Routledge
Taylor & Francis Group

Inquiry

An Interdisciplinary Journal of Philosophy

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/sinq20

Does predictive sentencing make sense?

Clinton Castro, Alan Rubel & Lindsey Schwartz

To cite this article: Clinton Castro, Alan Rubel & Lindsey Schwartz (02 Feb 2024): Does predictive sentencing make sense?, Inquiry, DOI: 10.1080/0020174X.2024.2309876

To link to this article: https://doi.org/10.1080/0020174X.2024.2309876

Published online: 02 Feb 2024.

INQUIRY
https://doi.org/10.1080/0020174X.2024.2309876

Routledge
Taylor & Francis Group

Does predictive sentencing make sense?

Clinton Castro[a], Alan Rubel[a] and Lindsey Schwartz[b]

[a]The Information School, University of Wisconsin-Madison, Madison, United States;
[b]Philosophy Department, Denison University, Granville, United States

ABSTRACT
This paper examines the practice of using predictive systems to lengthen the prison sentences of convicted persons when the systems forecast a higher likelihood of re-offense or re-arrest. There has been much critical discussion of technologies used for sentencing, including questions of bias and opacity. However, there hasn't been a discussion of whether this use of predictive systems makes sense in the first place. We argue that it does not by showing that there is no plausible theory of punishment that supports it.

ARTICLE HISTORY Received 20 September 2023; Accepted 26 November 2023

KEYWORDS Technology ethics; COMPAS; theories of punishment; criminal sentencing

1. Introduction

In 2013, the state of Wisconsin convicted Paul Zilly of stealing a push lawnmower and some tools.[1] His lawyer and the prosecutor agreed to a plea deal: one year in county jail with follow-up supervision. But the deal would not materialize.

In preparation for sentencing, the Department of Corrections prepared a presentencing investigation report on Zilly. The report included an algorithmically-generated forecast of the likelihood that Zilly would reoffend. The judge found the forecast to be, 'about as bad as it could be' and overturned the plea deal, sentencing Zilly to two years in state prison with three years of supervision.

CONTACT Clinton Castro clinton.g.m.castro@gmail.com The Information School, University of Wisconsin-Madison, Madison, WI, USA

[1]Details of Zilly's case obtained from Angwin et al. 2016. A well-documented and widely discussed case involving an algorithmic risk assessment tool in the context of criminal sentencing is *Wisconsin v. Loomis*, 881 N.W.2d 749 (Wisconsin Supreme Court 2016).

The Zilly case raises many questions, some of which have received ample attention. The software used to generate the risk score – COMPAS – calculates scores via statistical generalization. Was Zilly, then, robbed of his right to an individualized sentence?[2] COMPAS is owned by a private company – Northpointe, Inc. – and, thus, is under proprietary lock and key. Did Zilly suffer a wrong in being denied access to the formula that generated his score?[3] COMPAS has been accused of unfairness: it misclassifies Black defendants as high-risk at almost twice the rate it does White defendants; to what extent does this render its judgments illegitimate?[4] These are important questions. However, we are interested in a more foundational question: does it even make sense to use risk assessments in the way that they were used in Zilly's case?

In determining prison sentences, one is determining the length and severity of a legal punishment (hereafter we will generally refer to 'length' instead of 'length and severity' and to 'punishment' instead of 'legal punishment'). Punishment involves the intentional infliction of harm. That is, part of what it is to punish someone is to intentionally make them worse off in some way. For this reason, decisions of whether – and how severely – to punish must be underwritten by strong reasons that speak in favor of those decisions.

Those reasons – whatever they are – will offer guidance in determining the length of a sentence. For instance, consider the view that punishment is exclusively justified to the extent that it deters future crimes. If this is the correct theory, then we would be barred from using certain rationales – such as pure considerations of desert – as reasons for lengthening sentences.

Our contention in this paper is that there is no plausible theory of punishment that fits with the practice of *predictive sentencing*, the practice of lengthening the prison sentences of persons convicted of crimes just because they have high forecasts of re-offense or re-arrest. Before moving on, let us make an important note about this claim.

'Predictive sentencing' picks out something quite narrow. We are *not* saying that predictions of future re-arrest or re-offense have no place in the criminal justice system. Indeed, as we will later note, such predictions can play an important role in the allocation of scarce supervisory resources. We are instead specifically arguing that one cannot infer from the fact that someone is a likely re-offender that they should receive a longer sentence. While our focus is, indeed, narrow, the issue

[2]See, e.g. Freeman (2016).
[3]See, e.g. Pasquale (2015).
[4]See, e.g. Angwin et al. (2016) and Corbett-Davies and Goel (2018).

of whether predictive sentencing is justified is an important one. Software that predicts reoffense is widely available, and an intuitive application of this software is to use it as it was used in Zilly's case.

In what follows, we begin by clarifying the concept of punishment itself. We then take a piecemeal approach, working through various theories of punishment one-by-one, showing that none of them justify predictive sentencing. Aside from taking what, we hope, is an uncontroversial stance on what punishment consists in, our main argument will not take a stand on what the best, or most well-supported, theory of punishment is.

