Iudicium ex Machinae – The Ethical Challenges of Automated Decision-Making in Criminal Sentencing

Frej Klem Thomsen, Senior Researcher, DIHR

October 2020 draft

Abstract. *Automated decision making for sentencing is the use of a software algorithm to analyse a convicted offender's case and deliver a sentence. This chapter reviews the moral arguments for and against employing automated decision making for sentencing and finds that its use is in principle morally permissible. Specifically, it argues that well-designed automated decision making for sentencing will better approximate the just sentence than human sentencers. Moreover, it dismisses common concerns about transparency, privacy and bias as unpersuasive or inapplicable. The chapter also notes that moral disagreement about theories of just sentencing are plausibly resolved by applying the principle of maximising expected moral choiceworthiness, and that automated decision making is better suited to the resulting ensemble model. Finally, the chapter considers the challenge posed by penal populism. The dispiriting conclusion is that although it is in theory morally desirable to use automated decision-making for criminal sentencing, it may well be the case that we ought not to try.*

Keywords: Automated decision making; privacy; transparency; bias; criminal justice ethics; decision theory; moral uncertainty; penal populism

In this chapter, I will argue that the use of automated decision-making for sentencing is in principle morally permissible. By automated decision-making (ADM) generally, I mean a software algorithm whose output is or becomes a decision (MSI-NET 2017, Wagner 2019). In the context of sentencing this means an algorithm that is provided with appropriate input, for instance about the severity of a crime, and calculates an output in the shape of a just sentence. For it to be an automated decision-*making* system it is necessary that the algorithm's output ordinarily determines the sentence. An ADM for sentencing would thus occupy the role traditionally reserved for judges.

The use of algorithms in decision-making has been a controversial topic in recent years, particularly within criminal justice (Ferguson 2017, Veale, Van Kleek et al. 2018, Hannah-Moffat and Montford 2019). As such, I expect many will meet the initial claim of this chapter with scepticism. Much of the chapter will be devoted to considering a series of potential principled objections to the use of ADM in criminal sentencing. I argue that at least the most common objections are unpersuasive. However, I also engage with a currently underappreciated challenge to ADM for sentencing: the risk of misuse given the political climate of the societies that are currently poised to make use of ADM for criminal sentencing. Thus, despite defending the claim that the use of ADM for sentencing is in principle morally permissible, I will also argue that there are underappreciated reasons to be cautious about introducing it in practice.

The overall argument that I will pursue is as follows. In **section two** I discuss ADM for sentencing and argue that it is likely to better approximate the just sentence than human sentencing. In **section three** I discuss three general objections to ADM based on privacy, transparency and bias, and argue that none apply to ADM for sentencing. In **section four**, I review an assumption made in section two, that there are right answers in sentencing, dismiss two objections to this assumption, and discuss decisions under moral uncertainty as a practical obstacle to development of ADM for sentencing. Finally, in **section five**, I introduce penal populism as the most important challenge to ADM for sentencing. **Section six** summarises and concludes.

2. Rise of the robot judge

ADMs are a particular type of algorithm. The use of algorithms to support or deliver decisions is a much broader phenomenon, however, since an algorithm is simply a procedure for performing a particular set of logical or mathematical operations in order to translate input to a useful output (Hill 2016). As such, it is unsurprising that the use of algorithms in sentencing predates the recent rise of ADM based on machine learning. Earlier initiatives include so-called evidence-based sentencing, which base sentences (at least in part) on algorithmic risk assessments, and sentencing guidelines, which (broadly speaking) recommend sentences based on algorithms that combines various factors about the case.

An ADM for sentencing would resemble the former in certain respects and the latter in others. We might, for instance, feed the ADM the type of offence committed and details about the offender, and the ADM would respond with, say: “four months imprisonment”. Since we are here concerned with ADM, not merely with algorithms that support or recommend decisions, this sentence would then be executed by the court.

It is worth emphasising that an ADM for sentencing would by necessity preserve rather a large amount of human discretion. It would not, for instance, determine which offence to charge the offender with. Since it is notoriously often possible for prosecutors to charge the same wrongdoing as several different offences – offences which may have very different sentences associated with them – prosecutorial decision-making could still heavily influence sentencing (Shermer and Johnson 2010, Smith and Levinson 2012, see also Lippke in this volume). An ADM for sentencing will also realistically include features where determining the input value requires the exercise of judicial discretion. Suppose for instance that mitigating circumstances is a feature in the ADM. The algorithm would respond to the presence or absence of mitigating circumstances when determining the sentence, but it would not by itself be capable of determining whether there are in fact mitigating circumstances in the case. The number and influence of such decisions will depend on the input required by the ADM, but I cannot conceive of a plausible ADM for sentencing that would not require substantial judicial involvement.

With this loose picture of ADM for sentencing in mind, what are we to think of its use? Should we welcome our new robot overlords, or charge to the luddite chant of "Enoch shall break them"? In the limited debate so far responses have been divided (Roth 2016, Selbst 2017, Stobbs, Bagaric et al. 2017, Bagaric and Wolf 2018, Chiao 2018, Simmons 2018, Chiao 2019, Donohue 2019). In the remainder of this section we will first look at one possible concern, the analysis of which will help to further set the stage of the discussion. In the subsequent sections we will then review a series of common objections to ADM, and two difficulties that have received very little attention.

Can the machine judge?

Perhaps the most immediate concern with the use of ADM for sentencing might be whether a software algorithm is well suited to making this type of decision in the first place. Just sentencing involves the careful weighing of a range of different factors that vary between cases, a task that requires sound judgment and years of training and experience for human judges (Donohue 2019, see also Schwarze & Roberts in this volume). Call this **the objection from the inherent superiority of human sentencing**:

> *An ADM will be incapable of taking into account the unique features of each case and performing the careful weighing of these features to deliver a just sentence*

The objection may look appealing at first glance. After all, a criminal case involves human beings with individual histories and every case is unique. Taking account of the unique features of each case requires a certain cognitive flexibility, the opponent of ADM might say, which a software algorithm does not possess. Furthermore, all the relevant features must be taken into consideration and carefully weighed against each other, a task that is simply too complex for an algorithm. Hence, human sentencing is inherently superior to ADM for sentencing.

Whatever its immediate appeal, the objection is ultimately unpersuasive. The first part of the objection misunderstands the role of principles in sentencing, while the latter part gets the relative strength of humans and algorithms backwards.

