AI and Ethics
https://doi.org/10.1007/s43681-024-00442-5

ORIGINAL RESEARCH

Artificial intelligence at sentencing: when do algorithms perform well enough to replace humans?

Jesper Ryberg[1]

Received: 21 December 2023 / Accepted: 10 February 2024

Abstract

Artificial intelligence is currently supplanting the work of humans in many societal contexts. The purpose of this article is to consider the question of when algorithmic tools should be regarded as performing sufficiently well to replace human judgements and decision-making at sentencing. More precisely, the question as to which are the ethically plausible criteria for the comparative performance assessments of algorithms and humans is considered with regard to both risk assessment algorithms that are designed to provide predictions of recidivism and sentencing algorithms designed to determine sentences in individual criminal cases. It is argued, first, that the prima facie most obvious assessment criteria do not stand up to ethical scrutiny. Second, that ethically plausible criteria presuppose ethical theory on penal distribution which currently has not been sufficiently developed. And third, that the current lack of assessment criteria has comprehensive implications regarding when algorithmic tools should be implemented in criminal justice practice.

Keywords Algorithms · Artificial intelligence · Assessment criteria · Criminal justice · Sentencing

The use of algorithmic tools in the work of criminal courts constitutes a subject that is currently attracting increasing philosophical attention. This is not surprising. While some algorithmic instruments, such as risk assessment tools, have already been used as part of sentencing decisions for some time, the development of more advanced tools has opened up the possibility of a more comprehensive application. For instance, there have already been some initial attempts at using algorithms to provide sentence recommendations for serious types of crime, such as rape and drug possession.[1] And it is not unlikely that such an application concerning key decisions in the courts will become more widespread in the near future.[2]

At the same time, it has also become increasingly clear that the use of algorithms in sentencing gives rise to various sorts of ethical concern. It is generally recognized—and indeed this is the main impetus behind the ethics of punishment as a field of research—that state-imposed punishment on citizens is something that requires very careful consideration. Moreover, when it comes to the use of artificial intelligence in the criminal courts, several ethical problems have been identified. Thus, the fact that the use of artificial intelligence at sentencing is both practically pertinent and theoretically challenging makes it an obvious subject for ethical scrutiny. The purpose of this paper is to contribute to this discussion by directing attention to a challenge which, on the one hand, lies at the heart of the question of justified use of algorithms at sentencing but which, on the other, has so far received surprisingly little academic attention. More precisely, the question that will be considered is when (if at all) do algorithms used for sentencing purposes perform well enough to replace human decision-making in the criminal courts?

Suppose that computer scientists and engineers have been working hard to develop an algorithm that can be used to provide answers to questions that arise in a sentencing context. Suppose, further, that it turns out that the algorithm is apparently doing a very good job. At some point the question may then arise as to when the algorithm is performing sufficiently well to supplant human decision-making. When it comes to the use of algorithms as risk assessment tools,

✉ Jesper Ryberg
ryberg@ruc.dk

[1] Roskilde University, 4000 Roskilde, Denmark

[1] For instance, this is the case in Malaysia. See [1].

[2] The Chinese State Council has recently declared its intention to implement "intelligent courts" based on the use of AI for judicial decision-making, including sentencing. See Shi [2].

Published online: 29 February 2024

this point has already been passed. As noted, risk assessment algorithms—such as the COMPAS algorithm—have already been used in the US for more than two decades to determine the risk profile of offenders and, thereby, to carry out the work that was previously done by psychologists and medical doctors. When it comes to algorithms that are designed to determine the appropriate sentences for various crimes, the point has not yet been reached but, as indicated, it may not be far ahead. Thus, the question that arises is which are the appropriate criteria for assessment if one is comparing the performance of humans and algorithms? As will be argued in the following, this question turns out to be highly complicated.

