but for the act of accusation itself. I propose a case that shows that standard retributivism and rights forfeiture are both too weak. That is,

[8] Time is a relevant factor in the degree of punishment administered. For instance, if robbed the store twenty years ago, that will lessen the degree of punishment that I will receive. The example here is about a small range of times. It won't affect the degree of punishment, say, if the state says I robbed the store at 7:42 AM, when I actually did it at 7:43 AM.

merely forfeiting a right or merely having done a crime are insufficient for permissible punishment.

First, consider a Gettier case of legal proof from the literature:

> *Framed Defendant*: The police arrest a motorist and plant drugs in his car. He is convicted at trial of illegal possession based solely on testimony from the arresting officers and the planted drugs. As it turns out, the defendant did have illegal drugs in his car at the time that were never discovered. The verdict that the defendant possessed drugs is therefore both true and justified (that is, the evidence at the time of the trial is sufficient to establish conviction beyond a reasonable doubt), but the truth and the justifying evidence are disconnected. The truth of the verdict is purely coincidental or accidental. (Pardo 2010: 50)

Rightfully, this case shows that the disconnect between the evidence and the truth makes the verdict problematic. This case "suggests that whatever is required for *knowledge* beyond JTBs is also required for legal verdicts to achieve their goal" (50). Pardo therefore suggests some anti-luck condition on legal evidence, e.g. safety.[9] I want to suggest that safety is not merely a requirement for legal evidence. In the next section, I argue that the courts must "safely" connect the correct evidence to the correct crime, e.g. no accidents must occur in nearby possible worlds. That means that punishment is permissible in all nearby possible worlds.

Let us now turn to similar, Gettierized case:

> *Lexus*: In the middle of the night, I go to my local *Lexus* dealership, break into the offices, steal the car keys, and drive off in a brand-new *Lexus* worth $50,000. Unbeknownst to me, another car thief, Angela, was at the dealership at the exact same time as I was. Angela stole a *Lexus* model also worth $50,000. Suppose now that I am caught and that the state is in the process of charging me for my crime. Unbeknownst to me, however, the state is going to charge me for the car that Angela stole, rather than my own car. Suppose that, in an affidavit, a clerk must enter the VIN number of the car that was stolen. By mistake, he fails to enter the VIN number for Angela's car, and accidently enters the VIN number for the car that I stole. I am charged for stealing the correct car.

This case is interesting because it poses a problem for both rights forfeiture and for standard retributivism. The main feature here is that the state gets lucky. That is, I am (mistakenly) charged for taking the correct vehicle.

[9] It is noteworthy that Pardo does not explicitly endorse safety as a necessary condition on knowledge and, thus, legal proof. He merely states what *whatever* condition avoids Gettier cases is required for legal evidence too.

This, I think, is problematic. After all, it is an easy possibility that I am tried for the wrong crime, and this undermines the relatedness condition I argued for above. The state ought to be certain about the connection between the crime and the charge. Luck, that is, exposes that mere rights forfeiture (and the fact that I have done the crime) are insufficient to demonstrate permissible punishment. We need a modally *stable* approach to conviction.

Let me substantiate the following intuition about this case: it would be a problem if I were charged with the wrong car. This is because my charge would be inconsistent with the relatedness condition. Recall that the relatedness condition says that I must be charged for the crime that I actually committed. Thus if I were charged with stealing the wrong car, the retributivist and the rights forfeiture theorist—so long as they accept the relatedness condition (which they should)—must find the accusation problematic.

We may ask: "Why must the charge include the correct VIN numbers? Isn't the mere fact that I stole *a Lexus* sufficient for an accusation?" Recall the fine-grained claim. According to this claim, the accusation of the guilty must be, to some degree, fine-grained. We added two qualifications to the fine-grained claim: the relevance qualification and the discoverable qualification. I think both of these qualifications are met. The VIN number is relevant because it will help the state meet the relatedness condition. Recall that the relatedness condition says that the state, in order to correctly respond to me, must punish me for the reasons that make me a wrongdoer. Thus the state needs to identify the VIN number in order to substantiate facts relevant to the accusation. The state can also figure out the correct VIN number, which satisfies the discoverable condition. Thus the accusation needs to be fine-grained enough to get the VIN numbers correct.