The first part of the objection, recall, advances the claim that constructing an ADM that incorporates the infinite range of features relevant to individual cases is impossible (or at least currently unachievable).

Sentencing, it might be said, depends on such a range of different factors about the case, many of them particular to an individual case or a small set of cases, that any attempt to construct a model that incorporates them all is doomed from the outset.

This part of the objection seems to me clearly misguided. It is of course true that every case will have an infinite number of properties (the offender's hair colour, the victim's favourite brand of cereal, the temperature in the city of Marrakesh at the time of the offence, etc.). However, only a small number of these properties are relevant to sentencing, and it is neither chance nor personal preference that determines whether a particular property is relevant to sentencing, or how it affects the just sentence. The proper role in sentencing of any particular property in a case is determined by the moral reasons that apply to sentencing. Such reasons are what scholars attempt to determine in criminal justice ethics and summarise as what we might broadly call a theory of just sentencing. Current scholarship in criminal justice ethics is divided between competing theories of just sentencing, ranging from retributivist theories across compensation-based and restorative theories to utilitarian theories (E.g. Walker 1991, Moore 1997, Pettit 1997, Duff 2001, Braithwaite 2002, Von Hirsch and Ashworth 2005, Boonin 2008, Tadros 2011, for an overview see Duff and Hoskins 2019). But it is the theory of just sentencing we adopt that allows us to say which properties of cases affect just sentencing and how. This is true regardless of whether sentencing is carried by a human judge or by ADM. Hence, even if it is impossible to specify, for instance, an exhaustive list of mitigating circumstances, it will not be impossible to construct an ADM that gives mitigating circumstances a particular role in sentencing and allow the human judge to input whether mitigating circumstances are present or not (cf. Bagaric and Gopalan 2015). On any given theory, then, the individual properties of a case are either irrelevant, in which case it is no cause for concern that an ADM for sentencing does not include them as features, or specified by our theory of just sentencing, which allows an ADM to incorporate them, at least when humans provide the necessary input. (cf. Chiao 2019; for an illustration related to mercy, see Dagan & Baron in this volume)

The second part of the objection, which claimed that ADM is unsuited to the complex weighing of multiple factors necessary for sentencing, fares little better. Human judgement, particularly intuitive judgment of the kind sometimes labelled system 1 thinking by cognitive scientists, is eminently unsuited to the complex task of carefully weighing multiple factors (Kahneman and Tversky 2009, see also Chiao in this volume). We tend in such situations to rely on cognitive heuristics, and to later rationalise the answers provided by these shortcuts. Unlike human cognition, however, an ADM is perfectly capable of calculating the function for multiple, individually weighted features, including nonlinearities and interacting features. In fact, this difference in capacities is the fundamental reason for one of the common challenges raised by the use of ADM: the lack of transparency (I return to this point below). As the complexity of a model increases, it swiftly exceeds the human ability to simultaneously consider all of the components (Rudin 2019, see also Ryberg & Petersen in this volume).

The machine-learning shortcut as dead-end

On the other hand, fans of ADM may hope for too much in the case of sentencing. The hype currently surrounding "algorithms" is to a large extent based on the impressive achievements of models developed by machine learning. Unfortunately, sentencing is not a decision problem well-suited to machine learning. Machine learning excels in situations where we know the target value for a large number of examples but do not know how features produce or predict the target value. In these situations, a learning algorithm can detect correlations in the data and train a model that optimally solves the problem of predicting the target value for new examples. If, for instance, we want to train a model to predict rain, we can feed a learning algorithm with examples in which it did and did not rain. Importantly, we can rely on historical observations to determine whether it rained on any particular day, that is, we can establish the target value for our training examples without relying on the features of our model.

Sentencing, however, is not such a situation. We want to develop a model that will output just sentences, but there is little reason to think that historical cases have received just sentences. On the contrary, we have good reason to believe that current systems of criminal justice tend to overpunish, often dramatically so (cf. Bagaric and Wolf 2018). This follows if criminal responsibility cannot, as is commonly assumed, justify punishment (Boonin 2008, Zimmerman 2011, Thomsen 2018), if punishment is justified by its supposed deterrent effect (Von Hirsch, Bottoms et al. 1999, Doob and Webster 2003, Robinson and Darley 2004, Kennedy 2009, Braga and Weisburd 2012), and even on important theories that take offenders' just desert to impose upper limits on severity (Murphy 1979, Von Hirsch 1996, Tonry 2016, Husak 2019). If existing systems systematically overpunish, datasets of historical cases may consist mostly or even entirely of cases that have *not* received a just sentence. Training an ADM with machine learning on the historical data in this situation is practically pointless – at best the result would be a model that slightly more efficiently and consistently reproduced the fundamental injustices of our current sentencing practices.

Is this a decisive blow to ADM for sentencing? Not necessarily. The criminal justice ethics literature contains a well-established set of theories of just sentencing. Such theories detail the features that determine the just sentence.[i] We are thus perfectly capable of developing an ADM for sentencing that delivers just sentences through old-fashioned human programming. If the resulting model is complex, an ADM is likely to be better than humans at weighing the features that jointly determine the just sentence, and it will in any case eliminate individual human biases and randomness from sentencing (Kleinberg, Lakkaraju et al. 2017, Selbst

2017, Bagaric and Wolf 2018, Chiao 2019, see also Wingerden & Plesnicar in this volume). These advantages form the foundation of **the conditional argument for the use of ADM in criminal sentencing**:

1) A properly designed ADM for sentence will better approximate the just sentence than human judges.
2) We have a moral reason to employ the sentencing method that best approximates the just sentence.
3) We have a moral reason to employ ADM for sentencing (derived from 1 and 2).
4) We ought to employ whichever sentencing method is favoured by the balance of reasons.
5) A sentencing method is favoured by the balance of reasons if there is a reason in favour, and all else is equal.

C) All else equal, we ought to employ a properly designed ADM for criminal sentencing (derived from 3, 4 and 5).

The argument is valid. I have defended premise 1 above, and premise 2 will be acceptable to anyone who holds that there is a such a thing as a just sentence. Premise 3 is a conclusion from premises 1 and 2, and premises 4 and 5 are very plausible general principles of moral philosophy. Thus, I venture that the argument is also sound. The argument is conditional, however, so that opponents of ADM need only show that there are other reasons *against* using ADM. Thus, in the remaining sections of this chapter, I will evaluate a number of arguments to the effect that all else is, in fact, far from equal.

