Philosophy & Technology (2024) 37:9
https://doi.org/10.1007/s13347-024-00694-3

RESEARCH ARTICLE

# Criminal Justice and Artificial Intelligence: How Should we Assess the Performance of Sentencing Algorithms?

**Jesper Ryberg**[1]

Received: 19 April 2023 / Accepted: 30 December 2023 / Published online: 12 January 2024


## Abstract

Artificial intelligence is increasingly permeating many types of high-stake societal decision-making such as the work at the criminal courts. Various types of algorithmic tools have already been introduced into sentencing. This article concerns the use of algorithms designed to deliver sentence recommendations. More precisely, it is considered how one should determine whether one type of sentencing algorithm (e.g., a model based on machine learning) would be ethically preferable to another type of sentencing algorithm (e.g., a model based on old-fashioned programming). Whether the implementation of sentencing algorithms is ethically desirable obviously depends upon various questions. For instance, some of the traditional issues that have received considerable attention are algorithmic biases and lack of transparency. However, the purpose of this article is to direct attention to a further challenge that has not yet been considered in the discussion of sentencing algorithms. That is, even if is assumed that the traditional challenges concerning biases, transparency, and cost-efficiency have all been solved or proven insubstantial, there will be a further serious challenge associated with the comparison of sentencing algorithms; namely, that we do not yet possess an ethically plausible and applicable criterion for assessing how well sentencing algorithms are performing.

**Keywords** Algorithms · Crime · Criminal courts · Machine learning · Penal theory · Punishment · Sentencing

Artificial intelligence is becoming omnipresent. Various types of algorithmic tools are increasingly – indeed at a significant pace – permeating all domains of social life (see Ryberg and Roberts 2022a). This is also the case in contexts that involve what is characterized as high-stake societal decision-making; that is, decision-making that has a significant impact on the life and well-being of citizens. An obvious example of such a context is the criminal justice system.

✉ Jesper Ryberg
ryberg@ruc.dk

[1] Professor of Ethics and Philosophy of Law, Roskilde University, 4000 Roskilde, Denmark

Algorithmic tools are currently infiltrating all stages of criminal justice practice, from the work of the police to the final verdicts in court. For instance, risk assessment algorithms – such as the much-debated COMPAS algorithm – have for a long time been used in the US to inform the courts of the likelihood that offenders will fall back into crime. A more radical example is the use of algorithms designed to provide sentence recommendations in individual criminal cases. For instance, systems that determine sentences in cases of serious crime – such as rape and drug possession – have already been put into practice.[1] And some states have declared their intentions of introducing 'intelligent courts' based on the use of AI in judicial decision-making, including sentencing.[2] Thus, even though this is currently still only an aspiration, it may become reality in a not very distant future. But if it is the case that algorithms designed to determine sentences are about to find their way into the sentencing process in court, then one of the many questions one will be confronted with is what type of algorithms should be used in relation to such important decisions as the determination of sentences? For instance, would the use of machine-learning-based systems be desirable? This question has not yet been the subject of more comprehensive discussions. However, a few commentators have cursorily addressed the issue.

For instance, one of the concerns that have been expressed against the use of more complicated systems, such as machine learning algorithms, is that the predictive accuracy of these systems comes at a price: namely, a lack of interpretability of the inner workings of the systems. As observed by computer scientist David Cunning and his colleagues: 'There may be inherent conflict between ML [machine learning] performance (e.g. predictive accuracy) and explainability. Often, the highest performing methods (e.g. DL [deep learning]) are the least explainable, and the most explainable (e.g. decision trees) are the least accurate' (Gunning et al., 2019). A possible way of dealing with a lack of explainability of more complicated machine learning systems has been to draw on explainable artificial intelligence (xAI); that is, a second algorithm which is created to explain post hoc the workings of the black box system. However, this combination has been the subject of criticism. One of the strongest critics is Cynthia Rudin, who has argued that rather than relying on a combination of opaque algorithmic systems and xAI, what one should do is to use algorithmic tools that are 'inherently interpretable'; that is, systems that 'provide their own explanations, which are faithful to what the model actually computes' (Rudin, 2019). What she holds is that the current business model of machine learning often incentivizes firms to develop algorithms that are 'overly complicated' and that quite often 'organizations do not have analysts who have the training or expertise to construct interpretable models at all' (Rudin, 2019). On the grounds of the experience that she has 'not yet found a high-stakes application where a fully black box model is necessary', she ultimately suggests, as a general rule of thumb, that when it comes to high-stake societal applications, including the use of algorithmic tools in a criminal justice context, 'no black box should be deployed when there exists an

[1] For instance, this is the case in Malaysia. See Khazanah Research Institute 2021.