Now, the state got lucky, yet it meets the conditions of the theories outlined above. I've tried to show that getting lucky here is a problem because it violates the relatedness condition and the fine-grained claim that I outlined above. These theories, however, seem to suggest that it is permissible to punish the guilty in this way. That is, their criteria for what makes a punishment permissible problematically rules my punishment in as a permissible punishment. The rights forfeiture theory and standard retributivism are therefore too weak—or, at least, so I argue in the next two sections. I now examine rights forfeiture and this case.

5. RIGHTS FORFEITURE AND *LEXUS*

Let's turn now to the rights forfeiture theory. According to the rights forfeiture theory of punishment, I forfeited my right against some degree of hard treatment, in virtue of stealing the Lexus. I'll argue that this theory of punishment labels *Lexus* as an instance of permissible punishment. I'll then suggest that this is not right, by entertaining an objection from Wellman.

The rights forfeiture theorist must accept that *Lexus* is an example of permissible punishment. The problem is that it does not clearly matter that the state gets it right because the cars are of equal value and stealing them amounts to roughly equal punishment, which, according to this account, would be permissible. Thus, once I forfeit a right against hard treatment, it seems that so long as the state charges me for some crime proportional to the crime I committed, the state permissibly punishes. It looks like the theory, if what I said about the case above is correct, problematically labels *Lexus* as a case of permissible punishment. If this is right, then the rights forfeiture theory is too weak, i.e. merely forfeiting a right against hard treatment is insufficient to demonstrate permissible punishment.

Let me now field an objection. Wellman (2012) responds to the problem of relatedness. The problem of relatedness is that the rights forfeiture theory of punishment does not establish any necessary connection between the punishment and the crime that the guilty person actually did. Wellman fields this objection succinctly: "If a criminal forfeits her rights, then she may be punished for any reason" (380). Warren Quinn (1985) argues in favor of this objection. Quinn sets up a case similar to my own; call it *Boat*: suppose I am charged with stealing a $60,000 boat. I have not done this. However, I have, only recently, stolen a $60,000 car. Assuming that the punishments here are equal and that I have forfeited a right against hard treatment, the state seems permitted to punish me for a crime I did not commit.

Wellman, following Stephen Kershnar (2002), argues that rights forfeiture proponents have two avenues of response: a limited-reasons account and an unlimited-reasons account. According to the limited-reasons account, Kershnar writes that "if a person infringes or threatens to infringe the moral rights of another person, then she forfeits a moral right with regard to, and only to certain reasons for action" (77). If I have assaulted Angela, then I must be punished for this reason—it is problematic to punish me for any other reason. According to the unlimited reasons account, Kershnar writes that "if a person infringes or threatens to infringe the moral rights of another person, then she forfeits a moral right with regard to any reason for action" (77). The unlimited-reasons account contends that it would be permissible to be charged with stealing the boat,

instead of the car I actually stole.

In *Boat*, the limited reasons account gives us a sufficient response. The state must correctly connect the punishment to the reason that I have forfeited my right not to be punished, i.e. stealing a $60,000 car. Thus if we accept the limited reasons account, this account can handle *Boat*, but not *Lexus*, to which I now turn.

Consider why Wellman's response to the problem of relatedness is not sufficient to handle *Lexus*. Establishing a limited reasons account gives us a good reason to reject *Boat*. However, the proponent of the limited-reasons account is not so fortunate with *Lexus*. This is because in *Lexus* we have the correct connection between the punishment and the reason for which I am punished. The problem is the way in which the state has established this connection: it is by sheer luck. Thus the limited-reasons account does not fully get the rights forfeiture theorist out of the woods with cases of the problem of relatedness. We need, I suggest, an anti-luck condition, either in our limited-reasons account or for rights forfeiture.

Note the differences between *Lexus* and *Boat*. In *Lexus*, the state has, by sheer luck, established the correct connection between the punishment and the crime. So, by limited reasons accounts of rights forfeiture, the state has acted on the limited set of reasons. The state gets it right, but it seems clear that there is still a problem: punishing correctly by luck is not sufficient for permissible punishment.