3. Rage against the machine: privacy, transparency and bias

Critics of ADM tend to claim either that there is some property of human decision-making which ADM lacks, or that there is some property of ADM which does not (or at least need not) exist in human decision-making. In either case, the difference is held to count against the use of ADM. Arguably the three most prominent such objections to the use of ADM is that ADM infringes on privacy, lacks transparency, and can be biased against vulnerable minority groups. (Ji, Lipton et al. 2014, Mittelstadt, Allo et al. 2016, Jaume-Palasí and Spielkamp 2017, Lepri, Oliver et al. 2018, Chiao 2019) In this section, we will consider each of these objections in turn.

When faced with an objection along the above lines, the proponent of ADM has at least three potential responses: i) she can deny that ADM lacks the desirable property or possesses the undesirable property, ii) she can reply *tu quoque*, that human sentencing also lacks the desirable property or possesses the undesirable property, or iii) she can challenge the normative claim, that the difference makes the use of ADM for sentencing impermissible. Although to my mind the third of these responses is perhaps the most interesting and underexplored, it turns out that in the case of ADM for sentencing there are readily available responses along the first two lines for each of the three general objections.[ii] As such, I will argue here that, regardless of their strength as objections to ADM more generally, all three objections meet decisive versions of responses i) and ii) in the present context.

Privacy

It has already become a truism to point out that ADM specifically and the use of algorithms more generally often involves access to large amounts of data in both the development and use phases. It is also clear that collecting, sharing, and employing data can decrease the privacy of the persons to whom the data pertains (Crawford and Schultz 2014, Ji, Lipton et al. 2014). Thus, it might seem that a significant disadvantage of ADM for sentencing will be its cost in terms of privacy. Call this **the privacy-based objection to ADM for sentencing**:

> *The use of ADM for sentencing will cause a loss of privacy.*

To evaluate this objection it is necessary to briefly consider how data is employed in developing and using an ADM.

In the development phase, access to larger amounts of data typically allows the development of a model that can more accurately solve its task, particularly if the developer employs machine learning. This benefit applies to both the number of examples in the training data and the number of features associated with each example. A greater number of training examples allows easier separation of the valid statistical correlations (signal) from random correlations (noise). A greater number of relevant features – individual pieces of information, such as age, gender, occupation, educational level, or past offences – allows more fine-grained evaluation, and thus also promotes accuracy. A central strength of machine learning is its ability to discover hitherto unknown patterns in data. This gives developers an incentive to include *potentially* relevant features in the training data and let the learning algorithm sort useful features from chaff.[iii] The result is often a dataset for development which is both as broad (features) and as deep (examples) as is practically possible.

Meanwhile, in the use phase, the model relies on data about the target's features to evaluate the target value, whether it is the expected sales price of a house (a standard example), the chance of rain next weekend, or, as in the present context, the just sentence for a particular offence. Here again, a well-designed model will tend to be more accurate the more relevant features it employs, which gives developers an incentive to develop an ADM that employs as much relevant data as possible.

The powerful incentives to employ large amounts of data in both development and use of ADM has rightly raised concerns that ADM can infringe personal privacy. It remains an open question in moral philosophy how strong this objection is, at least in part because the underlying issue – the badness of privacy loss – is itself controversial and (I believe) underexplored (Ryberg 2007, Macnish 2018, Thomsen 2020). However, we do not need to solve this issue here, because the privacy-based objection to ADM for sentencing is vulnerable to a different and decisive response: a suitable ADM for sentencing need not employ more data or different data than a human judge. This is in large part because, as I have argued above, machine learning is unsuitable for an ADM for sentencing. Indeed, developing a suitable ADM on the basis of criminal justice ethics need involve no data, apart perhaps from data that is currently available in the criminological literature (e.g. to evaluate the effect on general deterrence of differences in sanctioning severity). Meanwhile, in the use phase an ADM for sentencing will rely on the same features as a human judge does (or ought to), and therefore will not need to access different data than a judge. This does not mean that privacy is irrelevant to sentencing. Perhaps there are features which we have some reason to consider in sentencing, but which we ought all-things-considered to not employ because doing so would diminish privacy. The point is that, even were that the case, the constraint would apply equally to human sentencing and ADM for sentencing.

At worst, therefore, an ADM for sentencing need be no more privacy-reducing than human sentencing. Arguably, however, access to data by an algorithm is *less* privacy-infringing than access to the same data by a person (Macnish 2012, Macnish 2017). While this point is controversial and will only be relevant where access by the ADM substitutes rather than supplements access by humans, it means that at best ADM for sentencing will enjoy an advantage with respect to preserving privacy, as compared with human sentencing.

Transparency

A second prominent general objection to ADM is that ADM systems can be opaque, and that this makes reviewing and challenging decisions made by ADM difficult or even impossible (Lepri, Oliver et al. 2018, Chiao 2019). In decisions of such importance as criminal sentencing, surely it is paramount that those subject to the decision can understand the basis of the decision, and critically evaluate whether the sentence is in fact just? If human sentencing is transparent while ADM for sentencing is opaque, it would seem that human sentencing enjoys a clear moral advantage (For further critical discussion of this claim, see chapters by Ryberg and Chiao in this volume). Call this **the transparency-based objection to ADM for sentencing**:

> *The use of ADM for sentencing will render sentencing decisions opaque, and thereby hamper review and challenge of unjust sentencing decisions*

To evaluate the objection, it is worth briefly discussing how transparency in ADM works. Although there is no mathematically precise definition, it is generally accepted that an algorithm is transparent if it is possible for the relevant person to understand how it works, a litmus test for which is that the person is able to reliably predict the algorithm's output for any particular input (Guidotti, Monreale et al. 2019, Molnar 2019). As such, algorithmic transparency is basically a function of two factors: access and complexity.

Access to information about the ADM, such as its source code or a mathematical representation of the model, is self-evidently necessary for transparency. Algorithmic opaqueness is frequently a consequence of the development of ADM by private companies who have a business interest in keeping the source code and model confidential. However, there is no obvious reason why an ADM for sentencing would need to be developed by private interests, or why the model would need to be kept private even if it was. On the contrary, it will presumably be an indispensable condition for both the development and use of an ADM for sentencing that the model is fully available to the public, just as current sentencing regulations and guidelines are.