[2] This has recently been declared by the Chinese State Council. See Shi (2022).

interpretable model with the same level of performance' (Rudin, 2019). Thus, in her view machine-learning sentencing algorithms would often not be recommendable.[3]

Another theorist who has briefly commented on the issue of what may constitute the desirable type of algorithm in a sentencing context is Frej Thomsen. In Thomsen's view, there are several advantages of introducing algorithmic tools into sentencing. In fact, he even argues that when it comes to the determination of sentences, there are strong reasons in favour of implementing 'fully automated decision-making' (Thomsen, 2022, p. 252). Notably, however, what he also suggests is that sentencing is not a decision problem well suited to machine learning (p. 256). Part of the explanation is the much-discussed problem that if there exist biases in the historical dataset on which an algorithm is trained, these may be reproduced or even exacerbated in the output of the algorithm. This is sometimes colloquially known as the 'garbage in, garbage out' problem. Therefore, Thomsen concludes that automated sentencing decisions should not be based on machine learning, but 'on old-fashioned human programming' (Thomsen, 2022).

What these briefly sketched considerations on what constitutes the preferable type of algorithmic systems in a sentencing context have in common is that they direct attention to – and seek to circumvent – some of the traditional challenges that have attracted attention in the discussion of algorithmic tools in criminal justice practice (and in other societal contexts). The first issue concerns the lack of transparency of some algorithmic systems; the second, the fact that algorithmic predictions can be biased (see also Lippert-Rasmussen, 2022; Davies & Douglas, 2022; Ryberg & Roberts, 2022). However, if the focus is on what type of algorithms would be suitable for use in a sentencing context, then these considerations do not provide an exhaustive answer. There would of course be other types of challenges that would have to be considered.[4] Moreover, and more importantly, the basic comparative question would remain even if these challenges had all been solved. Suppose *arguendo* that challenges concerning algorithmic transparency or algorithmic biases (and other similar collateral implications) have either been proven ethically insubstantial or have been handled through some sort of technical solution (see also Ryberg, 2024a). In this case, we would still be confronted with the question of what would constitute the preferable algorithm if such tools were to be implemented at sentencing. Suppose, for instance, that different computer scientists have either offered a highly complicated deep learning algorithm, a less complicated machine learning algorithm, or an algorithm not based on machine learning, and, further, that these algorithms all seem to be doing a fine job when it comes to the determination of sentences in individual criminal cases, then which system should be regarded as preferable?[5]

[3] For a more comprehensive discussion of Rudin's arguments and, more generally, of the conflict between accuracy and explainability, see also Ryberg and Petersen 2022.

[4] For instance, another obvious question will be whether one algorithmic system is more cost-effective than another in the sense of reducing case-processing time and resources spent in the courts (Hunter et al., 2020). Yet another question concerns challenges with the regard to the collection of data on which algorithms should trained (see e.g., Schwarze and Roberts 2022).

[5] The reason for assuming that the competing algorithms are doing "a fine job" is of course this is what generates the challenge of comparing the performance of the algorithms. If a sentencing algorithm would

When the question is phrased in this manner, that is, by initially excluding all the standard types of ethical challenge that are usually considered in comparisons between different algorithm systems, then it may seem that the answer becomes obvious: one should choose the algorithm that is best at doing the job it is designed to do; that is *in casu* the one that excels when it comes to the determination of sentences in criminal cases. However, even though this answer may sound almost like a truism, what will be argued in the following is that the application of this answer is in fact highly complicated. The answer presupposes solutions to a set of penal theoretical challenges that have not yet been provided. The implication, therefore, is that we do not yet possess the theoretical background that allows us to make justified assessments of which type of algorithm would be preferable for sentencing purposes. At least, so it will be argued.[6]

In order to sustain this conclusion, the article will proceed as follows. Section (1) considers a first possible candidate for a criterion for the assessment of the relative performance of sentencing algorithms. The criterion is based on indistinguishability between sentences determined by algorithms and human judges. It is argued that this criterion, despite immediate plausibility, should be dismissed. In Sect. (2), another criterion for the assessment of the performance of algorithmic tools is considered. This criterion is based on penal ethical considerations. It is argued that even though the criterion is indeed ethically plausible, it presupposes answers to a range of penal theoretical challenges that have not yet been provided. Section (3) considers a final possible assessment criterion based on considerations of overpunishment. It is argued that even though this criterion may provide some guidance, it is likely to be confronted with the same ethical challenge as the criterion considered in Sect. (2). Section (4) summarizes and concludes.

What the discussion will underline is the fact that the justified implementation of algorithmic tools – like any other type of technology – is contingent on ethical considerations and, therefore, if the relevant ethical theories have not yet been sufficiently developed, it may not be possible to determine whether such tools constitute an improvement.

Footnote 5 (continued)
recommend a life sentence in all cases of theft, then it would be an easy job to dismiss it as a tool in the sentencing process.