2. Punishment, what?

Before giving our arguments, it will be helpful to say a few words about what punishment is. Following Boonin (2008), we understand punishment as authorized reprobative retributive intentional harm.[5]

Let's first clarify our definition. As we stated in the introduction, we take it that it is part of the concept of punishment that it involves intentional harm, where we take harming someone to involve making them worse off in some way, and intentionally harming them to involve aiming for the harm to be *experienced as* harm (whether or not it is, in fact). We often associate punishment with incarceration, but the intended harm can be financial (fines) or the loss of some kinds of privileges (e.g. to hold certain positions of public trust). Further, we take it that the intentional harm has to be retributive in the sense that must be done in response to the commission of a legal offense. Further, the intentional harm has to be reprobative, that is, it has to express disapproval. Finally, it must be authorized, that is, it has to be carried out by legally authorized officials acting in their capacities as officials. If an action is missing one of these elements, it is best understood as something other than punishment.[6] If an action isn't harmful, it is mere censure. If the harm is not intended, the conceptual distinction between, say, parking fees and parking fines or between benevolent institutionalization and punitive incarceration collapses. If an action isn't retributive, it is either arbitrary harm or vicarious harm. If an action isn't authorized, it is vigilantism. If the action isn't reprobative, it is based on bad law.

[5]For a thorough (and in our view, decisive) defense of this definition of punishment, see Boonin (2008) section 1.1.

[6]Though the scope of this paper is limited to assessing the compatibility of predictive sentencing with theories of punishment, we acknowledge that assessment of the same with respect to alternative theories of criminal justice is a worthy future pursuit. Pereboom's (2019; 2020) quarantine model of criminal justice, for instance, is one popular alternative to punishment that might be worth assessing along these lines.

Before we proceed, one thing worth clarifying is that our *definition* of punishment neither commits us to nor forecloses any plausible *theory* of punishment, that is, any theory of what (if anything) justifies punishment. Though punishment is, on our definition, retributive, it is retributive in the *legal* sense and not in any *substantive* sense of the word. The difference is not nominal. It is the difference between claiming that lawbreaking itself justifies hard treatment and merely acknowledging that, if the law is to have any force, lawbreaking cannot go unanswered. An action is legally retributive if and only if it is done in response to a legal offense. Put another way, any action taken against a person that is *not* done in response to the commission of a legal offense does not, and cannot, count as *legal* punishment. Punishment is imposed to answer clearly delineated legal infractions for which a person has been formally determined to be responsible. It does not make sense, for instance, to punish someone for a crime that no one is sure has transpired. Even for inchoate crimes, a person has to have taken reasonable steps toward actual commission for punishment to be appropriate. This is an essential feature of legal punishment insofar as it differentiates the concept from condemnatory harms imposed by authorized entities for other reasons like the breach of a social norm, or a non-legal rule or regulation. Importantly, nothing about this sense of the word, retributive, (or this conceptual component of our working definition of punishment) precludes a consequentialist *theory* of punishment. One might well argue that we should hold those who commit crimes responsible for them by subjecting them to some form of harm because doing so will promote the good, via deterring crime. Holding them to account *for their crimes* by subjecting them to authorized reprobative intentional harm is legally retributive; subjecting them to that treatment in the service of deterrence-based aims is substantively consequentialist.

Similarly, our identification of punishment as reprobative does not mean that we think that punishment must be justified on expressivist grounds. Following the previous example, we might express disapproval *as a means* to deter crime. It is an essential part of the definition of punishment insofar as it differentiates punishment from authorized harms, enacted in response to lawbreaking, that are not intended to be condemnatory. Take, for example, a civil judgment of monies owed by a defendant to a plaintiff. The payment may leave the defendant monetarily worse off, and may do so intentionally, but it is merely a judgment of restitution owed, free from any condemnation for owing. So, our definition of punishment as a practice does not commit us to any particular theory

of punishment. With this clarification in place, we turn to the case against predictive sentencing.

3. Predictive sentencing and consequentialist theories

Consequentialist theories of punishment hold that legal punishment is justified because of its beneficial consequences. Consequentialist theories differ on their accounts of what the relevant valuable outcome is (e.g. deterrence) and exactly how to evaluate punishment (e.g. by evaluating *each* punishment or the *practice* of punishing).

Despite its many varieties, we have – at root – one argument for the conclusion that predictive sentencing does not make sense on any plausible consequentialist theory of punishment

(1) If predictive sentencing makes sense on some plausible consequentialist theory of punishment, then imposing longer prison sentences on persons in response to the crimes for which they have been convicted on the further basis that they have higher forecasts of recidivation leads to better consequences.
(2) It's not the case that imposing longer prison sentences on persons in response to the crimes for which they have been convicted on the further basis that they have higher forecasts of recidivation leads to better consequences.
(3) So, predictive sentencing does not make sense on some plausible consequentialist theory of punishment.

The first premise follows from the definition of predictive sentencing. Allow us, then, to develop the key premise: premise 2.

In a recent study of the evidence on the impact of incarceration on crime, David Roodman (2017) concludes that

> the best estimate of the impact of additional incarceration on crime in the United States today is zero. And, while that estimate is not certain, there is as much reason overall to believe that incarceration increases crime as decreases it. (7; emphasis removed)

Similarly, Daniel Nagin – the author of a separate large-scale overview of the empirical literature on deterrence – finds, 'there is little evidence that increases in the length of already long prison sentences yield general deterrent effects that are sufficiently large to justify their social and economic costs' (Nagin, 3).