The second requirement for understanding how an ADM works is that the model is not overly complex. Some ADM, particularly when developed by machine-learning, can use hundreds or thousands of features and contain complex representations of non-linearities and interactions between features. The resulting model can be impossible to grasp in its entirety even for experts, simply because there are too many interlocking parts (Molnar 2019, Rudin 2019).

Depending on our theory of just sentencing, complexity may be a challenge for ADM for sentencing. Our theory of just sentencing may contain many features, the function for each feature may be complex, and there may be interactions between features. If so, an ADM based on the theory may well be too complex to be interpretable. Any such complexity, however, will be a result not of employing ADM but of our theory of just sentencing. The complexity, in other words, will apply equally to human sentencing that employs the same theory of just sentencing. And conversely, if we believe that human sentencing ought to deviate from our theory of just sentencing in order to make sentencing transparent, then presumably exactly the same will apply to ADM for sentencing. In terms of opaqueness produced by complexity, therefore, there is no reason an ADM should do worse than human sentencing.

Furthermore, it is worth repeating, as has been frequently observed, that human decision-making is far from transparent (Kleinberg, Lakkaraju et al. 2017, Zerilli, Knott et al. 2018). It is, of course, impossible to observe directly how another person makes a decision. Indeed, it is very difficult to observe with any precision how one makes one's *own* decisions. Thus, it is no surprise that even when required to provide reasons for a decision, humans have a lamentable tendency to provide reasons that are (to put it diplomatically) inaccurate. In the transparency competition between human sentencing and ADM for sentencing, ADM may therefore have an advantage. The two parties can employ the same potentially complex and opaque theory of just sentencing, but at least in the case of an ADM we can verify that the theory – the whole theory, and nothing but the theory – has been rigorously applied in every single case.

Bias

The third, and perhaps most prominent common challenge to ADM is that ADM can be biased against vulnerable minorities, despite the developer's best intentions. This has been clearly demonstrated in the literature on risk assessment in criminal justice. (Angwin, Larson et al. 2016, Kleinberg, Mullainathan et al. 2016, Chouldechova 2017, Ensign, Friedler et al. 2017, Berk, Heidari et al. 2018, Chouldechova and Roth 2018, Dressel and Farid 2018, Chiao 2019) Such biases are often the result of existing inequalities, either in treatment or structurally, and there is no doubt that criminal justice is an area where there are ample examples of both treatment inequalities and relevant structural inequalities. (Thomsen 2011, Ferguson 2017) Thus, should we not expect that an ADM for sentencing will be vulnerable to biases? And will using ADM for sentencing therefore not reproduce or exacerbate existing and morally repugnant inequalities? (For further discussion of bias-based objections, see also chapters by Douglas & Davis and Lippert-Rasmussen in this volume) Call this **the bias-based objection to ADM for sentencing**:

> *An ADM for sentencing will be biased against vulnerable minorities, and its use will therefore reproduce or exacerbate inequalities*

As with the previous two objections, it is necessary to briefly explore how ADM can become biased against vulnerable minorities in order to evaluate the objection.

An ADM can become biased in essentially two ways: it can reproduce biases in the training data, or it can employ features whose values vary systematically across relevant populations. For illustration consider the prediction of recidivism risk. A common way of measuring recidivism is by recorded re-arrests. However, police may treat different population groups very differently when deciding whom to arrest. Offenders who are members of religious or ethnic minorities, in particular, may be arrested at much higher frequencies than offenders from majority groups. This difference in treatment will inflate the perceived recidivism of minority members in the data, and lead an ADM trained on such data to predict higher recidivism for members of these groups. In this case, a bias in the training data has been reproduced by the ADM (cf. Ensign, Friedler et al. 2017).

An ADM may also be biased simply because it employs features whose values vary systematically across populations. In risk assessment, it may well be the case that recidivism varies across racial or ethnic groups, and that important risk predictors, such as poverty, educational level, employment history, or criminal record, similarly vary. In such cases an ADM will treat the two groups differently, predicting higher risks for members of some groups, making more mistakes when evaluating some groups, or making different *types* of mistakes when evaluating different groups (Barocas and Selbst 2016, Kleinberg, Mullainathan et al. 2016, Chouldechova 2017, Chouldechova and Roth 2018, Kleinberg, Ludwig et al. 2019).

Now, let us consider bias in an ADM for sentencing. An ADM for sentencing will not reproduce biases in the training data, since as I have argued, machine learning on historical data is not suitable for developing ADM for sentencing. The first source of biases therefore does not (or at least should not) apply to ADM for sentencing. An ADM for sentencing will, however, be vulnerable to biases of the second type. One solution would be direct discrimination in favour of the vulnerable minority, that is to make ethnic or racial identity a feature in our theory of just sentencing such that sentences would in at least some cases be made more lenient for minority members (Lipton, Chouldechova et al. 2018). This would be a form of affirmative action, and although controversial such policies are arguably frequently justified (Lippert-Rasmussen 2020). However, we need not pursue an argument to the effect that ADM for sentencing can avoid the problem posed by biases by resorting to affirmative action. This is because once again the problem emerges not because of our use of ADM but as a result of structural inequalities and our theory of just sentencing. Thus, any biases that result from our use of ADM will apply equally to human sentencing that employs the same theory. In fact, scholars in the literature on bias in machine-learning have been keen to emphasise that since these biases are mathematically unavoidable they apply equally to human decisions (including all of our past decisions) (Kleinberg, Lakkaraju et al. 2017, Kleinberg, Ludwig et al. 2019). ADM may in that perspective serve to helpfully force us to confront these biases and develop principled ways of dealing with them. Furthermore, ADM for sentencing will be able to avoid the biases that demonstrably affect human judges (Rachlinski, Johnson et al. 2009, Kang, Bennett et al. 2012, Liu and Li 2019, see also Chiao in this volume). In conclusion, it appears that with respect to biases too, ADM for sentencing is at least as good as and potentially enjoys an advantage over human sentencing.

4. Is there a right answer in criminal sentencing?

In the previous sections I have argued that an ADM based on a theory of just sentencing will be at least as good as and plausibly better than human sentencing in several important respects. I have been assuming throughout that there is such a thing as a just sentence, that is, that there is a correct answer to the question of what sentence the court ought to impose upon any specific offence. Some readers may have balked at this assumption for any one of several related reasons. Perhaps there are no right answers in criminal justice ethics, or indeed in ethics more generally. Or perhaps there is a right answer, but we cannot currently plausibly claim to know what it is. In this section, we will deal with these two objections in turn, showing that the former is is unpersuasive, and that the latter is plausibly resolved in a way that favours ADM over human sentencing.