[6] As noted, this article will only discuss algorithms designed to provide sentence recommendations (e.g., that a particular offender should have 6 months in prison), not algorithms designed to produce risk assessments. If it possible to show that an ethical assessment of risk assessments is contingent on the question of what constitute the ethically right sentences of offenders, then it might the case that the ensuing considerations would also be relevant for the discussion of the comparative performance of risk assessment algorithms. For arguments in support of this contention, see e.g. Ryberg 2020a and Ryberg and Thomsen 2022. However, within the scope of the present article, there will not be space to consider such an extrapolation of the arguments.

## 1 The Indistinguishability Criterion

Suppose that two different sentencing algorithms have been developed.[7] For instance, these could be a system based on machine learning and one that does not involve machine learning. For reasons of ease in exposition, let us call the competing systems α and β. In the comparison between these systems, what does it mean to say α is performing better β (or vice versa)? A first possible criterion for the comparative assessment that might come to mind would be to hold that we should prefer the algorithm which comes closest to determining the sentences that would have been determined by human judges. That is to say, more precisely, that assessment should be based on:

> *The indistinguishability criterion*: α is preferable to β if and only if the sentences determined by α to a larger extent than those determined by β equal those that would have been determined by human judges.

This criterion has some immediate appeal. Particularly if we are considering the use of algorithms considered by Thomsen and others, which involves the replacement of human judges in favour of fully automated sentencing decisions, it seems plausible to suggest that the question of whether an algorithm will be able to do the job usually carried out by judges must constitute a proper yardstick for the assessment of an algorithm. Therefore, the comparison of the merits of competing algorithmic systems must also depend upon the system's ability to mirror the decisions of human judges. Obviously, a criterion that makes indistinguishability from decisions by judges the parameter for the comparison of algorithmic merits needs some clarification in order to work in practice. For instance, judges do not always mete out the same sentences in similar cases. Many studies have established the existence of sentence disparity within the same jurisdictions.[8] However, even if we disregard this sort of challenge, there is a more basic ethical reason to be sceptical with regard to an indistinguishability-based criterion for the comparative assessment of algorithm systems.

The main problem is that, on closer inspection, it is difficult to see why the sentences determined by human judges should constitute a plausible parameter for the assessment. The criterion implies that the best possible algorithmic system would be the one that achieves complete indistinguishability from sentences determined by judges. For instance, if human judges would give two years in prison for a burglary, four years for an assault, and six years for a rape, then the best we can hope for is that an algorithm would reach the very same sentences. However, in many other contexts in which algorithmic tools are implemented, the ambitions are much higher. For instance, when machine learning algorithms are used to analyse scan images in

[7] The discussion of the criteria in this and the following sections draws on thoughts that have been presented in relation to a discussion of when algorithms are performing better than humans, and of when it would be justified to replace human judges with algorithms in sentencing, see Ryberg 2024a.

[8] For a brief overview of some of the studies that have been conducted on disparity in sentencing in the US and in Europe, see e.g. Ryberg 2023.

medical contexts, the idea is not only that these tools should perform in the same way as human radiologists. Rather, the goal is that they will be able to out-perform radiologists by producing assessments that are correct in cases where radiologists fail. The same is the case in many other contexts in which algorithmic instruments are currently taking over the work of humans. Algorithms are implemented not only to maintain levels of human decision-making, but also to improve them. Could there be room for the same ambition if we are considering the introduction of algorithms into sentencing?

The answer must be in the affirmative and it does not require much reflection to sustain this contention. For instance, there are strong reasons to believe that the decisions made by human judges may sometimes be biased. Many studies have established that black and brown people are treated more harshly in the US and in other countries (von Hirsch and Roberts, 1997; Clair & Winter, 2017; Veiga et al., 2023). Moreover, there are many studies that have supported the conclusion that sentencing decisions may also be skewed in other ways than by racial biases.[9] Therefore, suppose that we have a case in which a human judge would give eight months in prison for a certain crime, but that the unbiased sentence would have been six months. In this case, it is very hard to maintain the view that indistinguishability from human decisions should constitute the most plausible criterion for the assessment of algorithms.[10] Rather, what we should ideally want from a sentencing algorithm is that it recommends a six-month prison term. In fact, the hope that algorithmic tools may help us to avoid some of the biases of human decision-making has been used as an argument in favour of replacing human judges with fully automated decision-making (see Thomsen, 2022). In other words, when it comes to sentencing algorithms, it also seems plausible to believe that there is room for out-performing human judgements. Human sentencing decisions cannot be assumed to always be ethically perfect (in fact, as we shall see shortly, there are several other reasons beyond those illustrated in the bias example underpinning this contention). Therefore, indistinguishability from human sentencing decisions cannot constitute a necessary condition for the comparative assessment of different types of sentencing algorithms.