In order to show what sort of complications are at stake, the paper will proceed as follows. In Sect. 1, assessment criteria will be considered with regard to the performance of risk assessment algorithms. The question that will be discussed is when do such tools do a better job than humans in terms of the task they are designed to handle? In Sect. 2, the same question will be considered regarding sentencing algorithms. That is, when can such algorithms be considered to perform better than human judges when it comes to the determination of sentences? In both sections it is argued that the most obvious candidates for assessment criteria are implausible and that criteria which are ethically more plausible make it very hard, in theory and practice, to undertake the requisite assessments. The purpose will not be to provide a defence of particular criteria, but rather to reveal the complexity of the task of making comparative assessments of the work performed by algorithms and by humans. Section 3 then discusses the practical implications of the previous considerations. It is shown that the lack of plausible and applicable assessment criteria does not only have implications for the discussion on when it is justified to replace humans with algorithms in the work of criminal courts, but also with regard to the possibility of evaluating the work of such tools once they have been put into practice. Moreover, and more importantly, it is also argued that the lack of such criteria will have consequences beyond the question of when it is justified to replace humans with algorithms. It might be held that one of the reasons that the question of what constitutes plausible criteria for the comparative assessment of humans and algorithms has hardly been addressed is that, at least when it comes to the determination of sentences, this task should never be left entirely to the work of algorithms. Human judges should always be involved.[3] However, what will be argued is that the problem of the lack of plausible assessment criteria will have implications even if one is no longer considering the replacement of humans with algorithms but only the less revisionary scenario of using the recommendations of algorithmic tools to inform judges' decisions. Section 4 consists of the summary and conclusion.

However, before embarking upon these considerations, a final comment concerning the scope of the ensuing discussion is required. The question of when an algorithm performs better than humans may obviously give rise to many different considerations; for instance, some of the questions that have been at the centre of current discussions concern algorithmic transparency, that is, the issue of what sort of insight should be required into the inner workings of algorithms that provide court assistance (see e.g. [5–8]). Another important issue is the fact that algorithmic recommendations may be biased and lead to discriminatory decisions which, in a court context, may of course have very serious implications (see e.g. [9, 10]). In addition to these questions, there are also practical challenges regarding what sort of dataset should be used to train algorithms [3]. Moreover, it is also important to consider whether the implementation of algorithms will be cost-effective in the sense of reducing case-processing time and resources spent in the courts [11]. However, in what follows, these questions—which are of course all important in an all-things-considered comparative assessment—will be left out of consideration. The focus will be solely on the question of how well algorithms perform with regard to the work they are designed to carry out: that is, either the delivery of risk assessments or the determination of sentences. Thus, for the sake of the discussion we can imagine that all the other challenges have either been proven insubstantial or have been solved. Obviously, this way of bracketing a whole range of ethical and practical challenges has implications for the overall conclusion that can be drawn on the grounds of the ensuing considerations, we will return to this later. However, as we will see, this does not make the question of the comparative assessment of the performative merits of algorithms and humans any less important.

1 Risk assessment: algorithms vs humans

The use of risk assessment tools in the criminal courts has given rise to various discussions.[4] For instance, an important question has been when risk predictions are sufficiently accurate to be used at all (see [13]). Various studies have shown that risk predictions carried out by humans are often far from accurate (see e.g. [14, 15]). Similarly, meta-studies of algorithmic risk assessments have found only moderate levels of predictive accuracy [16, 17]. Other noteworthy

[3] For a defence of this view, see for instance Schwarze and Roberts [3] or Wingerden and Plesnicar [4].

[4] For a more broader disuccsion of various types of risk assessment tools used in criminal justice contexts, see [12], chapter 5).

studies have questioned the quality of risk assessment algorithms that are currently in use by showing that the predictions made by these tools are no more accurate than the predictions most of us would be able to make. For instance, in a recent paper published in Science Advances, Dressel and Farid have shown that the widely used COMPAS algorithm is no more accurate than predictions made by people with little or no criminal justice expertise. Moreover, they found that even though the COMPAS algorithm incorporates 137 distinct features in order to predict recidivism, it is possible to yield the same level of predictive accuracy on the grounds of only two features: age and total number of previous convictions [16]. However, despite such interesting results and discussions, the question of when an algorithmic risk assessment tool should be regarded as performing better or worse than humans has hardly been addressed. Thus, what would a plausible criterion for such comparative assessments look like?