I have argued that the forfeiture theorist should accept a relatedness condition for permissible punishment. If she accepts this, then she can also respond to *Boat*, because the reasons for my punishment and the culpable crime which I have actually committed are not the same. Adding a relatedness condition to permissible punishment, however, does not get the forfeiture theorist out of the woods: *Lexus* still poses a problem. I turn now to another solution the forfeiture theorist may identify.

Consider the unlimited reasons account (URA). According to this view, I forfeit rights proportional to whatever crime I committed. If I am charged with stealing a $5,000 four-wheeler, when I actually stole a $5,000 dirt bike, this is not a big deal according to the URA theorist. But note that these crimes are proportional. The URA theorist, I think, would reject an instance where, instead of the minor speeding charge I actually committed, I am charged with murder. Disproportional punishments are still problematic.

Now, in *Lexus* the URA theorist would not care if I were charged with stealing Angela's car, since the punishment is proportional to the crime I actually committed. According to Wellman: "We could still criticize the person who punishes for the wrong reasons, but this does not mean that

she violated my rights, and it does not mean that she acted impermissibly" (2012: 383). Thus the URA theorist would not condemn—at least not for the reason that the punishment was impermissible—the state for getting lucky.

It is clear that the URA theorist rejects the relatedness condition. This means that once I have culpably committed a crime, the state can punish me for any reason: the state needs no accurate connection between my actual culpable action and the reasons for punishment. But recall my reasons for thinking that the standard retributivist and forfeiture theorist must accept a relatedness condition. It ensures that the state correctly identifies the reasons that make me a wrongdoer, and it also condemns systemized framing, when, say, the prosecution believes that I did some crime it could not sufficiently substantiate. Such practices are clearly unjust. The URA account implies that such practices are permissible. This, I think, is sufficient to reject this account. I turn now to showing that *Lexus* undermines standard retributivism.

6. STANDARD RETRIBUTIVISM AND *LEXUS*

Lexus also poses a problem for standard retributivism. According to this theory of punishment, the crime itself provides sufficient reason for punishment. Retributivists are committed to the relatedness condition, as I tried to show in the first section. Yet the state has accidentally punished me for the correct reason in *Lexus*. The above case, I argue, demonstrates that this account of punishment is too weak—we have met the conditions for punishment, yet the punishment does not appear to be permissible.

Could the retributivist argue that *Lexus* gives us an instance impermissible punishment? I suppose, but it is unclear why this would be true. The state has correctly identified the reasons for my punishment, e.g. the exact VIN number, my being at the car dealership at the time the car was stolen, etc. Given this fine-grained and accurate accusation, it is difficult to see how retributivism, as I have construed it, could show that it is an impermissible punishment. I generated a reason for punishment in virtue of stealing the car. The state has identified this reason and punished me accordingly. Such a picture lies at the very heart of retributivism. It is therefore plausible to think that the retributivist must say that *Lexus* gets us a permissible instance of punishment.

I want to suggest a claim about retributivism: When there is insufficient relatedness between the state's reasons for punishment and my actual, culpable crime and the reasons that make me a wrongdoer, the intrinsic

goodness of the punishment is undermined. Recall that (desert-based) retributivists are committed to the claim that it is good that the guilty get their just deserts. I think that this intrinsic goodness is contingent upon the punishment adhering to the relatedness condition. Though there is some goodness in the guilty being punished, punishing her for an incorrect reason (or incorrect set of reasons) undermines this goodness because it fails to identify what makes her a wrongdoer, as I contended in the grounds argument.

Let me now take this claim a step further. As *Lexus* points out, the state gets lucky in its accusation. Thus the state could have easily gotten its accusation wrong. I think this feature of the case also undermines the intrinsic goodness of my punishment, since it could have easily been that the relatedness condition was not met. The state must be certain about its accusation—this is one reason why standards of proof, e.g. beyond a reasonable doubt, are so important and controversial in the philosophy of law. Thus my punishment for stealing the Lexus—which is conferred upon me by luck—is less good than if the state had been certain about its accusation.