The objections at stake in this section focus on the way the ADM for sentencing I have sketched requires a theory of just sentencing. ADM for sentencing, I have suggested, should be developed by constructing a model of the features (and weights) our theory of just sentencing specifies as the determinants of the just sentence. This type of ADM for sentencing is therefore impossible if there are no right answers in sentencing. Why might that be the case? The first, and most sweeping suggestion is what I will call **the objection from moral scepticism**:

> *The right answers required by ADM for sentencing do not exist, because there are no right answers in (criminal justice) ethics*

In response, it is worth first recalling that debate within criminal justice ethics tends to proceed on the assumption that there are in fact right answers to questions within ethics. Metaethical scepticism is a viable position, but it is hardly the consensus view that scholars outside of Philosophy sometimes assume.[iv] Given the complexity of the debate in metaethics, however, it would be preferable if we could find a response to the objection that does not involve a full-fledged defence of moral realism. Fortunately, it seems clear that we can.

The most readily available response is, I believe, that if there is no right answer in sentencing, *because* there are no right answers in ethics, then it becomes difficult to see on what grounds the objection is meant to stand. After all, developing an ADM cannot be morally impermissible if there are no right answers in ethics. The objection can show, if we accept the sceptical claim, that we cannot develop ADM for sentencing that delivers just sentences, but not only is human sentencing similarly incapable of delivering just sentences, but this failure cannot count against the moral permissibility of either. At most, the opponent of ADM who offers this objection could thereby express their personal dislike of ADM for sentencing – the influential emotivist metaethics championed by Alfred Ayer, which holds that this is precisely how moral language works, is often memorably referred to as “boo-hurrah-theory” (Ayer 2002) – but again, supposing that there are no right answers in ethics, this dislike does not give the proponent of ADM a moral reason to alter the ADM or refrain from using it. As the grounds of a moral objection to ADM for sentencing, moral scepticism is self-defeating.

A related and more powerful objection is based upon the idea that given theoretical disagreements in contemporary criminal justice ethics we do not *know* what the right answer is. It follows, the opponent might say, that if ADM for sentencing requires us to specify the right theory of just sentencing, then we cannot develop ADM for sentencing. Call this **the objection from theoretical disagreement**:

> *Given current scholarly disagreement on the correct theory of just sentencing, we cannot specify an appropriate theory of just sentencing for the development of ADM for sentencing*

The objection is partly correct but faces two convincing responses. It is arguably correct in claiming that given current disagreements, we cannot confidently identify one theory of just sentencing as the true theory. First, however, the problem of adopting a particular theory of just sentencing in such circumstances applies equally to human sentencing. As with so many objections before, ADM fares no worse, it simply forces us to confront unpleasant difficulties that we might otherwise prefer to ignore. Second, the problem has an interesting solution in the shape of decision-theory for moral uncertainty. Since it can significantly affect how we ought to develop ADM for sentencing, it seems to me worthwhile to briefly sketch how we can apply such decision-theory to solve the problem.

Decision-making under moral uncertainty has been recognized as a metaethical problem for several decades, but it has drawn increased attention in recent years (Bykvist 2017, MacAskill, Bykvist et al. Forthcoming). The basic problem will be clear to most: even if we think that a particular moral theory is true, we typically recognize that there is at least some probability that other theories could be true instead. What should we do in the light of this moral uncertainty?

One response would be to avoid the problem by looking for theoretical common-ground. It is possible that there are areas of just sentencing where one answer or policy dominates others, that is, the answer or policy is held to be at least as good as all alternatives by *every* plausible theory of criminal justice ethics. In these cases, competing theories are “climbing the mountain” – approaching the issue from different sides, only to find themselves emerging at the same summit, in the shape of identical moral conclusions (Parfit 2011). However, I suspect that we are going to encounter dominant answers infrequently, if at all.

Another tempting response might be to follow one’s favourite theory. If nothing else, making decisions in accord with whatever theory we think most likely to be true and ignoring the rest has the virtue of simplicity. Unfortunately, this strategy has intuitively unappealing implications in many situations. Consider the following version of a common case:

> ***Hedging.*** A judge believes that (some form of) consequentialism is true or (some form of) deontology is. She considers the former slightly more likely than the latter (55% vs 45% chance). The judge now faces a choice between handing down one of two sentences, S_1 and S_2. S_1 is permissible but only very slightly better than S_2 according to her consequentialist theory. S_2 is permissible according to her deontological theory, while S_1 is impermissible, indeed morally horrendous.

Should the judge inflict S_1? Intuitively, this seems wrong. The judge ought to “hedge her bets” and inflict S_2, because doing so will only be very slightly worse if her consequentialist theory is true, while allowing her to avoid a 45% chance of doing something morally horrendous.

A decision-theory capable of accommodating this intuition is the principle of maximising moral choice worthiness (MacAskill and Ord 2020, Riedener 2020). For each possible act, we ask how good or bad the competing moral theories hold that act to be. The moral choice worthiness of the act is then calculated as the sum of the moral value each theory ascribes to the act weighted by our subjective probability that the theory is true, and we choose the option that has the highest moral choice worthiness.

In the context of just sentencing, the result would be an ensemble model. An ensemble model contains several different models, each of which individually evaluates the target value, and combines their results by some mechanism, e.g. weighted voting. The most familiar example is probably so-called random forests, which combine a set of individual decision trees, each of which has been trained slightly differently. Ensemble models are common in machine-learning because they often significantly outperform individual models. In the sentencing context, a simple ensemble ADM might aggregate the decisions of a series of individual models based on competing theories of just sentencing, weighting each decision by our subjective probability for the relevant theory. Sentencing is, in that respect, a well-suited decision problem, since sentences are (in theory, at least) located on a continuous scale of severity.

As an illustration, consider an ensemble model of three competing theories T_1, T_2, and T_3 , and suppose that we believe that there is a respectively 45%, 30% and 25% chance of their being true. If T_1 recommends a sentence of 30 days imprisonment, T_2 a sentence of 180 days imprisonment, and T_3 a sentence of 135 days imprisonment, then weighted aggregation will deliver a sentence of approximately 101 days.

However, the simple application presupposes that all theories value deviations from their recommended sentence linearly and equally. This assumption does not hold in practice since many theories of just sentencing hold both that the disvalue of deviations from the recommended sentence increases non-linearly and that it increases much more steeply for deviations in the direction of overpunishment (Duus-Otterström 2013). Thus, a proper ensemble model will introduce individual functions for the disvalue of deviations from the just sentence, which might lead to very different results in aggregation.