To vet premise 2., we will present key details of the empirical research on deterrence, following Roodman's scheme of understanding the importance of *deterrence*, *incapacitation*, and the *aftereffects* of imprisonment. Note that under Roodman's rubric, the discussion of deterrence will primarily (but not exclusively) address what many would call *general* deterrence (deterring the public at large from committing crimes). Discussion of incapacitation and aftereffects will primarily (but not exclusively) address *specific* deterrence (deterring specific offenders from committing crimes).

Begin with *deterrence*. There is very little evidence that marginal increases of sentences have significant general deterrence effects (Roodman 2017; Nagin 2013). Surprisingly, there are very few quality studies on this topic. Of the few that support the conclusion that longer sentences deter, the deterrence effect of longer sentences is found to be small and not worth the cost (see, e.g. Helland and Tabarrok 2007).

One of these studies – Helland and Tabarrok (2007) – analyzes California's three-strikes law and finds that it does in fact deter crime: the program deters about 31,000 crimes per year. But this comes at a high cost of about $4.6 billion, by way of increasing the sentences of about 8,000 prisoners by about 16.6 years per prisoner (which is, of course, a source of tremendous disutility for those prisoners, their families, their children, and so on). Klick and Tabarrok (2010) estimate that for roughly the same price, around 1,000,000 crimes could be prevented if that $4.6 billion was transferred to new police hires. They further argue that those new police would provide greater overall benefit than harm.

Now, one could respond to our use of this study by pointing out that the three-strikes law is a blunt instrument, nothing like COMPAS or similar sophisticated products. This very well might be true, but that doesn't undermine our use of the study. It's plausible that most defendants facing a third strike would be deemed high risk by a program like COMPAS. Further, no quality evidence suggests that simply giving longer sentences to higher risk defendants has general deterrence effects that could justify them. As we will soon show, this stands to reason: Individuals prone to committing crimes tend to have a psychological make-up that makes longer sentences the wrong kind of thing to deter them from committing crimes. So, we are happy to admit that the Helland and Tabarrok (2007) study is not decisive evidence for our case. But it does provide at least *some* evidence. Further, it is not our only evidence.

The fact that increased sentences have little effect on crime is at least partially explained by the fact that many of those who are prone to committing crime are also prone to future discounting (Mastrobuoni and Rivers 2016). That is, individuals who commit crimes also tend to see punishments (and rewards) in the distant future as less severe – all things being equal – than those in the present or near future. As Latessa, Listwan, and Koetzle (2015) puts it, 'The problem is that most street-level criminals act impulsively; have a short-term perspective; are often disorganized […] and are not rational actors' (Latessa, Listwan, and Koetzle 2015, 47). This, in effect, means that many of the people who are supposed to be deterred by longer sentences tend to be less sensitive to that form of deterrence. This helps to explain why – dollar for dollar – Klick and Tabarrok's (2010) suggestion of hiring more police as opposed to increasing sentences is a more effective deterrent. Police presence is one way of increasing the *certainty* of punishment, which – it turns out – *is* an effective deterrent (Nagin 2013). Note that this isn't to say that we favor increased policing as an alternative to punishment as opposed to, say, investment in social services. It is simply an example of a noncustodial deterrent that is more potent, dollar-for-dollar, than increased sentences, i.e. one that the consequentialist should prefer.

Turn now to *incapacitation*. While longer prison sentences do not appear to be an effective deterrent, they do serve to incapacitate the incarcerated for a longer period of time. Incarceration *is* an effective means to preventing prisoners from committing crimes *outside* of prison *while they are incarcerated* (Roodman 2017). This stands to reason: it is hard to commit crimes outside of prison when you are in it.

This may seem like a boon to the (specific) deterrence case for predictive sentencing, but the positive criminogenic effects of incapacitation are offset by incarceration's *aftereffects*, that is, the effects of imprisonment on individuals. This too stands to reason. As the National Institute of Justice puts it,

> Prisons are good for punishing criminals and keeping them off the street, but prison sentences (particularly long sentences) are unlikely to deter future crime. Prisons actually may have the opposite effect: Inmates learn more effective crime strategies from each other, and time spent in prison may desensitize many to the threat of future imprisonment. (National Institute of Justice 2016)

Moreover, longer sentences cannot prevent inmates from committing crimes *inside* of their prison facilities, and, to reiterate the evidence

against the deterrent effects of longer sentences, being subject to a longer sentence weakens the reasons they may otherwise have to refrain from participating in criminal activity while incarcerated. It is for these reasons that incapacitation is not an effective specific deterrent from a consequentialist point of view.

A further reason that imposing longer sentences on persons in response to the crimes for which they have been convicted on the further basis that they have higher forecasts of recidivation does not lead to better consequences stems from, and builds on, a familiar objection to consequentialist theories of punishment in general. Predictive sentencing results in disparate treatments meted out in response to identical crimes. It thus detracts from the uniformity of legal sanctions. Moreover, it does this on extra-legal grounds.