## 2 The Penal Ethical Criterion

If the previous considerations are true, that is, if it is correct that the sentences reached by human judges do not constitute a plausible parameter for the comparative assessment of different types of sentencing algorithms, then what does it mean to hold that one sentencing algorithm is performing better than another? On the basis of the previous considerations, there is a criterion for the comparative assessment that seems straightforward. Given the challenge that there may be some sentences that would be ethically preferable even if they deviate from those that would have

[9] For instance, studies have been conducted on how legal decisions may be affected by anchor effects (Englich et al., 2006), hindsight biases (Harley 2007), and perspective effects (Lassiter et al., 2009).

[10] For a parallel discussion concerning replacement of human judges with algorithms, see Ryberg 2024a.

been determined by human judges, an obvious possibility would be to base a criterion directly on what constitutes an ethical approach to the imposition of sentences on offenders. More precisely, what might be suggested is:

> *The penal ethical criterion*: α is preferable to β if and only if α to a larger extent than β succeeds in determining sentences that accord with our best ethical theory of punishment.

This criterion for comparison is clearly more plausible than the indistinguishability criterion. While the possibility of ethically preferable deviations from the sentences determined by judges showed that indistinguishability could not constitute a necessary condition for the assessment of algorithms, the penal ethical criterion is not vulnerable to the same objection. All that matters, according to this criterion, is what would be ethically desirable, not what sentences would have been determined by judges. It might perhaps therefore also be said that the criterion comes close to a truism: how could a sentencing algorithm fail to be preferable, if it determines sentences that are ethically better than those determined by another algorithm? However, even though this criterion is indeed hard to dispute as long as we have initially excluded all types of collateral consequences that might be part of an all-things-considered assessment of sentencing algorithms, this does not imply that it is devoid of challenges. The problem facing this criterion is that it is very hard to apply in practice. There are, as we shall now see, several reasons why this is the case.

The first problem is that it is far from clear what should be considered the best ethical theory of punishment. The story of how the ethics of punishment has developed has often been told. During much of the nineteenth and twentieth centuries, the utilitarian approach to punishment dominated the philosophical discussion. The retributivist approach was often regarded as an inhumane or even barbarous position far distant from what could possibly be seen as an enlightened approach to the issue. However, in the early 1970s the picture started to change. An increasing number of penal theorists declared their approval of retributivist thinking and references to the revival or renaissance of retributivism became part of the standard refrain in titles and opening lines of works on penal theory (Duff and Garland, 1994; Ryberg, 2004). While retributivism dominated the field in the subsequent decades, the picture today has become much more diverse. There are still many theorists who defend retributivism as the most plausible theory of punishment. However, it is also a well-known fact that retributivism does not denote a single theory of punishment. It is more apposite to regard retributivism as an umbrella concept comprising a range of theories which, even though they share the view that desert plays a crucial role in the justification of punishment, have nevertheless been developed in many different directions. Moreover, many theorists also advocated mixed theories which seek to combine elements of both consequentialist and retributivist thinking. These theories also exist in many different versions. Furthermore, a range of other theories that do not readily fit the standard distinction between retributivism and utilitarianism have found their way into the modern discussion. These theories include: consequentialist (non-utilitarian) theories, right-forfeiture theories, self-defence theories,

restitutionist theories, restorative justice approaches, and different accounts of abolitionism.[11] Thus, it is fair to say that there exists much disagreement regarding what constitutes the best theory of punishment. In fact, the field seems more diverse than ever and, importantly, there is no reason to believe that the different theories would have the same implications when it comes to penal distribution. Therefore, the comparison of sentencing algorithms against the backdrop of the sentences that would follow from the 'best ethical theory of punishment' is not an easy task.

Second, even if we disregard existing theoretical disagreements between different theories of punishment and focus instead, more narrowly, on what some of the most influential theories have to offer with regard to punishment distribution, it turns out that the answer is very poor. For instance, from a utilitarian point of view it is not easy to tell precisely how severely different offenders should be punished. Obviously, the consequences of the imposition of punishment on offenders will vary to some extent with the social context. Thus, from such a view one should not expect any universal answers. However, within different contexts it is often empirically underdetermined whether a particular offender should have, say, six or seven months in prison. From a retributivist perspective the problem is somewhat different, namely, that desert-based theories have yet had very little to say about what constitute the appropriate penal levels. Much has been written on the retribution-based idea of ordinal proportionality.[12] However, this principle only says something about the *relative* punishments for crimes of varying degrees of seriousness. It does not tell how severely a particular crime should be punished. A few retributivists have taken up the challenge of trying to develop theories designed to determine the appropriate punishment for different crimes (see e.g. von Hirsch, 1993; Scheid, 1997; Lippke, 2012). Though I cannot here enter a thorough discussion of these theories, it is fair to hold that they have so far only provided the contours of very general frameworks for how questions of penal distribution should be approached.[13] No theories have provided precise answers to how severely different crimes should be sentenced. Therefore, when it comes to the question of what constitutes the ethically right sentence for a particular crime, current consequentialist and retributivist theories have not yet provided answers which the penal theoretical criterion can draw on in the comparison of the merits of sentencing algorithms. The same is the case with regard to other penal theories in the field.