The principle of maximising expected choice worthiness faces important theoretical challenges (Tarsney 2018, Riedener 2019, Riedener 2020). My purpose here is not to defend it, but only to illustrate how there are promising ways of solving the problem posed for a theory of just sentencing by theoretical disagreement in criminal justice ethics. Beyond the theoretical challenges, there is the daunting practical challenge of constructing an actual ensemble model. This would undoubtedly require herculean efforts even for a skilled and devoted group of criminal justice ethicists.[v] There is no guarantee, however, that their efforts would be appreciated. This is the ultimate and most important objection to the use of ADM for sentencing.

5. The penal populist challenge

In the above I have presented a reason that supports the use of ADM for sentencing and dismissed a range of common concerns as unfounded. So far, it seems as if there is a fair case for the use of ADM in criminal sentencing. In this penultimate section, I will temper this conclusion by presenting what seems to me the greatest threat to ADM for sentencing: the influence of penal populism, and the resulting likelihood that ADM for sentencing will be *worse* at approximating the just sentence.

Consider the following three claims:

1. The general trend in penal policies in for the past decades has been towards ever more severe punishment.
2. An important driver of the above trend is penal populism, i.e. the political use of appeals to uninformed punitive attitudes among the public (Roberts, Stalans et al. 2002, Pratt 2007, Wood 2014).
3. The judiciary plays a moderating role in this development, exercising their judicial discretion to reduce the impact of politically mandated increases in severity of punishment.

The three claims will not be true of every society, but they seem to me to be true of some societies, perhaps many. When true, they have important implications for the conditional argument for the use of ADM for sentencing, since they ground **the penal populist objection**:

> *ADM for sentencing will be worse than human sentencing at approximating the just sentence, because ADM for sentencing will be developed under the influence of penal populism.*

The argument is simple. If ADM for sentencing is developed under the influence of penal populism, then the theory of just sentencing on which the model is based is likely to be at least as draconian as the existing sentencing regime.[vi] Furthermore, since judges no longer exercise a moderating influence, the resulting sentences will be more severe even if the same theory of just sentencing is employed. An increase in sentencing severity would, I have previously argued, be bad, since there is good reason to believe that European and Anglophone societies already overpunish (see also Thomsen 2014).

The real danger in using ADM in criminal sentencing is therefore not, I believe, that an ADM will make mistakes, reduce privacy, be opaque or biased. In all these respects we have reason to believe that a properly developed ADM will do as well as or better than human sentencing. The real danger is that we won't get that ADM, but a model tailored to suit penal populism that makes our current woes worse.

6. Summary and conclusion

Over the course of this chapter I have sketched what an ADM for sentencing might look like, and argued that the use of such an ADM is in principle morally desirable. ADM for sentencing, I have suggested, should not be developed with machine learning, but we can employ our theories of just sentencing to handcraft a suitable model. Such a model will do no worse than and is likely to enjoy several advantages over human sentencing. Notably, I have shown why prominent concerns that are often raised about ADM do not apply to the ADM for sentencing I have proposed. If anything, an ADM for sentencing should perform better than human sentencing at preserving privacy, providing transparency, and avoiding bias. I have also argued that it is difficult to see how moral scepticism can support an objection to ADM for sentencing, and that the problem posed by theoretical disagreements in criminal justice ethics has an interesting solution amenable to the use of ADM, in the shape of an ensemble model based on decision theory for moral uncertainty. Finally, I have noted an important, and I believe underappreciated, difficulty for the use of ADM for sentencing: the risk that penal populism will produce an ADM that is worse than current sentencing practices at approximating the just sentence.

In Greek drama, a deus ex machina is the sudden resolution of a difficult plot point by implausible means, such as impromptu divine intervention. The concept carries clear negative connotations. A deus ex machina is disillusioning – a shortcut to the resolution that robs the preceding drama of its intensity and meaning. The influence of penal populism means that we risk having a similar experience with a iudicium ex machina. Rather than resolving our (many) difficulties in criminal justice, ADM for sentencing may be a technical fix that only delivers more of the injustice with which we are so familiar.

Taking another cue from the roots of philosophy, the Delphic temple of Apollo contained the inscribed maxim "Γνῶθι σεαυτόν" – know thyself. As moral agents we must make decisions in the light of our self-knowledge, including knowledge of our proclivity for making particular types of mistake. Our susceptibility to penal populism and our awareness of this fact leaves us in something akin to the classical dilemma of Goldman's Professor Procrastinate, who ought to accept the task of reviewing a paper, but knows that if she does, she will fail to produce a timely review. (Goldman 1978) If we could trust ourselves to do it right, it would be best to employ ADM for sentencing. However, given what we know about our recent history with criminal justice, it may well be that we are obligated to not even try.

References

Angwin, J., J. Larson, S. Mattu and L. Kirchner 2016. "Machine Bias." *ProPublica*.
Ayer, A. J. 2002. *Language, Truth and Logic*, New York: Dover Publications Inc.

Bagaric, M. and S. Gopalan. 2015. "Saving the United States from Lurching to Another Sentencing Crisis: Taking Proportionality Seriously and Implementing Fair Fixed Penalties." *Saint Louis University Law Journal 60* (2): pp. 169-242.

Bagaric, M. and G. Wolf. 2018. "Sentencing by Computer: Enhancing Sentencing Transparency and Predictability, and (Possibly) Bridging the Gap Between Sentencing Knowledge and Practice." *George Mason Law Review 25* (3): pp. 653-709.

Barocas, S. and A. D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review 104* (3): pp. 671-732.

Berk, R., H. Heidari, S. Jabbari, M. Kearns and A. Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research*. Doi: 10.1177/0049124118782533

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

Braga, A. A. and D. L. Weisburd. 2012. "The Effects of Focused Deterrence Strategies on Crime: A Systematic Review and Meta-Analysis of the Empirical Evidence." *Journal of Research in Crime and Delinquency 49* (3): pp. 323-358.

Braithwaite, J. 2002. *Restorative Justice & Responsive Regulation*. Oxford: Oxford University Press.

Broome, J. 2013. *Rationality Through Reasoning*. Chichester: Wiley-Blackwell.

Bykvist, K. 2017. "Moral uncertainty." *Philosophy Compass 12* (3).