Consider an objection. Patrick Tomlin (2014) argues for the principle of comparative proportionality. According to this thesis, if a state deems crime A worse than crime B, then the state must punish A more harshly than crime B. The case I have given meets this principle. Regardless of which car I am charged with stealing, I am punished proportionally. This proportionality, an objection may go, is sufficient for a permissible punishment.

Two observations offer a response: first, this objection misses the relatedness condition that I have argued for above. According to the relatedness condition, a punishment is permissible only if the state has correctly identified me as having done the crime I actually committed. I cannot be charged for a different crime that I have actually done. Second, the principle of comparative proportionality, on a reasonable interpretation, is a necessary condition on permissible punishment, not a sufficient condition. If I am charged with a crime that I did not commit, this may fail to meet other necessary conditions on permissible punishment—for instance, the relatedness condition. Thus, even though the punishment in *Lexus* may meet the principle of comparative punishment, it is not a permissible punishment.

If these thoughts are plausible, then we need a mechanism to account for them. My suggestion is that the safety condition from epistemology provides such a mechanism. I turn now to defining the safety condition and working it into standard retributivism and rights forfeiture.

7. SAFETY AND PUNISHMENT

The safety condition goes some way to dispel luck-related worries in epistemology. For instance, I do not have knowledge in Fake Barn Country if there is a safety condition on knowledge. Though some have argued for "unsafe" knowledge, these responses have ultimately not been successful.[10] I argue in this section that adding a safety condition to retributivism and rights forfeiture theories of punishment relieves the worries outlined in the *Lexus* case. But first, what is the safety condition?

Ernest Sosa offers the following first stab at the safety condition on knowledge:

> Call a belief by S that p "safe" iff: S would not believe that p without it being so that p. (Alternatively, a belief by S that p is "safe" iff: as a matter of fact, though perhaps not as a matter of strict necessity, S would not believe that p without it being so that p.) (Sosa 1999: 378)

He writes that "in order to… constitute knowledge, a belief must be safe" (Sosa 1999: 378). Note that this is a necessary and sufficient condition for knowledge. A belief counts as knowledge if and only if it is safe. Knowledge on this view is safe belief.

Duncan Pritchard and John Hawthorne also offer accounts of the safety condition on knowledge:

> If a believer knows that p, then in nearly all, if not all, nearby possible worlds in which the believer forms the belief that p in the same way she does in the actual world, that belief is true. (Pritchard 2005: 163)

On this view, we examine possible scenarios where S believes p in the same way that S actually believes p and examine whether or not p is true in those scenarios. For example, I do not know that my lottery ticket is a loser because there is a world, equidistant to the actual world where my ticket is a winner, even if the chances are low. So the belief is not safe, and I do not know my ticket is a loser. Hawthorne also offer an account of safety:

> Insofar as we withhold knowledge in Gettier cases, it seems likely that 'ease of mistake' reasoning is at work, since there is a very natural

[10] See, for instance, Juan Comesaña (2005). The famous Halloween case purportedly gives us an instance where S knows that P and S could have easily been wrong (or there are nearby possible worlds where S believes falsely). The problem with the case is that Comesaña assumes that being unsafe at one moment makes you unsafe at the next moment. This is not right. Thus, in the case, when Juan appears Juan-looking, rather than Michael-looking, his belief that the party is down the right road is safe, even though, just moments before, his belief would have been unsafe. Once Juan appears looking like Juan, Judith instantly becomes a reliable and safe source of testimony. Others have tried, unsuccessfully, to alleviate this temporal issue. *See Atomic Clock* (Bogardus 2014). For responses to Bogardus, see Coffman (2015).

> sense, in such cases, in which the true believer forms a belief in a way that could very easily have delivered error. (Hawthorne 2004: 56 n. 17)

Hawthorne give us a nonmodal account of safety. The idea here is that if a belief could have easily been false, then we fail to meet the safety condition. Though I prefer safety construed in terms of easy possibilities, I will stay neutral about the account of safety: both will offer helpful resources for the literature on punishment.