Perhaps the main reason that this question has not been thoroughly considered is that the answer seems straightforward. If the purpose of it is to produce predictions of the risk of recidivism and if one knows that risk predictions are not infallible—which indeed is the case for all existing risk assessment tools—then it seems reasonable to hold that what matters in the comparative assessment of human and algorithms is the accuracy of the predictions. Accuracy, as already indicated, is usually considered the “gold standard” in considerations of risk assessment. Thus, following this view, the obvious criterion for assessment would be to contend that, in the comparison of algorithmic and human risk assessments, replacing the latter with the former would be justified if and only if the algorithm produces more accurate predictions than do humans (or vice versa). If, as initially assumed, we leave out other factors that might count in an all-things-considered comparative assessment and focus only on the performance of the risk assessment tools, then it certainly seems appealing to suggest that one tool must be preferable to another if it performs better in the sense of delivering more of what it is designed to—that is, accurate predictions. However, as we shall now see, there is only one problem—namely, that this criterion does not stand up to closer ethical scrutiny.

The problem is that the ethical quality of a risk assessment tool is not only a function of predictive accuracy. This can easily be demonstrated by a simple analogy. In relation to the recent pandemic, many people have tested for Covid-19. It is a well-known fact that such tests are not infallible. Does this imply that the preferable test is the one that provides the most accurate assessment of whether one is infected? The answer is in the negative. The reason is that the quality of a test does not only depend upon the accuracy of the test results, but also upon what type of error a test produces. A false positive error—that is, the error of showing that a person is infected when this is not the case—is usually not very serious. All it implies is perhaps that the person will stay at home for a brief period for no genuine reason, that is, without being a risk to others. Comparatively, a false negative result is usually much worse. It may imply that the person continues business as usual and thereby transmits the virus to other people. Therefore, from a societal perspective, false negatives will usually be more serious than false positives. But this also means that one type of test may be preferable to another if it has a preferable error profile (i.e., ratio of false positives and false negatives) and, importantly, this may be so even if the test is less accurate. The same may be the case with regard to other types of predictive tools such as those mentioned in our discussion of risk assessments.

The prediction of recidivism may be fallacious either by making a false positive assessment, that is, by predicting that an offender will re-offend when this is not the case, or by making a false negative assessment, that is, by predicting no re-offending when this would in fact occur. How, then, should we ethically assess these two types of error? The answer depends upon how the criminal justice system reacts to such predictions, which also implies that no universal answer can be provided (see also [18]). Consider first a false negative prediction. Suppose that the criminal justice system reacts to this prediction by giving the offender a non-custodial rather than a custodial sentence, and that this results in an instance of re-offending. In that case the cost of the mistaken prediction will be related to the harmful consequences of the crime. However, this may be very difficult to estimate in advance. An offender can re-offend in many ways, committing more or less serious crimes (see e.g. [19]). Thus, it is not an easy task to estimate the costs of this sort of error. The same is the case with regard to the costs of false positive predictions, but for very different reasons. Suppose that an offender, because of a positive prediction, receives a longer prison term. In that case the most immediate costs of a false prediction relate to the unnecessary extra harm that is imposed on the offender by keeping him or her behind bars for a longer period. However, the picture is more complicated than that. Many theories of punishment imply that, depending on the seriousness of the crime, there are upper limits to how severely an offender should be punished. For instance, this is a view defended by both so-called negative retributivist and limiting retributivist theories of punishment.[5] But this means that if a false positive prediction implies that an offender receives a more severe sentence, then it will be crucial whether this person