What these considerations show is of course not that it is impossible to apply the penal theoretical criterion in the assessment of sentencing algorithms. It could be the case that penal theory will in the future be developed in a way that suffices to provide answers to the detailed questions of penal distribution which the application of the criterion requires. However, it is fair to say that we are currently very far from possessing the theoretical (and empirical) resources which the application of

[11] For an overview and discussion of many of these theories, see e.g. Ryberg 2024b.

[12] See, for instance, von Hirsch 1993; von Hirsch and Ashworth 2005; Ryberg 2004; Tonry 2020.

[13] Furthermore, it should be mentioned that the few theories that have attempted to provide a theoretical framework for providing answers to how severely different crimes should be punished, have been subject of massive criticism. See, for instance, Ryberg 2004 and 2020b; Tonry 2020.

the penal ethical criterion presupposes. Thus, even though the criterion is indeed plausible in principle, it is still of little use in actual cases that require the comparison of the merits of competing sentencing algorithms.

## 3 The Over-punishment Criterion

Even though it is a fact that theories of punishment have so far had very little to offer with regard to the severity of the sentences that should be imposed on different offenders, it might perhaps still be felt that the previous considerations are somewhat premature. It may be a fact that current penal theory is unable to prescribe whether a thief should spend four or five months in prison, or whether a drunk driver should pay a fine of 500 or 600 dollars. However, this does not necessarily imply that penal theories are devoid of the possibility of giving any sort of direction with regard to what constitute the appropriate penal levels. In fact, there seems to be a remarkable agreement amongst penal theorists on the fact that many offenders, not only in the US but also in many other countries, are currently being punished much too severely.

One of the reasons that have been given in support of this view is that there is a problem of over-criminalization. That is, that there are simply too many ways of acting that should not have been criminalized in the first place. For instance, Douglas Husak, who has comprehensively considered the issue, characterizes over-criminalization as 'the most pressing problem with the criminal law today' (Husak, 2008, p. 3). But if there are too many laws on the books, then it follows that there are citizens who are being over-punished. They are being punished for acts that do not warrant any punishment in the first place. In this way, over-criminalization produces over-punishment.

The second reason is that even if we are only considering acts on which it is generally agreed that they should be legally prohibited, there are numerous theorists who hold that the criminal sanctions are much too harsh. In particular, many theorists contend that incarceration is being massively over-used. For instance, this point has been made in relation to penal theories that draw on consequentialist considerations. Numerous studies on incapacitation and deterrence have shown that there is no gain in terms of crime prevention by locking offenders up for longer periods and, consequently, that periods of incarceration should be reduced. Michael Tonry has recently summarized his review of studies on crime prevention by contending that: 'it is not controversial to assert that the crime prevention effects of mass incarceration have been much less than many people supposed or hoped. That there is little or no reason to believe that harsher punishments have greater deterrent effects than milder punishments, that incapacitating people by locking them up for lengthy periods is an ineffective crime prevention strategy, or that the experience of imprisonment makes many offenders more not less likely to commit crime later in their lives' and furthermore that 'the implications of the literatures on deterrence and incapacitation are straightforward: few convicted offenders should be sent to prison and for shorter times' (Tonry, 2016, pp. 453 and 459). Similar support for the over-punishment theses can be found among theorists belonging to the retributivist camp. For

instance, Richard Singer contends that it is a mistake to think of the desert model as a derivation of a 'throw away the key' approach to punishment. He suggests that incarceration should only be reserved for the most serious crimes, and even then, the duration should be relatively short (Singer, 1979, p. 44). Along the same lines, Jeffrie Murphy holds that if desert theory were to be followed consistently, one would punish less and in more decent ways than one actually does (Murphy, 1979, p. 230). And Saul Smilansky underlines that there is 'evidence indicating that the overpunishment of guilty people – punishing them more than they morally deserve for the crime they are convicted of – is a widely prevalent practice in many western countries' (Smilansky, 2021, p. 1) In fact, Andreas von Hirsch has even argued that terms of imprisonment even for the most serious crimes should seldom exceed five years – a view of penal levels that deviates radically from current practice in all Western countries (von Hirsch, 1993).