The enhancement to the sentence in the case of predictive sentencing as we have defined it – the extra time tacked on to the end of the 'regular' sentence – is imposed on the basis of the forecast for recidivation, which means it is imposed in response not to the crimes for which a person has been convicted, but for those which he is more likely to commit in the future. The extent to which a consequentialist theory of punishment is a theory of (legal) *punishment* is exactly the extent to which it justifies imposing a burden on a criminal offender *for* or *in response to* his having committed a criminal infraction. That is what makes it a theory of *punishment*: the application of a general normative theory to the restricted domain of punishment. Extended sentences may make sense on some other application of consequentialism. A consequentialist theory of crime control, for instance, can certainly countenance enhanced sentencing on the basis of behavioral forecasts. Extending a sentence, as we mentioned above, extends the period for which an offender is incapacitated. If incapacitating a person for a greater period of time is likely to prevent bad behavior going forward, that would be a legitimate method of crime control from the consequentialist perspective. But what may be justified as a method of crime control differs significantly from what may be justified as a legitimate aim of legal punishment, i.e., punishment *in response to the commission of a crime*.

A recidivation forecast is about the likelihood of future criminal activity. It essentially specifies the odds of a person committing another crime at some time in the future, once the criminal sentence for *this* crime has expired. That one is likely to reoffend does not guarantee that he will in fact reoffend, and there's a sense in which the presumption as it stands is (and ought to be) that the odds are his to beat. If they

weren't, there would be no point in offering an expiration date for the sentence in the first place.

All of this might lead us to ask, then, whether risk scores are of *any* use to the consequentialist; that is, it might lead us to ask what, if anything, is to be done with high-risk individuals in place of longer sentences. As it turns out, risk scores are quite useful for determining who should get which services. High risk individuals have been shown to benefit from cognitive behavioral theory, for example, with high-risk graduates of these programs having 50% lower recidivism rates than their counterparts who did not participate (Lowenkamp et al. 2009). Since low-risk individuals, by definition, are not likely to recidivate – and, as it happens, might actually be more likely to recidivate if they receive anti-recidivism treatment (Lowenkamp and Latessa 2002) – risk scores can play a crucial role in the effective allocation of socially beneficial services.

For these reasons, we think that 2. is true and, thus, that predictive sentencing does not make sense on the assumption of consequentialism. Note that we *do not* take this argument to show that consequentialists cannot endorse a close relative to predictive sentencing, namely – as the previous paragraph indicates – there may be room to endorse the use of risk scores to inform the *quality* of a sentence; e.g., whether it involves cognitive behavioral therapy.

Let us turn, then, to another theory of punishment: retributivism.

4. Predictive sentencing and retributivist theories

The consequentialist case for predictive sentencing ultimately failed because there is little empirical reason to think that longer prison sentences are an effective way to promote good consequences in terms of deterrence or overall welfare. So, the empirical facts rule out longer prison sentences as a means of achieving positive outcomes in the future. However, there may be other reasons to opt for longer sentences. One such path would turn on retributivism being correct. Where the consequentialist looks forward to justify punishment in terms of future consequences, the retributivist looks back to past wrongs to justify punishment. Different retributivists look backward to justify punishment in different ways: some claim that punishment gives criminals what they deserve, others that it is justifiable on the grounds that people forfeit certain rights by committing crimes, and others that punishment corrects for unfair advantages people gain through criminal activity (Boonin 2008). The question for us is whether lengthening the sentences of persons

convicted of crimes when they have higher forecasts of recidivation makes sense as a response to what those persons have done.

Our view is that imposing longer prison sentences on 'high-risk' criminals does not make sense as a response to what they have done. Our argument against retributivist predictive sentencing runs as follows:

(4) If predictive sentencing makes sense on the assumption of retributivism, then, on some form of retributivism, imposing longer prison sentences on persons in response to the crimes for which they have been convicted on the further basis that they have higher forecasts of recidivation is a fitting response to the offense they committed.
(5) There is no form of retributivism according to which imposing longer prison sentences on persons in response to the crimes for which they have been convicted on the further basis that they have higher forecasts of recidivation is a fitting response to the offense they committed.
(6) So, predictive sentencing does not make sense on the assumption of retributivism.

The crucial premise in this argument is 5, and it will take us some time to develop it in full. The underlying problem, though, is easy to state. Risk assessments – when they are accurate – estimate a person's disposition to commit crimes, not the severity of an offense. Proportionality, a core doctrine of retributivist theories of punishment, links the severity of a punishment to the gravity of the offense (von Hirsch 2007). Committing a crime while being disposed to commit crimes is not the same thing as committing a more severe crime. Unless there is a plausible sense in which a disposition to commit crimes makes the commission of a *specific* crime more egregious, giving longer prison sentences to 'high-risk' individuals doesn't make retributivist sense, as risk assessments measure the magnitude of the wrong thing.

Let us begin our justification for 5 by explaining why predictive sentencing makes no sense for the most popular form of retributivism, desert-based retributivism. The desert-based retributivist holds that punishment is justified because it gives criminals what they deserve (Boonin 2008; for defenses of this view see Kershnar 2000; Moore 1987).