[5] For presentations and discussions of these positions, see e.g. [20–22]. Obviously, traditional positive retributivist theories also imply that there are upper limits to how severely offenders should be sentenced. But according to such theories, risk assessments do not play a role when it comes to the determination of appropriate sentences.

ends up being sentenced in a way that violates such upper proportionality constraints; to put it more simply, whether the person is punished in an unacceptably severe manner. However, this is not an easy question to answer. Theorists in the modern retributivist tradition have had very little to say about precisely how severely different crimes should be punished. Some theories concerning the appropriate penal levels for different crimes have been developed (e.g. [23–25]). But these theories have also been heavily criticized. Furthermore, such theories, even if they are taken for granted, usually only provide a very general structure in terms of how appropriate sentences could be determined. What they imply in practice when it comes to how severely a burglar, a rapist, or a murderer should be punished remains unclear.

The point of these considerations is twofold. The question we have been considering is what would constitute a plausible criterion for the assessment of the relative performance of algorithms and humans with regard to risk assessments. What we have seen, firstly, is that perhaps the most obvious answer—namely, that the performance of algorithms and humans should be compared on the grounds of the accuracy of the risk predictions that are provided—is premature. An adequate assessment also needs to take the error profile of such predictions into account. This may imply that an algorithm could be preferable to a human risk assessment (or vice versa), even if both provide equally accurate risk prediction, if the algorithm has a preferable error profile.[6] Moreover, it may even imply that there could be cases, analogous to the Covid-19 test example, where an algorithm is preferable to human predictions (or vice versa), even if it provides less accurate predictions.[7] Once again, this will depend on the types of error it produces.

The second thing we have seen is that once we try to include both considerations of accuracy and error profiles into the assessment of the comparative performance of algorithmic and human risk predictions, then this sort of comparison is complicated by the fact that it hinges both on empirical estimates that are difficult to make and on ethical considerations which are not currently sufficiently developed. It is important to underline that the point of these considerations is not to suggest that it will be impossible to make the relevant sort of comparison. For instance, it may be the case that the requisite penal ethical resources with regard to how severely different offenders should be punished will at some point be developed. The point, therefore, is the more modest one, namely, that we cannot currently say that we possess adequate criteria for the comparative assessment of the performance of different methods of risk assessment and, consequently, that we cannot yet tell when it would be justified to replace human assessments with those provided by algorithmic tools. What *prima facie* seemed a very manageable type of comparison, on closer scrutiny turned out to be a highly complicated task.

2 The determination of sentences: algorithms vs humans

While considerations of the justified use of risk assessment algorithms are pertinent in the sense that such instruments are currently used every day in criminal justice practice in a range of jurisdictions, the same is not the case if we turn to the possibility of using algorithms to determine sentences in criminal cases. Algorithms that provide sentence recommendations have been implemented in a few places, as already noted, but fully automated sentencing decisions have not yet been put into practice. However, this may become a possibility in the future. The question to which this possibility gives rise, then, is when should a sentencing algorithm be regarded as performing sufficiently well to replace a human sentencing judge? Once again, it should be kept in mind that we are not here engaging in all-things-considered comparisons of algorithms and humans but only comparisons with regard to the performance of meting out appropriate sentences.[8] Given this focus, it is easy to imagine cases in which it would certainly *not* be justified to replace a judge with an algorithm. For instance, if it turned out that the algorithm would give a life sentence every tenth time it was dealing with an instance of theft, then obviously the algorithm has not yet been properly calibrated. But when should we say that an algorithm is performing well enough to supplant human decision-makers in the distribution of sentences?[9]

As was the case in the discussion on risk assessment algorithms, here there is also a potential assessment criterion

[6] For a more detailed discussion of accuracy and error profiles in relation to the discussion of algorithmic fairness, see e.g., [26] and [27].