Therefore, suppose that, despite the fact that penal theories have not been able to determine precisely how severely different crimes should be punished, it is nevertheless the case that there are theoretical grounds for the judgement that offenders are currently being over-punished. In that case, a modified version of the penal theoretical criterion for comparative assessment of sentencing algorithms might be the following:

> *The over-punishment criterion*: α is preferable to β if and only if α determines sentences that are more lenient than those determined by β.

It is interesting to note that considerations based on this sort of criterion have been used to reject the possibility of applying sentencing algorithms based on machine learning. As we have seen, an algorithm that draws on historical data might reproduce patterns of discrimination. However, a similar reproduction problem will of course exist when injustices arise from the fact that sentences are too harsh. This point is clearly made by Thomsen who contends that: '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 [automated decision-making] 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' (Thomsen, 2022, p. 257). This observation seems true. However, beyond directing attention to a challenge associated with the use of machine learning, can the suggested criterion serve as a helpful guide when it comes to the overall question of how the relative merits of sentencing algorithms should be assessed?

There is no doubt that this criterion could, in some cases, serve as a guide for the comparison of some sentencing algorithms. For instance, if α tends to recommend sentences slightly more severe than the current penal level, while β determines sentences that are more lenient than the current level, then following the criterion β would be preferable to α. However, in many real-life cases the picture may be more complicated. Suppose that some of the sentences determined by α are more severe than those determined by β, while in other cases involving other types of crime, the sentences determined by β are more severe than those determined by α. In this case, the application of the criterion easily becomes complicated. From a consequentialist

perspective, it may well be empirically underdetermined which algorithmic system will be ethically preferable. And seen from a retributivist perspective, one will be faced with the highly complicated and theoretically under-explored problem of how one should ethically compare different deviations from the ethically right punishment.[14] It is in the connection worth noting that in real-life penal practice, it is to be expected that this problem will arise, because it is quite unlikely that different algorithms that are considered for implementation will provide radically different sentencing recommendations. For instance, a comparison between an algorithm that recommends sentences that are close to those hitherto determined by judges, and another algorithm which reflects von Hirsch's contention that even the most serious crimes should not be punished with more than five years in prison, would hardly ever occur in real life. The recommendations of the latter algorithm would constitute a revolution of penal practice which very few decision-makers would be willing to accept.[15] Thus, in real life it is much more likely that the algorithmic tools that will be considered in relation to sentencing will reflect the existing penal order. But it is precisely when algorithms do not differ much in the recommendations they provide that the suggested problems in determining which set of sentence recommendation is preferable will arise.

In summary, the *over-punishment criterion* may seem like a plausible attempt to account for the fact that current penal theory has not yet been able to provide precise answers to how severely different offenders should be punished. And the criterion may in some cases be able to provide guidance on which algorithms would be preferable. However, as we have just seen, it is also reasonable to expect that in real-life comparison between competing algorithmic systems, the sentences that each system recommends will not deviate significantly, and it is precisely when this is the case that the criterion becomes vulnerable to the underexamined and indeed theoretically demanding problem of ethically comparing varying patterns of deviation from what constitute the ideal penal levels.

[14] In order to deal with this question, one would have to engage in considerations of how one should, from a retributivist perspective, compare sets of sentences with differing patterns of deviations from the ethically right sentences. To take a very simple example: suppose that α recommends that offender A gets 3 months, that offender B gets 6 months, and that offenders C gets 8 months, while β recommends that A gets 4 months, that B gets 3 months, and that C gets 9 months. Suppose further that all these sentences would constitute instances of over-punishment. In that case, which of the two algorithms would be ethically preferable? To my knowledge, modern retributivists have not yet provided any sort of guidance on how such comparisons should be made (see also Ryberg 2014).

[15] The problem is captured in what I have elsewhere called *the dilemma of AI-based sentencing*: namely, that a machine-learning sentencing algorithm may work on the basis either of a dataset that is built on actual sentencing decisions of the judiciary within a jurisdiction – in this case, the system may be practically workable but it is not clear that this would result in an ethical improvement of sentencing practice – or of a database built on hypothetical assessments of what would have constituted the ethically right sentences in various cases. In the latter case, the system would in principle provide an ethically proper guideline but would hardly be politically workable under real-life circumstances where there is a significant gap between penal theory and penal practice (see Ryberg 2023).

## 4 Conclusion

The use of algorithmic tools to determine sentences has already become a reality in some jurisdictions and given the explicit intentions of drawing on artificial intelligence in criminal justice practice it is certainly highly likely that the use of such technological tools will increase in the near future. Moreover, there is also a growing number of researchers who, despite acknowledging that there are ethical challenges that should be addressed, defend the implementation of such tools in the work of the courts (see Chiao, 2018). Thus, it seems realistic to expect that we will in the future be confronted with situations where one will have to compare which algorithmic systems should be regarded as preferable. It is this question that has been considered in the previous discussion. More precisely, the focus has been on the question of what it means to hold that one sentencing algorithm is performing better than another. By disregarding all the traditional challenges associated with the use of algorithmic systems – which would of course have to be included in an all-things-considered assessment – the purpose has been to give content to the truism that one should opt for the sentencing algorithm that performs the best.