How exactly to determine the severity of a punishment one deserves on a desert-based view is a matter of debate. But in general the idea is that punishments should be proportional to the seriousness of the

offense. The idea is, roughly, as follows: the more serious an offense, the greater one's moral debt; these debts are paid via punishment; so, the greater the debt, the greater the punishment.

With this in mind, we can explain why predictive sentencing does not make sense from a desert-based retributivist point of view. While some crimes are made more serious by certain facts about perpetrators (their intent, and so on), a person's propensity to reoffend in the future doesn't itself make their past offense any more or less serious.

To see this, consider a case. Suppose Hanlon is caught speeding. In this case – and any other – the seriousness of the offense is a function of two factors: the effects of the activity and the offender's culpability. Suppose we are desert-based retributivists and we think that he is eligible for punishment in the form of a fine because of the danger he has exposed to those surrounding him. That is, this danger is the source of the moral debt that justifies his punishment. We learn that he is likely to speed again, even after being caught. Can the desert-based retributivist use this as a reason to increase his fine? Though we do feel some frustration with Hanlon, we do not think that this can translate into a higher penalty if we are desert-retributivists. To see this, we can work through both factors relevant to the gravity of the offense.

First ask: Does the fact that Hanlon is disposed to speed again make his previous offense any more dangerous? We think the answer is clearly no. *This* offense was as dangerous as it was, regardless of any propensity to commit some *other* dangerous offense. His propensity might be offensive in its own right, but it does not change the severity of his past offense. Further, being prone to speeding in the future is not a legal offense. Nor should it be.

Now ask: Does the fact that drivers like Hanlon are disposed to repeat offenses make him any more culpable for his action? Again, we think not. This is because we do not know *how* his propensity to speed in the future is related to his past speeding. Consider two markedly different cases. In the first case, Hanlon's propensity is caused by a brain lesion that he is unaware of. In the other, Hanlon is fully rational and speeds whenever he thinks he can get away with it. We can imagine that the propensity is the same in both cases, in the sense that it leads to the same behavior. But – knowing the nature of the propensity – we should treat Hanlon differently in these cases. It's plausible that in the brain lesion case, Hanlon's action is less offensive and that he deserves *less* punishment (if any) than in the case where his choice is free.

This raises two concerns about giving defendants with higher risk scores longer sentences. One is that just knowing that a defendant has a propensity is not enough to know the nature of that propensity; giving defendants with higher risk scores longer sentences might be deeply unjust – it might involve giving longer sentences to those who deserve shorter ones. Second, we worry that the propensities that COMPAS tracks are of this sort. As Rubel, Castro, and Pham (2020) note:

> [T]he [COMPAS] questionnaire asks about the age at which one's parents separated (if they did), whether one was raised by biological, adoptive, or foster parents, whether a parent or sibling was ever arrested, jailed, or imprisoned, whether a parent or parent-figure ever had a drug or alcohol problem, and whether one's neighborhood friends or family have been crime victims. (557)

To the extent that these are factors that defendants are not responsible for, a high risk score is, if anything, reason to think defendants are *less* culpable and, if anything, good candidates for social services, not additional punishment. Responsibility itself is a key element of retributivist justifications of punishment. But a person cannot be held legally responsible for something he has not yet done, or has not taken demonstrable steps toward doing. On the view that offenders deserve to be held responsible for their criminal actions, then, predictive sentencing does not make sense.

Perhaps it is possible to use predictive technologies like COMPAS to determine whether an offender's propensity to criminal behavior is due to factors outside of his control or whether it is indicative of a bad moral character. If an offender's high-risk forecast is due to his own bad character, there might be a sense in which he deserves to be treated worse than offenders whose high-risk forecasts are due to factors beyond their control. If, for instance, a person has demonstrated a tendency to commit crimes but scores as significantly low-risk according to other factors, it might be reasonable to extend the length of that person's sentence, since outside factors cannot be blamed for this propensity. If what justifies punishment is moral desert, then a longer sentence based on an algorithmically-generated forecast which indicates a bad moral character may be a fitting response.

The problem with this justification is that it does not justify *legal* punishment, that is, punishment in response to the offense committed. It justifies treatment on the basis of what the offender is *like* rather than what he has *done*. It is thus retributive only in the substantive sense; it is not retributive in the sense required for the act to count as legal

punishment. It is at best moral punishment, which is not something to which the criminal justice system responds. Thus, predictive sentencing does not make sense on the assumption of desert-based retributivism.

With these ideas in mind, it is fairly easy to explain why predictive sentencing does not make sense on the assumption of a different form of retributivism, forfeiture-based retributivism. The basic idea behind forfeiture-based retributivism is that when an individual breaks the law, they forfeit the rights that would make it wrong for the state to intentionally harm them (Boonin 2008; for a defense see Kershnar 2002). The exact details of this story will vary from account to account, but the basic idea remains the same: to the extent that some person violates other's rights, they lose certain of their own.

Now, for the forfeiture account to be at all plausible, it needs a principle of proportionality. That is, it needs – like desert-based retributivism – a principle that amounts to the claim that the severity of punishment must be proportional to the seriousness of an offense. The forfeiture-based account has a ready explanation of this principle, which we have already mentioned: to the extent that some person violates others' rights, they lose certain of their own. But now, forfeiture-based retributivism and predictive sentencing cannot be bedfellows. As we just explained, one's propensity to commit crimes in the future does not increase the seriousness of one's past offenses.