I propose the following necessary condition on rights forfeiture and retributivism:

> *Safe Punishment*: S's punishment of p for crime C with evidence E is permissible only if S safely accuses p of C with E.

But what is a safe accusation? A safe accusation is an accusation where there is a safe connection between the crime and the accusation. In all nearby possible worlds where the state accuses p of crime C with evidence E, the state correctly connects the crime and the accusation. Alternatively, we could say: the state *could not have easily* mistaken the connection between the crime and the accusation. This, I think, is distinct from Pardo's (2010, 2011) requirement, as he thinks that an anti-luck condition is necessary for the standard of evidence. In contrast, I want to consider safety as a requirement of permissible punishment. Unlike Pardo, however, I think that the anti-luck condition bears on the circumstances—namely the act of accusation—for punishment, and not necessarily the standard of evidence. Specifically, I think that permissible punishment implies that punishers, in all nearby possible worlds, correctly connect the crime to the criminal. This is distinct from the epistemic sense of safety, where in nearby possible worlds we believe truly.

What does this mean for *Lexus*? *Lexus* is an example of an unsafe punishment. There are nearby possible worlds where I am accused of stealing the wrong car. The state, moreover, could very easily have gotten its punishment incorrect. These close possibilities make it the case that the punish is impermissible. That is, in order for the state to punish permissibly, it must get it right in scenarios that are sufficiently similar to the actual world. Just as the safety condition says that we must believe truly in all nearby possible worlds, the state must get it right in nearby possible worlds where it would have punished with its particular evidence.

How do we know whether a possible world is nearby or far away? This has been understood in terms of the number of propositions you must change between different states of affairs. For instance, a possible world that is identical to ours—except where I have dyed my hair purple—is pretty close. Everything between the actual world and this possible world

is identical, with the exception of my hair color.

Now, could it have very easily been the case that I have purple hair? Given my current state, the answer is no. I would be quite reluctant to dye my hair purple. Understood in terms of easy possibility, it seems to be a pretty remote possibility that I would dye my hair. Thus the two accounts of safety will get us slightly different results because different possible scenarios will be relevant. I think, overall, that the "easy possibility" construal of safety is better because it will rule out far-fetched scenarios. Yet, I remain neutral on this point; retributivists and forfeiture theorists can debate this point.

My point is that safety is a necessary condition on permissible punishment, for both standard retributivism and forfeiture theory. This condition can be construed in at least two ways. Standard retributivism can be construed as follows: state S permissibly punishes only if S's accusation could not have easily been false (or in all nearby possible worlds, S's accusation is true). It is important to note that the safety condition could also connect to the relatedness condition. That is, we could modify the relatedness condition to capture anti-luck concerns: state S permissibly punishes only if S's accusation is *safely* related to the wrongdoer's reasons for being a wrongdoer. Thus we have two different formulations: safety as a necessary condition on permissible punishment, and safe relatedness as a necessary condition. Either, I think, is sufficient for an anti-luck condition.

8. OBJECTIONS

Let me now consider several objections. I first address the demandingness objection. I then respond to a related, yet distinct objection that most punishment is actually unsafe. I also address an objection from nonideal theory.

Adding a safety condition to punishment may be too demanding. This condition on permissible punishment seems to make the standards of permissible punishment much higher than we otherwise would have thought. In *all* possible worlds where the state punishes, the state must correctly connect the crime to the accusation. This is a matter of necessity, i.e. there is no possible state of affairs where the state gets the accusation wrong. Thus state S would not punish, unless it punishes for the right reasons. This is much too demanding; most punishment probably does not meet these standards.

In response, we need to turn back to Sosa's construal of safety. He

writes, "a belief by S that p is 'safe' iff: as a matter of fact, *though perhaps not as a matter of strict necessity*, S would not believe that p without it being so that p" (1999: 378). Sosa is not arguing that in all possible worlds where S believes, S believes truly. This, as the objector points out, would be too strong. Sosa's account of safety is weaker than this: it is merely a "matter of fact" that S believes truly in possible worlds sufficiently similar to the actual world. Thus we may think that, as a matter of fact and not necessity, the state will correctly connect the accusation to the relevant details of the crime committed. Thus I think this helps relieve the demandingness objection.