Chen, L. 2009. Curse of Dimensionality. In *Encyclopedia of Database Systems*, edited by L. Liu and M. T. Özsu. Boston, MA: Springer, pp. 545-546.

Chiao, V. 2018. "Predicting Proportionality: The Case for Algorithmic Sentencing." *Criminal Justice Ethics 37* (3): pp. 238-261.

Chiao, V. 2019. "Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice." *International Journal of Law in Context 15* (2): pp. 126-139.

Chouldechova, A. 2017. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *Big Data 5* (2): pp. 153-163.

Chouldechova, A. and A. Roth. 2018. "The Frontiers of Fairness in Machine Learning." *arXiv e-prints*.

Crawford, K. and J. Schultz. 2014. "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms." *Boston College Law Review 55* (1): pp. 93-128.

Crisp, R. 2006. *Reasons & the Good*. Oxford: Oxford University Press.

Donohue, M. 2019. "A Replacement for Justitia's Scales? Machine Learning's Role in Sentencing." *Harvard Journal of Law and Technology 32* (2): pp. 657-678.

Doob, A. N. and C. M. Webster. 2003. "Sentence Severity and Crime: Accepting the Null Hypothesis." *Crime and Justice 30*: pp. 143-195.

Dressel, J. and H. Farid. 2018. "The accuracy, fairness, and limits of predicting recidivism." *Science advances 4* (1).

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

Duff, A. and Z. Hoskins. 2019. "Legal Punishment." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Duus-Otterström, G. 2013. "Why Retributivists Should Endorse Leniency in Punishment." *Law and Philosophy 32* (4): pp. 459-483.

Ensign, D., S. A. Friedler, S. Neville, C. Scheidegger and S. Venkatasubramanian. 2017. "Runaway Feedback Loops in Predictive Policing." *arXiv e-prints*..

Ferguson, A. G. 2017. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. New York: NYU Press.

Goldman, H. S. 1978. "Doing the Best One Can." In *Values and Morals*, edited by A. Goldman and J. Kim, pp. 185-214. Dordrecht*:* Springer Netherlands.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi and F. Giannotti. 2019. "A Survey Of Methods For Explaining Black Box Models." *ACM Computing Surveys 51* (5).

Hannah-Moffat, K. and K. S. Montford. 2019. "Unpacking Sentencing Algorithms: Risk, Racial Accountability and Data Harms." In *Predictive Sentencing - Normative and Empirical Perspectives,* edited by J. W. De Keijser, J. V. Roberts and J. Ryberg, pp. 175-196. Oxford: Hart Publishing.

Hill, R. K. 2016. "What an Algorithm Is." *Journal of Philosophy & Technology 29* (1): pp. 35-59.

Huemer, M. 2007. *Ethical Intuitionism*. New York: Palgrave Macmillan.

Husak, D. 2019. "Why Legal Philosophers (Including Retributivists) Should Be Less Resistant to Risk-Based Sentencing." In *Predictive Sentencing - Normative and Empirical Perspectives,* edited by J. W. De Keijser, J. V. Roberts and J. Ryberg, pp. 33-50. Oxford: Hart Publishing.

Jaume-Palasí, L. and M. Spielkamp. 2017. "Ethics and algorithmic processes for decision making and decision support." AlgorithmWatch.

Ji, Z., Z. C. Lipton and C. Elkan. 2014. "Differential Privacy and Machine Learning: a Survey and Review." *arXiv e-prints*.

Kahneman, D. and A. Tversky. 2009. *Choices, Values, and Frames*. New York: Cambridge University Press.

Kang, J., M. Bennett, D. Carbado, P. Casey, N. Dasgupta, D. Faigman, R. Godsil, A. G. Greenwald, J. Levinson and J. Mnookin. 2012. "Implicit Bias in the Courtroom." *UCLA Law Review 59*: pp. 1124-1186.

Kennedy, D. M. 2009. *Deterrence and Crime Prevention: Reconsidering the Prospect of Sanction*. London: Routledge.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig and S. Mullainathan. 2017. "Human Decisions and Machine Predictions." *NBER Working paper series*.

Kleinberg, J., J. Ludwig, S. Mullainathan and C. R. Sunstein. 2019. "Discrimination in the Age of Algorithms." *arXiv e-prints*.

Kleinberg, J., S. Mullainathan and M. Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *arXiv e-prints*.

Lepri, B., N. Oliver, E. Letouzé, A. Pentland and P. Vinck. 2018. "Fair, Transparent, and Accountable Algorithmic Decision-making Processes." *Philosophy & Technology 31* (4): pp. 611-627.

Lippert-Rasmussen, K. 2020. *Making Sense of Affirmative Action*. Oxford: Oxford University Press.

Lipton, Z. C., A. Chouldechova and J. McAuley. 2018. "Does mitigating ML's impact disparity require treatment disparity?" *32nd Conference on Neural Information Processing Systems*.

Liu, J. Z. and X. Li. 2019. "Legal Techniques for Rationalizing Biased Judicial Decisions: Evidence from Experiments with Real Judges." *Journal of Empirical Legal Studies 16* (3): pp. 630-670.

MacAskill, W., K. Bykvist and T. Ord. 2020. *Moral Uncertainty*. Oxford: Oxford University Press.

MacAskill, W. and T. Ord. 2020. "Why Maximize Expected Choice-Worthiness?" *Noûs 54* (2): pp. 327-353.

Macnish, K. 2012. "Unblinking eyes: the ethics of automating surveillance." *Ethics and Information Technology 14* (2): pp. 151-167.

Macnish, K. 2017. *The Ethics of Surveillance: An Introduction*. London: Routledge.

Macnish, K. 2018. "Government Surveillance and Why Defining Privacy Matters in a Post‐Snowden World." *Journal of Applied Philosophy 35* (2): pp. 417-432.

Mittelstadt, B. D., P. Allo, M. Taddeo, S. Wachter and L. Floridi. 2016. "The ethics of algorithms: Mapping the debate." *Big Data & Society 3* (2).

Molnar, C. 2019. *Interpretable machine learning. A guide for making black box models explainable*.

Moore, M. S. 1997. *Placing blame: a general theory of the criminal law*. Oxford: Oxford University Press.

MSI-NET. 2017. *Algorithms and Human Rights - Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*. Council of Europe.

Murphy, J. G. 1979. *Retribution, Justice, and Therapy*. Dordrecht: Reidel.

Parfit, D. 2011. *On What Matters*. Oxford: Oxford University Press.