What we have seen is that a criterion that turns the sentences determined by human judges into the parameter for the comparative assessment of sentencing algorithms should be dismissed. No penal ethicists would hold that all sentences that are determined by judges in court are ethically right. But this means that a plausible criterion will have to be based not on actual sentences determined in court, but on penal ethical considerations of what constitute the ethically right sentencing levels. However, as argued, an approach along these lines turned out to have two important implications.

First, if the comparative assessment of the merits of competing algorithmic systems is not based on actual sentences but on what would constitute the ethically right sentences for different crimes, then it makes little sense to make algorithmic sentencing recommendations contingent on previous sentencing decisions.[16] This means that there will be no point in using machine-learning-based algorithms which are trained on and, therefore, reproduce historical sentences. Second, and more importantly, if a plausible criterion for the assessment of algorithmic performance must be based on penal ethical considerations, then this obviously presupposes that it is possible to get sufficiently precise answers to how severely different crimes ought to be punished. In the absence of such answers, the criterion would remain empty. However, as argued, current penal theory has so far had very little to offer when it comes to the question of the distribution of punishment. Theories of punishment distribution are either empirically underdetermined or theoretically underdeveloped and, therefore, incapable of providing the requisite answers. Thus, it seems

[16] It might perhaps be suggested that historical sentences are ethically important because it is crucial to maintain consistency in the form of ordinal proportionality over time. However, as suggested, if it is the case that offenders are currently being over-punished, then this argument presupposes that ordinal proportionality has primacy over considerations of what constitute the right penal levels non-relatively speaking. As argued elsewhere, this is hardly a plausible position. For instance, it is not plausible to go on over-punishing offenders simply because previous offenders have been over-punished (see Ryberg 2023; and Duus-Ötterstroöm 2020).

fair to conclude that we do not currently possess the penal theoretical resources that are necessary to determine whether one algorithmic system performs better than another (at least not in realistic scenarios where the algorithmic recommendations do not differ significantly). It should certainly be hoped that future penal theories will be able to fill the gaps in current discussions of punishment distribution. This will of course be important because the question of how severely the state should punish offenders in itself constitutes an urgent ethical challenge. But, as we have seen, it will also be crucial in order to be able to prepare the ground for a principled implementation of algorithmic sentencing tools in the criminal courts.

**Authors Contributions** Not applicable (there is only one author of this submission).

**Funding** Open access funding provided by Roskilde University The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

**Data Availability** Not applicable (this is a purely philosophical study).

## Declarations

**Ethics and Approval to Participate** This is a philosophical study that does not involve human or animals subjects, or any sort of data management.

**Consent for Participate and Publish** Not applicable.

**Competing Interests** The author has no competing interests to declare that have relevance to the content of this article.



## References

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

Clair, M., & Winter, A. S. (2017). How Judges Can Reduce Racial Disparities in the Criminal-Justice System. *Court Review: The Journal of the American Judges Association, 598*, 158–160.

Davies, B., & Douglas, T. (2022). Learning to discriminate: The perfect proxy problem in artificially intelligent sentencing. In J. Ryberg & J. V. Roberts (Eds.), *Sentencing and Artificial Intelligence* (pp. 97–120). Oxford University Press.

Duff, A., & Garland, G. (Eds.). (1994). *A Reader on Punishment*. Oxford University Press.

Duus-Otterström, G. (2020). Weighing Relative and Absolute Proportionality in Punishment. In M. Tonry (Ed.), *Of One-Eyed and Toothless Miscreants* (pp. 30–50). Oxford University Press.

Englich, B., et al. (2006). Playing Dice with Criminal Sentences. *Personality and Social Psychology Bulletin, 32*, 188–200.

Gunning, D., et al. (2019). XAI - Explainable artificial intelligence. *Science Robotics, 4*, 1–2.