[7] For instance, this could be the case if a sentencing system reacts forcefully to false positive predictions of re-offending and if the crimes committed by those who are false negatives are not very serious. Under these circumstances, the introduction of an algorithm might be preferable to human risk assessment if the algorithm's ratio of false positives and false negative predictions is such that there is a lower rate of the former type of errors but a higher of the latter. Conversely, if the crimes that would be committed by those who re-offend are rather serious and the sentencing system does not react very forcefully to predicted positives, then the algorithmic prediction might be preferable if its balance of false positives and false negatives is such that it makes fewer of the latter errors but more of the former. In both of these cases the algorithmic predictions could be preferable to human predictions even if the algorithms score less in overall accuracy (see also [18], p. 255].

[8] For a recent discussion of how the use of algorithms may affect the sentences process at the courts, see [28].

[9] For discussions of this issue, see also [29].

that easily comes to mind. It might be suggested that an algorithm is working sufficiently well to replace a human judge if and only if it is able to determine the very same sentences that would have been determined by human judges. For instance, if a judge would have given a $1000 fine for a theft, two years in prison for a burglary and five years in prison for a rape, and if it turns out that the algorithm reaches the very same decisions, then one could say that it has reached a performance level at which it would be justified to replace the decisions of a human judge with those of the algorithm.[10] In practice, there may of course be some problems associated with this sort of comparative assessment. For instance, it may not be clear how many instances of cases where the algorithm provides the same answer as a human judge should be regarded as sufficient to reach the conclusion that the criterion has been satisfied. Moreover, there is the well-known challenge that it is not always the case that judges within the same jurisdiction reach the same sentences. Many studies have over the years demonstrated the existence of sentencing disparity—that is, that the same crimes are not always punished with the same sentences—between judges within the same jurisdiction (see [29, 32]). However, if for the sake of the argument we leave such more practical issues aside, should we then maintain that indistinguishability between sentences determined by human judges and by algorithms constitutes an ethically plausible criterion for justified replacement?

Despite its immediate appeal, it turns out that, on closer scrutiny, indistinguishability does not constitute a plausible criterion. The problem is that it is based on the assumption that sentences determined by human judges are ethically perfect, so to speak, and therefore that any deviation from the sentences that a judge would mete out must be regarded as an ethically erroneous judgment. However, this assumption is highly dubious. It is today a well-known fact that there exist many types of cognitive bias, and many studies have been conducted that support the conclusion that such biases may also influence the decisions of judges.[11] Now, suppose that in a particular case the human judge has reached the conclusion that an offender should spend eight months in prison. Suppose, further, that this sentence was influenced by a certain bias and that the proper sentence, that is, the sentence the judge would have reached had he or she not been biased, would have been six months in prison. Suppose, finally, that this is the very sentence which the algorithm would have reached. In this case, it would certainly be absurd to contend that it is the algorithm which is involved in an unacceptable deviation. Rather, the obvious conclusion would be that the algorithm is on the right track and, consequently, that deviations from the sentences that would have been determined by a human judge cannot always be assumed to be ethically undesirable.

More generally, it is also worth noting that many penal ethicists today subscribe to the view that penal systems are not working as they ideally should. For instance, many believe that offenders are currently being over-punished and that the imposition of long prison terms should be used much more sparingly, not only in the US but also in other countries.[12] But if such views are plausible, then they clearly give further support to the conclusion that one cannot take it for granted that the sentences reached by human judges are always ethically perfect and, therefore, that deviations from the sentences imposed by judges must always be regarded as ethically undesirable. In fact, in this respect the work of sentencing algorithms will have a direct affinity with the work carried out by many other types of algorithm. For instance, the aspiration when algorithms are used to analyse scan images in a medical context is not only that these algorithms should be able to reach the same judgements as human radiologists. Rather, the aspiration is that such algorithmic tools will be able to outperform humans (see [40]). If, as argued, the sentences determined by human judges cannot be assumed to be ethically perfect, then a similar aspiration also seems plausible when we are considering sentencing algorithms. Therefore, in short, indistinguishability does not, after all, constitute a plausible criterion for the assessment of the comparative performance of human judges and algorithms.