The last step of our explanation of premise 5 addresses fairness-based retributivism. On this view, punishment is justified to the extent that it corrects for the unfair advantage that a person gained through criminal activity (Boonin 2008; for a defense see Morris 1968 or Sher 1987). As before, we can explain why predictive sentencing does not make sense with a simple observation: the amount of advantage gained through some offense is in no way affected by the perpetrator's disposition to reoffend in the future, much in the same way that Hanlon's act of speeding on one occasion is made no more or less serious by the mere fact that he's disposed to doing it again.

Our analysis for each of these retributivist theories has a common thread. In each case, it ultimately seems to make more sense to call a legal sanction imposed on the basis of recidivation forecasts *pre*tributive than it does to call it retributive. Any sanction imposed on the probability of future offense secures retribution for exactly nothing. It may express an increase in frustration with the incorrigibility of a repeat offender over time, but its relationship to the severity of the crime committed is effectively nil.

5. Predictive sentencing and other theories

We have argued that two of the most popular theories of punishment – consequentialism and retributivism – are incompatible with predictive sentencing. We now turn to less popular theories, briefly describing each and showing why it is incompatible with predictive sentencing.

5.1. Compromise theories

So far, we have dealt with pure views, views that justify punishment by purely referencing consequences or retribution. We now turn to Hart's (1968) compromise account, on which the 'general justifying aim' of punishment is to produce good consequences but whose distributive principles – the principles determining who gets punished – are retributive.

While this theory has some benefits in virtue of combining retributivism and consequentialism, justifying predictive sentencing is not one of them. There are two key reasons for this.

One is that even though Hart keeps questions of the general justifying aim of punishment and its distributive principles more separate than other theories we have considered, this does not mean that the two are entirely separate. Hart thought that the general justifying aim of punishment was the reduction of crime. This would mean, then, that the distributive principles will be justified to the extent that they tend to reduce crime. As we showed in section 3, predictive sentencing does not seem to reduce crime. So, by the lights of the compromise theory, it isn't justified.

The other is that to the extent that the distributive principles are retributivist, they have to make sense from a retributive point of view. But – as we demonstrated in section 4 – predictive sentencing does not make sense from a retributive point of view. So, predictive sentencing and compromise theories are a poor match.

Note that these remarks also apply to other views that blend retributivism and consequentialism, such as negative retributivism (Mackie 1982) and minimalist retributivism (Golding 1975), which hold that, 'wrongdoers forfeit their right not to suffer proportional punishment, but that the positive reasons for punishment must appeal to some other goods that punishment achieves, such as deterrence or incapacitation' (Walen 2021).

5.2. Reprobative theories

Reprobative theories of punishment claim that punishment is justified because the state has an entitlement to censure offenders and that

punishment is an effective means to this end (Boonin 2008; for a defense see Duff 2001). The problem with extending the reprobative justification of punishment to predictive sentencing is that – as we noted in the section on retributivism – risk assessments don't tell us whether someone has committed a crime or how severe that crime was; rather, they tell us about something else: one's disposition to commit a crime. So, further argumentation is needed to justify the state's entitlement to use punishment to censure this disposition.

To do this, the reprobative theory would need a principle, such as

> (A) if we are justified in censuring (via punishment) those who commit offense *O*, we are justified in censuring – via punishment – one's having the disposition to commit offense *O*.

But (A) is extremely implausible. Among other things, it would justify punishing those *merely* prone to committing *O*, i.e. those who are disposed to but nevertheless will not even attempt – or even form the intention – to *O*. But this is odious; we should not punish those who are merely prone to committing a crime but do not commit it.

One might object that the reprobative justification of predictive sentencing could be made to work if we simply refine (A). But, it is difficult to see how this could be done because risk scores are a rather blunt instrument; they cannot separate the merely prone from those who will actually go on to commit crimes (and, for that matter, those who are not prone at all but merely mistaken as such). This puts the reprobativist in a difficult position: they must refine (A) to avoid our objection while leaving it weak enough to permit the use of risk scores in punishing the disposition. This seems hopeless given the bluntness of risk scores, not to mention the problematic nature of punishing people for crimes before they even plan or attempt them.

Thus, we do not think that a reprobative justification of predictive sentencing could work.

5.3. Moral education theories

According to the moral education theory of punishment, punishment is justified in virtue of its benefit to the incarcerated (Boonin 2008; for a defense see Duff 1986). That is, the bringing about the harms of being incarcerated is justified by educational benefits of incarceration.

The problem with this justification of predictive sentencing is that longer sentences do not teach the incarcerated valuable lessons. As we

stated above in our discussion of consequentialist theories, longer sentences appear – if anything – to do the opposite. In fact, moral education is compatible with the intended use of algorithmic systems like COMPAS, namely, recognizing particular characteristics of defendants in order to tailor treatment (such as CBT).

5.4. Consent theories

Finally, according to the consent theory of punishment, punishment is justified because offenders give tacit consent to being punished when they commit crimes with the understanding that in so doing they make themselves eligible for punishment (Boonin 2008; for a defense see Nino 1983).

There are a few problems with pairing this theory of punishment with predictive sentencing.