Also consider Pritchard's account of safety again: "If a believer knows that p, then in nearly all, if not all, nearby possible worlds *in which the believer forms the belief that p in the same way she does in the actual world*, that belief is true" (2005: 163, added emphasis). Pritchard only considers possible worlds that are sufficiently similar, i.e. in the belief formation process, to the actual world. Consider two scenarios to highlight Pritchard's thought: First, I watch the Discovery Channel and learn that some lizards reproduce asexually. Second, I am high on LSD and fortuitously form the belief that some lizards reproduce asexually. Even though I form the same true belief in both cases, I have formed the belief in very different ways. Pritchard says that the only possible worlds that are relevant are the ones in which I have a similar belief formation process. A similar claim can be made for punishment: We are only examining possible worlds where the state has a similar body of evidence and raises an accusation based on that evidence. This is why I included "for evidence E" in my account of safe punishment. Thus we are confined to a subset of the possible worlds where the state accuses, and not all of those possible worlds. If we confine the subset of possible worlds as Pritchard suggests, the safety condition is much less demanding.

Consider another objection: If it's so easy to get the VIN numbers wrong, the clerk's writing down the number seems to be unsafe a lot of the time. Suppose that I have been identified as stealing the correct car. The clerk writes down the correct VIN numbers, and the state has the appropriate connection between the accusation and the crime I committed. Could it not have easily been the case that the clerk wrote down the numbers wrong? Thus, even when the state gets it right, the accusation is still unsafe. If it is unsafe, then this is an instance of unsafe (and impermissible) punishment. But this is too strong, an objector may argue, because there are some instances of safe punishment. Thus the safety condition on permissible punishment is too strong.

In response, so long as the appropriate measures are taken at the front

end—the clerk double checks her work, the investigators are confident about their findings, etc.—the punishment is safe. That is, when precautions are met, the state gets it right by their own capacities and not by luck.

Consider another objection: Following Alice Ristroph (2020), the safety condition runs into ideal theory about the nature of punishment. Ristroph argues that "substantive criminal law" fails to acknowledge that the law is a human construct, subject to human errors. Consequently, it primes young attorneys to be needlessly retributive in their future prosecutions. The safety condition fails to consider various unjust circumstances in the real world and especially the U.S. legal system, such as mass incarceration and racially disparate punishments. We should, therefore, reject the safety condition, since it runs the risk of contributing to these nonideal problems.

In response, the theories we are working with—rights forfeiture and retributivism—are themselves ideal theories of punishment. This discourse—about the nature of permissible punishment—is inherently idealizing. Thus there is nothing about the safety condition itself that is ideal; I merely use it in a discussion that is already centered on ideal theory. If this objection stands, it may be evidence for rejecting rights forfeiture and retributivism outright. A substantive defense of ideal theory lies beyond the scope of this paper: the lesson from this objection is that we should give up more than the safety condition if the objection stands.

Consider one last objection: My project here confuses the following distinction: (1) the conditions under which punishment is objectively permissible or permissible as a matter of fact, and (2) the conditions under which we would be justified in believing that punishment is so permissible. The concern with the anti-luck condition, the objection goes, is primarily concerned with (2), whereas my argument here has been concerned with (1). Safety is not required for permissible punishment as I suggest; rather it is required for, say, standards of proof or justified belief about punishment.[11]

In response, the contribution of this paper is that safety is required for (1). In all nearby possible worlds, the state must correctly connect the crime to the criminal punished. But insofar as meeting a certain standard of proof is required for permissible punishment, I will at least say that safety is a requirement for meeting a certain burden of proof. Thus even if safety is just a requirement for (2), (2) is going to play some role in (1).

9. CONCLUSION

[11] Thanks to the reviewer for pointing out this objection.