Harley, E. M. (2007). Hindsight Bias in Legal Decision Making. *Social Cognition, 25*, 48–63.
Hunter, D., et al. (2020). A Framework for the Efficient and Ethical Use of Artificial Intelligence in the Criminal Justice System. *Florida University State Law Review, 47*, 749–800.
Husak, D. (2008). *Overcriminalization: The Limits of the Criminal Law*. Oxford University Press.
Khazanah Research Institute. (2021). *#NetworkedNation: Navigating Challenges, Realising Oppurtunities of Digital Transformation*. Kula Lumpur: Khazanah Research Institute.
Lassiter, G. D., et al. (2009). Evidence of the Camera Perspective Bias in Authentic Videotaped Interrogations: Implications for Emerging Reform in the Criminal Justice System. *Legal and Criminological Psychology, 14*, 157–170.
Lippert-Rasmussen, K. (2022). Algorithmic-based sentencing and discrimination. In J. Ryberg & J. V. Roberts (Eds.), *Sentencing and Artificial Intelligence* (pp. 74–96). Oxford University Press.
Lippke, R. (2012). Anchoring the Sentencing Scale: A Modest Proposal. *Theoretical Criminology, 16*, 463–480.
Murphy, J. G. (1979). *Retribution, Justice, and Therapy*. Kluwer Academic Publishers.
Rudin, C. (2019). Stop Explaining Black Box Machine Learning for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence, 1*, 296–215.
Ryberg, J. (2004). *The Ethics of Proportionate Punishment: A Critical Investigation*. Kluwer Academic Publishers.
Ryberg, J. (2014). When Should Neuroimaging be Applied in the Criminal Court? *The Journal of Ethics, 18*, 81–99.
Ryberg, J. (2020a). Risk Assessment and Algorithmic Accuracy. *Ethical Theory and Moral Practice, 23*, 251–263.
Ryberg, J. (2020b). Proportionality and the Seriousness of Crimes. In M. Tonry (Ed.), *Of One-Eyed and Toothless Miscreants* (pp. 51–75). Oxford University Press.
Ryberg, J. (2022). Sentencing and Algorithmic Transparency. In J. Ryberg & J. V. Roberts (Eds.), *Sentencing and Artificial Intelligence* (pp. 13–33). Oxford University Press.
Ryberg, J., & Roberts, J. V. (Eds.). (2022). *Sentencing and Artificial Intelligence*. Oxford University Press.
Ryberg, J. (2023). Sentencing Disparity and Artificial Intelligence. *The Journal of Value Inquiry, 57*, 447–462.
Ryberg, J., & Petersen, T. S. (2022). Sentencing and the Conflict between Algorithmic Accuracy and Transparency. In J. Ryberg & J. V. Roberts (Eds.), *Sentencing and Artificial Intelligence* (pp. 57–73). Oxford University Press.
Ryberg, J., & Roberts, J. V. (2022). Sentencing and Artificial Intelligence: Setting the Stage. In J. Ryberg & J. V. Roberts (Eds.), *Sentencing and Artificial Intelligence* (pp. 1–13). Oxford University Press.
Ryberg, J. (2024a). Punishment and Artificial Intelligence. In J. Ryberg (Ed.), *The Oxford Handbook of the Philosophy of Punishment*. Oxford University Press, (forthcoming).
Ryberg, J. (2024b). *The Oxford Handbook of the Philosophy of Punishment*. Oxford University Press, (forthcoming).
Scheid, D. E. (1997). Constructing a Theory of Punishment, Desert, and the Distribution of Punishments. *The Canadian Journal of Law and Jurisprudence, 10*, 441–506.
Schwarze, M., & Roberts, J. V. (2022). Reconciling Artificial and Human Intelligence: Supplementing Not Supplanting the Sentencing Judge. In J. Ryberg & J. V. Roberts (Eds.), *Sentencing and Artificial Intelligence* (pp. 206–229). Oxford University Press.
Shi, J. (2022). Artificial intelligence, algorithms and sentencing in Chinese Criminal Justice: Problems and Solutions. *Criminal Law Forum, 33*, 121–148.
Singer, R. G. (1979). *Just Deserts*. Ballenger Publishing Company.
Smilansky, S. (2021). Overpunishment and the Punishment of the Innocent. *Analytic Philosophy, 63*, 232–244.
Thomsen, F. K. (2022). Iudicium ex Machinae: The Ethical Challenges of Automated Decision-making at Sentencing. In J. Ryberg & J. V. Roberts (Eds.), *Sentencing and Artificial Intelligence* (pp. 254–278). Oxford University Press.
Tonry, M. (2016). Making American Sentencing Just, Humane, and Effective. *Crime and Justice, 46*(1), 441–504.
Tonry, M. (Ed.). (2020). *Of One-Eyed and Toothless Miscreants*. Oxford University Press.
Veiga, A., et al. (2023) Racial and ethnic disparities in sentencing: What Do we Know, and Where Should We Go?, *The Howard* Journal *of Crime and Justice*, 2, 167-182.
von Hirsch, A. (1993). *Censure and Sanctions*. Clarendon Press.
von Hirsch, A., & Roberts, J. V. (1997). Racial Disparity in Sentencing: Reflections on the Hood Study. *The Howard Journal, 36*, 227–236.

Von Hirsch, A., & Ashworth, A. (2005). *Proportionate Sentencing*. Oxford University Press.