The first is that predictive sentencing is not enough of a widespread practice for offenders to know about it. Most offenders cannot expect that, were they caught, an algorithmically generated forecast might be used to inform their sentence. So, it isn't at all clear that they consent to it as a response to their criminal activity.

One might respond on behalf of the consent theory that offenders needn't consent to the particular way in which their punishment is executed. Rather, they consent to being punished. Further, using algorithms is just one way to (partially) determine the punishment.

This, however, can't be the whole story. There must be some constraints on what offenders tacitly consent to. One reasonable refinement on what the objector says is to claim that offenders needn't consent to the particular way in which their punishment is executed, just that they consent to being punished *in ways that make sense*. But this raises a further question: does predictive sentencing make sense?

In light of everything we've argued, we think it does not. We might think that the treatment makes sense if it was something offenders explicitly consented to or if the practice was widespread enough for it to be reasonably expected as a consequence of committing crime. But neither or these conditions hold. We might then think that the treatment makes sense if some other rationale could be given, linking it to something that offenders might have tacitly consented to. Stand-ins here might include being subjected, for example, to treatments that will reduce crime or give them what they deserve and so on. But, as we have argued, predictive sentencing does not meet the aims of these

rationales. So, it is hard to see how offenders could be understood as consenting to predictive sentencing.

The second, and perhaps more important problem is that – if our previous arguments are correct – even if offenders are aware of the practice, they couldn't consent to it in any way that could justify it. For offenders' consent to justify their punishment, their consent would have to meet some normative standard, such as being reasonable. Given our above arguments – that predictive sentencing does not promote the good, that isn't called for by retributivist considerations, and so on. It is hard to see how offenders could reasonably consent to predictive sentencing.

6. Predictive sentencing and punishment

To this point, we have systematically argued that predictive sentencing does not square with the principles of any plausible theory of punishment, and thus does not make sense on those views. Empirical evidence shows that it does not meet the consequentialist aims of deterrence or incapacitation. Nor does it square with retributivist aims and principles including desert, forfeiture, and the fair balance of social benefits and burdens. It is subject to combinations of these criticisms on compromise theories; it permits too much on reprobative theories; it does not provide a valuable moral lesson; and there are no good reasons for people to consent to it, tacitly or otherwise.

There is a deeper issue, though, that we have as yet only mentioned in passing. It is this: even if there were a way for predictive sentencing to sensibly square with the principles of a plausible theory of punishment, the practice still could not be justified by that (or *any* such) theory, because a plausible justificatory theory of *punishment* cannot justify a practice that itself does not meet the criteria required for an act to count as punishment.

Predictive sentencing does not itself square with the concept of legal punishment. It is not retributive in the right sense of the word. As noted in the section on desert-based retributivism, there may be a sense in which lengthening a sentence according to predictions that indicate a bad moral character is giving that person what he morally deserves, but that sense has nothing to do with law or legal retribution. If it is punishment, it is moral, not legal, punishment. The state is not in the business of punishing people for what they are like, especially since the state's shortcomings often contribute to people's character development in less than ideal ways (Murphy 1973; 2011).

For predictive sentencing to make sense on any plausible theory of legal punishment, it must count as legal punishment. For an act to count as legal punishment, it must be done in response to the (actual or reasonably established) commission of a crime. Predictive sentencing addresses not the crime that has been committed, but the possibility of future crimes. It is therefore not legal punishment, and cannot possibly be justified as a legal punishment by any plausible theory thereof. In fact, *no* practice of lengthening sentences on the basis of recidivation forecasts – whether by algorithm, by judicial reasoning, by police record, or whatever else – can be justified on any plausible theory of legal punishment. As punishment, predictively enhanced sentencing simply does not make sense.

7. Lessons and loose ends

The argument now is this: risk assessment systems such as COMPAS, as used in the Zilly case and in others, use risk of future crimes to determine punishment length. However, no plausible account of punishment links the justifiability of a sentence's severity to risk of future crimes.

We have focused on the use of a particular kind of technology – algorithmically generated forecasts of recidivism – but our arguments are about the moral justifiability of using risk at all in sentencing, regardless of whether it's a machine or a judge. Thus, in addition to putting pressure on certain applications of COMPAS (and similar technologies), it puts pressure on the intuition, which is widely held and written into sentencing law and practice, that defendants' likelihood of committing future crimes is an appropriate basis for the length of prison sentences. Hence, reflecting on the application of new technologies gives us a chance to revisit an old practice. Having powerful tools to predict – well! – which defendants are indeed more likely to commit further crimes forces us (and gives us the opportunity) to turn a critical eye towards a longstanding practice that, if our arguments are successful, lacks the justification it needs to stand.

Acknowledgements

We are grateful to participants at the seventh meeting of the Philosophy, Politics, and Economics Society, participants at the University of Florida Workshop on Ethics in Criminal Justice AI, and several anonymous referees. Open access was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias: There's Software Used across the Country to Predict Future Criminals and It's Biased against Blacks." *ProPublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessmentsin-criminal-sentencing.

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

Corbett-Davies, Sam, and Sharad Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." CoRR, abs/1808.00023. http://arxiv.org/abs/1808.00023.

Duff, R. A. 1986. *Trials and Punishments*. Cambridge: Cambridge University Press.