But if this is the case, what would a plausible criterion look like? A possible general answer, and an answer that would be able to handle the challenges confronting a criterion based on indistinguishability, would be to hold that a sentencing algorithm is performing better than a human judge (or vice versa) if and only if it reaches sentences that are preferable according to our best ethical theories of punishment. Following this idea, we can safely assume that an algorithm that decides on six months of imprisonment rather than eight months in the above example of biased human sentencing would no longer constitute a problem, but rather an improvement. Likewise, deviations in a more lenient direction might constitute improvements within a sentencing framework in which human judges tend to over-punish offenders. Thus, such a criterion seems much more plausible

[10] For a futher discussion of the indistinguishability criterion, see also [30] and [31].

[11] To mention a few examples, studies have been conducted on the significance of anchor effects [33], hindsight biases [34], and perspective effects [35].

[12] This view has been defended by many penal theories on the grounds of varying penal ethical positions. See e.g. [25, 36–39]. However, this view has been defended without providing clear answers to how severely different crimes should be punished.

than the one based on indistinguishability. In fact, it may sound almost like an ethical truism. How can one possibly reject the contention that an algorithm is performing better than human judges (or vice versa) if it is determining sentences that are preferable according to our best ethical theories of punishment? However, although it is indeed a plausible criterion, problems arise once we try to spell out this criterion in detail.

The first obvious challenge is that there is today no consensus when it comes to the question of what constitutes the most plausible ethical theory of punishment. It is true that retributivist theories have dominated the field over the previous decades. However, as often described, retributivism does not denote a single theory of punishment. Rather, it is more apposite to regard the term as an umbrella concept covering a range of different desert-based theories which diverge in various respects. Furthermore, it seems fair to say that the penal theoretical field is today even more diverse than in previous periods. Not only does the field comprise many accounts of retributivism, but there are also penal theorists who subscribe to different versions of consequentialism, restitutionism, self-defence theories, right-forfeiture theories and other theories as well.[13] Thus, one cannot plausibly hold that research within the ethics of punishment has yet been able to identify the best ethical theory in the field.

The second challenge is that most of these theories have had very little to say about the more precise question of how severely different offenders should be punished for their misdeeds. What these theories have mainly been concerned with is the basic question of the justification of punishment. However, when it comes to the distribution of punishment, various problems arise. For instance, with regard to consequentialist theories, it is fair to say that it has been empirically underdetermined precisely how severely offenders should be punished. And if one turns to different accounts of retributivism, the question has often been theoretically underdetermined. It is indisputable that, in the modern retributivist-dominated area, there has been a move from the "why punishment?" question to the question of "how much?" (see [43]). But it is also fair to hold that very little has been achieved when it comes to precise answers to how severely different crimes should be punished. Retributivists have usually underlined the importance of maintaining (ordinal) proportionality in sentencing. But all this implies is that more serious crimes should be punished more severely than less serious crimes and that equally serious crimes should be responded to with equally severe sentences. This does not say anything about how severely crimes should be punished. Furthermore, as previously noted, it is a fact that some retributivists have attempted to address the "how much?" question. For instance, so-called "anchor theories" have been suggested concerning how a scale of crimes, ranked in ascending order of seriousness, and a scale of punishments, ranked in severity, should be linked. However, there is currently absolutely no consensus with regard to which of these attempts is the most plausible. The theories have been the subject of various sorts of theoretical criticism (see also [21]). And they have typically only provided the overall contours of how one might reach answers to questions of how severely different crimes should be punished. Thus, more precise answers cannot be supposed to be available. And even though we do not have the space here to enter considerations of all alternative penal theories, it is fair to say that no other theories have yet succeeded in delivering precise answers to how severely crime should be punished.[14] However, and importantly, this is precisely the sort of answers that would be needed in order to apply the suggested criterion. If a sentencing judge metes out a sentence of eight months in prison, whereas a sentencing algorithm prescribes six months behind bars to the same offender, then we need to know whether this deviation should be regarded as ethically unacceptable or rather as an improvement. But at this point we do not seem to possess the requisite theoretical resources.