This paper reveals that the rights forfeiture theory of punishment and standard retributivism are both in need of an anti-luck condition. *Lexus* demonstrated that these theories alone are insufficient for permissible punishment. I proposed—following the view of many epistemologists—that we need a modal, anti-luck condition on permissible punishment. I have proposed safety as such a condition. Thus, in order for an instance of permissible punishment to occur, in all nearby possible worlds where the state punishes in a similar way, the state must correctly connect the guilty person's crime to the accusation. Or it could not have easily been the case that the state punishes wrongly. Adding such a condition to permissible punishment make it safe against anti-luck worries, such as *Lexus*.

BIBLIOGRAPHY

Bogardus, T., 2014: "Knowledge Under Threat", *Philosophy and Phenomenological Research* 88: 289-313.

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

Coffman, E. J., 2015: *Luck: Its Nature and Significance for Human Knowledge and Agency*, New York: Palgrave Macmillan.

Comesaña, J., 2005: "Unsafe Knowledge", *Synthese* 146: 395-404.

Dagger, R., 1993: "Playing Fair With Punishment", *Ethics* 103: 473-88.

Dolinko, D., 1991: "Some Thoughts About Retributivism", *Ethics* 101: 537-59.

Gardiner, G., 2022: "Legal Evidence and Knowledge", in *The Routledge Handbook of the Philosophy of Evidence*, ed. C. Littlejohn and M. Lasonen-Aarnio, Abingdon: Routledge.

Harel, A., and Porat., A., 2009: "Aggregating Probabilities Across Cases: Criminal Responsibility for Unspecified Offenses", *Minnesota Law Review* 94: 261-310.

Hawthorne, J., 2004: *Knowledge and Lotteries,* Oxford: Oxford University Press.

Kershnar, S., 2002: "The Structure of Rights Forfeiture in the Context of Culpable Wronging", *Philosophia* 29: 57-88.

Levmore, S., 2001: "Conjunction and Aggregation", *Michigan Law Review* 99: 723-56.

McBride, M., 2011: "Reply to Pardo: Unsafe Legal Knowledge?", *Legal Theory* 17: 67-73.

Menashe, D., 2014: "On the Inadmissibility of the Aggregated Probabilities Principle", *The International Journal of Evidence and Proof* 18: 291-309.

Moore, M., 1993: "Justifying Retributivism", *Israel Law Review* 27: 15-49.

Moss, S., 2022: "Knowledge and Legal Proof", in *Oxford Studies in Epistemology*, vol. 7, ed. T. S. Gendler, J. Hawthorne, and J. Chung, Oxford: Oxford University Press.

Nau, R., 2001: "The Aggregation of Imprecise Probabilities", *Journal of Statistical Planning and Inference* 105: 265-82.

Nozick, R., 1981: *Philosophical Explanations*, Cambridge, MA: Harvard University Press.

Pardo, M., 2005: "The Field of Evidence and the Field of Knowledge", *Law and Philosophy* 24: 321-92.

—, 2010: "The Gettier Problem and Legal Proof", *Legal Theory* 16: 37-57.

—, 2011: "More on the Gettier Problem and Legal Proof: Unsafe Nonknowledge Does Not Mean that Knowledge Must Be Safe", *Legal Theory* 17: 75-80.
Posner, E., and Porat., A., 2012: "Aggregation and Law", *Yale Law Journal* 122: 2-51.
Prichard, D., 2005: *Epistemic Luck*, Oxford: Oxford University Press.
Quinn, W., 1985: "The Right to Threaten and the Right to Punish", *Philosophy & Public Affairs* 14: 327-73.
Ristroph, A., 2020: "The Curriculum of the Carceral State", *Columbia Law Review* 120: 1631-708.
Schauer, F., and Zeckhauser, R., 1996: "On the Degree of Confidence for Adverse Decisions", *Legal Studies* 27: 41-51.
Sosa, E., 1999: "How Must Knowledge Be Morally Related to What is Known?", *Philosophical Topics* 26: 373-84.
Tadros, V., 2011: *The Ends of Harm: The Moral Foundations of Criminal Law*, Oxford: Oxford University Press.
Tomlin, P., 2014: "Retributivists! The Harm Principle Is Not for You!", *Ethics* 124: 272-98.
Wellman, C., 2012: "The Rights Forfeiture Theory of Punishment", *Ethics* 122: 371-93.