Duff, R. A. 2001. *Punishment, Communication and Community*. Oxford: Oxford University Press.

Freeman, K. 2016. "Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State V. Loomis." *North Carolina Journal of Law & Technology* 18 (5): 75–106.

Golding, Martin P. 1975. *Philosophy of Law*. Englewood Cliffs, NJ: Prentice-Hall.

Hart, H. L. A. 1968. "Prolegomenon to the Principles of Punishment." In *Punishment and Responsibility: Essays in the Philosophy of Law*, 1–27. New York: Oxford University Press.

Helland, E., and A. Tabarrok. 2007. "Does Three Strikes Deter? A Nonparametric Estimation." *The Journal of Human Resources* 42 (2): 309–330. https://doi.org/10.3368/jhr.XLII.2.309. Accessed February 23, 2021. http://www.jstor.org/stable/40057307.

Kershnar, Stephen. 2000. "A Defense of Retributivism." *International Journal of Applied Philosophy* 14 (1): 97–117. https://doi.org/10.5840/ijap20001416.

Kershnar, Stephen. 2002. "The Structure of Rights Forfeiture in the Context of Culpable Wrongdoing." *Philosophia* 29 (1–4): 57–88. https://doi.org/10.1007/BF02379901

Klick, Jonathan, and Alexander Tabarrok. 2010. "Police, Prisons, and Punishment: The Empirical Evidence on Crime Deterrence." Chapters, In *Handbook on the Economics of Crime, Chapter 6*, edited by Bruce L. Benson and Paul R. Zimmerman, 127–144. Edward Elgar Publishing.

Latessa, E. J., S. J. Listwan, and D. Koetzle. 2015. *What Works (and doesn't) in Reducing Recidivism*. London: Routledge.

Lowenkamp, C., D. Hubbard, M. Makarios, and E. J. Latessa. 2009. "A Quasi-experimental Evaluation of Thinking for a Change: A 'real world' Application." *Criminal Justice and Behavior* 36 (2): 137–146. https://doi.org/10.1177/0093854808328230.

Lowenkamp, C. T., and E. J. Latessa. 2002. *Evaluation of Ohio's Halfway House and Community Based Correctional Facilities*. Cincinnati, OH: University of Cincinnati.

Mackie, J. L. 1982. "Morality and the Retributive Emotions." *Criminal Justice Ethics* 1 (1): 3–10. https://doi.org/10.1080/0731129X.1982.9991689.

Mastrobuoni, Giovanni, and David Rivers. 2016. "Criminal Discount Factors and Deterrence." IZA Discussion Paper No. 9769, SSRN. https://ssrn.com/abstract=2742557.

Moore, Michael S. 1987. "The Moral Worth of Retribution." In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, edited by Ferdinand Schoeman. Cambridge: Cambridge University Press.

Morris, Herbert. 1968. "Persons and Punishment." *The Monist* 52:475–501. https://doi.org/10.5840/monist196852436

Murphy, Jeffrie G. 1973. "Marxism and Retribution." *Philosophy & Public Affairs* 2 (3): 217–243.

Murphy, Jeffrie G. 2011. "Some Second Thoughts on Retributivism." In *Retributivism: Essays on Theory and Policy*, edited by Mark D. White, 93–106. New York: Oxford University Press.

Nagin, D. 2013. "Deterrence in the Twenty-first Century: A Review of the Evidence."

National Institute of Justice. 2016. "Five Things About Deterrence." June 5, 2016, nij.ojp.gov. https://nij.ojp.gov/topics/articles/five-things-about-deterrence.

Nino, C. S. 1983. "A Consensual Theory of Punishment." *Philosophy & Public Affairs* 12 (4): 289–306.

Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press.

Pereboom, D. 2019. "Free will Skepticism and Prevention of Crime." In *Free Will Skepticism in Law and Society*, 1st ed., edited by E. Shaw, D. Pereboom, and G. D. Caruso, 99–115. Cambridge, UK: Cambridge University Press.

Pereboom, D. 2020. "Incapacitation, Reintegration, and Limited General Deterrence." *Neuroethics* 13 (1): 87–97. https://doi.org/10.1007/s12152-018-9382-7.

Roodman, David. 2017. "The Impacts of Incarceration on Crime.". SSRN. Accessed September 25, 2017. https://doi.org/10.2139/ssrn.3635864 or https://ssrn.com/abstract=3635864.

Rubel, Alan, Clinton Castro, and Adam Pham. 2020. "Algorithms, Agency, and Respect for Persons." *Social Theory and Practice* 46 (3): 547–572. https://doi.org/10.5840/soctheorpract202062497.

Sher, George. 1987. *Desert*. Princeton, NJ: Princeton University Press.

von Hirsch, A. 2007. "The Desert Model of Sentencing: Its Influence, Prospects, and Alternatives." *Social Research* 74 (2): 413–443. https://doi.org/10.1353/sor.2007.0040.

Walen, Alec. 2021. "Retributive Justice." In *The Stanford Encyclopedia of Philosophy*, Summer 2021 Edition, edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2021/entries/justice-retributive/.

Wisconsin Supreme Court. 2016. *Wisconsin v. Loomis*, 881 N.W.2d 749.