Pettit, P. 1997. "Republican Theory and Criminal Punishment." *Utilitas 9* (1): pp. 59-79.

Pratt, J. 2007. *Penal populism*. London: Routledge.

Rachlinski, J. J., S. Johnson, A. J. Wistrich and C. Guthrie. 2009. "Does Unconscious Racial Bias Affect Trial Judges?" *Cornell Law Faculty Publications Paper 786*.

Riedener, S. 2019. "Constructivism about Intertheoretic Comparisons." *Utilitas 31* (3): pp. 277-290.

Riedener, S. 2020. "An axiomatic approach to axiological uncertainty." *Philosophical Studies 177*: pp. 483-504.

Roberts, J. V., L. J. Stalans, D. Indermaur and M. Hough. 2002. *Penal populism and public opinion: Lessons from five countries*. Oxford: Oxford University Press.

Robinson, P. H. and J. M. Darley. 2004. "Does Criminal Law Deter? A Behavioural Science Investigation." *Oxford Journal of Legal Studies 24* (2): pp. 173-205.

Roth, A. 2016. "Trial by Machine." *Georgetown Law Journal 104* (5): pp. 1245-1306.

Rudin, C. 2019. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence 1* (5): pp. 206-215.

Ryberg, J. 2007. "Privacy Rights, Crime Prevention, CCTV, and the Life of Mrs. Aremac." *Res Publica 13* (2): pp. 127-143.

Scanlon, T. M. 2014. *Being Realistic about Reasons.* Oxford: Oxford University Press.

Selbst, A. D. 2017. "A Mild Defense of Our New Machine Overlords." *Vanderbilt Law Review 70*: pp. 87-104.

Shafer-Landau, R. 2005. *Moral Realism: A Defence*. Oxford: Clarendon Press.

Shermer, L. O. N. and B. D. Johnson. 2010. "Criminal Prosecutions: Examining Prosecutorial Discretion and Charge Reductions in U.S. Federal District Courts." *Justice Quarterly 27* (3): pp. 394-430.

Simmons, R. 2018. "Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System." *University of California, Davis Law Review 52* (2): pp. 1068-1118.

Smith, R. J. and J. D. Levinson. 2012. "The Impact of Implicit Racial Bias on the Exercise of Prosecutorial Discretion." *Seattle University Law Review 35*: pp. 795-826.

Stobbs, N., M. Bagaric and D. Hunter 2017. "Can sentencing be enhanced by the use of artificial intelligence?" *Criminal Law Journal 41* (5): pp. 261-277.

Tadros, V. 2011. *The Ends of Harm: The Moral Foundations of Criminal Law*. Oxford: Oxford University Press.

Tarsney, C. 2018. "Intertheoretic Value Comparison: A Modest Proposal." *Journal of Moral Philosophy 15* (3): pp. 324-344.

Thomsen, F. K. 2011. "The Art of the Unseen - Three Challenges for Racial Profiling." *The Journal of Ethics 15* (1): pp. 89-117.

Thomsen, F. K. 2014. “Why Should We Care What the Public Thinks? A Critical Assessment of the Claims of Popular Punishment.” In *Popular Punishment*, edited by J. Ryberg and J. V. Roberts, pp. 119-145. Oxford: Oxford University Press.

Thomsen, F. K. 2018. "Good Night and Good Luck - In Search of A Neuroscience Challenge to Criminal Justice." *Utilitas 30* (1): pp. 1-31.

Thomsen, F. K. 2020. "The Teleological Account of Proportional Surveillance." *Res Publica 26*: pp. 373-401.

Tonry, M. 2016. "Making American Sentencing Just, Humane, and Effective." *Crime and Justice 46*: pp. 441-504.

Veale, M., M. Van Kleek and R. Binns. 2018. "Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*.

Von Hirsch, A. 1996. *Censure and Sanctions*. Oxford: Oxford University Press.

Von Hirsch, A. and A. Ashworth. 2005. *Proportionate Sentencing*. Oxford: Oxford University Press.

Von Hirsch, A., A. E. Bottoms, E. Burney and P.-O. Wikström. 1999. *Criminal Deterrence and Sentence Severity*. Cambridge: Hart Publishing.

Wagner, B. 2019. "Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems." *Policy & Internet 11* (1): pp. 104-122.

Walker, N. 1991. *Why Punish? Theories of punishment reassessed*. Oxford: Oxford University Press.

Wood, W. R. 2014. “Punitive Populism.” In *The Encyclopedia of Theoretical Criminology*, edited by J. M. Miller, pp. 678-682. Chichester: John Wiley & Sons.

Yui, T. 2019. "The Curse of Dimensionality." *Towards Data Science*.

Zerilli, J., A. Knott, J. Maclaurin and C. Gavaghan. 2019. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" *Philosophy & Technology 32*: pp. 661-682.

Zimmerman, M. J. 2011. *The Immorality of Punishment*. Toronto: Broadview Press.

[i] Given that we have theories of just sentencing, could we not apply them to distinguish between historical cases of just and unjust sentencing and then apply machine-learning to the relabelled set? We could, but doing so would be pointless since all we have done is apply the model that machine-learning was supposed to help us discover. Any model developed by machine learning on the basis of the resulting dataset would be at best identical to our theory of just sentencing, and realistically just a less accurate version of it.

[ii] For very insightful discussion of the normative claims of the transparency and bias objections, see respectively Ryberg and Lippert-Rasmussen in this volume.

[iii] Note, however, that the incentive for maximal inclusion of features is subject to the constraint imposed by the so-called "curse of dimensionality": the data-space grows exponentially with the number of features, while the ability to distinguish signal from noise grows only linearly with the number of training examples (Chen 2009, Yui 2019). Adding features therefore either requires also adding an ever-greater number of training examples or imposes a cost in the shape of an increased number of non-generalizable correlations.

[iv] Examples of recent influential defences of metaethical cognitivism and/or moral realism include Shafer-Landau 2005, Crisp 2006, Huemer 2007, Parfit 2011, Broome 2013, and Scanlon 2014.

[v] Cf. Donohue 2019 on the difficulties encountered by the committees of the 1986 US sentencing guidelines commission. For a more encouraging illustration of the feasibility of such a project, see Bagaric and Gopalan 2015.

[vi] See Roth 2016 for analysis of how the introduction of technology in criminal justice has in previous cases been tailored to serve a penal populist agenda of more frequent and more severe punishment.