In summary, what we have seen in this section is the same pattern as in the discussion on the criterion for the assessment of the performance of risk predictions in the previous section. A plausible criterion for assessing the performance of sentencing algorithms and, hence, for answering the question of when it is justified to replace human sentencing judges with sentencing algorithms, presupposes ethical answers which current penal theory has not provided. Once again, it should be underlined that the point is not to suggest that it is impossible to provide such answers. In fact, we would hope that much more research will be conducted on the complicated question of punishment distribution. The point simply is that we cannot plausibly say that we currently possess the theoretical goods that are required to make an ethically justified assessment of the comparative performance of algorithms and human judges.

3 Implications for the use of algorithmic tools

Let us assume that what has been argued in the previous sections is true; that is, that the comparison of the performance of both risk assessment algorithms and sentencing algorithms with that of medical professionals and sentencing

[13] For recent overviews of the many competing penal theories, see e.g. [41] or [42].

[14] For some of the most recent considerations of the "how much?" question, see [44] and [45].

judges, respectively, presupposes assessment criteria that are contingent on penal ethical considerations and, furthermore, that we cannot yet be said to possess ethical theories of punishment distribution that are sufficiently developed to provide the sort of guidance that the comparative assessments require. What then are the implications? Should this be regarded as a minor inconvenience or a more serious problem?

The most immediate implication, obviously, is that if we are considering the implementation of algorithmic tools instead of humans in the sentencing process, then we do not yet possess the theoretical resources to tell whether an algorithm would provide answers that would be preferable to those provided by humans. Of course, this does not imply than any kind of comparison will remain impossible. For instance, as already indicated, there may be cases where more fine-grained ethical theory is not required. If a risk assessment algorithm provides clearly absurd predictions, such as the predictions that all offenders are at high risk of re-offending, or if a sentencing algorithm prescribes that all offenders should receive a life sentence regardless of the crime that has been committed, then obviously it would not be a problem to rule out the possibility of introducing such tools in sentencing practice. However, it is precisely when the use of such instruments cannot simply be rejected—because they provide answers that are not clearly absurd—that more precise comparisons of the performance merits of these instruments relative to those of humans will require the sort of criteria that have been considered and which we have seen are not yet available.

This should be regarded as an important conclusion if one is considering the implementation of sentencing algorithms in penal practice. However, in a way this conclusion is more noteworthy with regard to risk assessment algorithms because these tools have already been implemented. If what we have seen so far is correct, then the implementation of these tools has happened without the proper ethical criteria for judging whether they are performing in a way that is preferable to that of humans. Perhaps it is precisely the feeling of an absence of firm ethical grounds for comparative assessment that have led some commentators to characterize the implementation and current use of such tools as an "experiment" and to contend that some US states have "been likened to policy laboratories" [19], p. 214).

However, it is also important to note that the previous conclusions are not only of relevance if one is considering the justified implementation of algorithms in lieu of humans. They are of equal relevance if one wishes to reconsider the use of algorithms once they have been implemented. If one has implemented risk assessment algorithms or sentencing algorithms in penal practice, then it seems obvious at some point to consider whether these new procedures work as they should; that is, whether "the experiment" has succeeded. However, the absence of proper ethical criteria for comparative assessment of the performance of algorithms and of humans also implies that it will not be possible to make this sort of evaluation. Thus, the previous conclusions do not only have implications with regard to prospective considerations of whether or not risk assessment and sentencing algorithms should be implemented in penal practice. They have equally serious implications in terms of the possibility of conducting a retrospective assessment of such tools once they have been put into practice. In fact, as we shall now see, the implications are even more far-reaching.

A possible reaction to the previous considerations might be that even if it is true that we do not yet possess the necessary criteria for assessments of the performance of algorithms and humans at sentencing, it is not clear that the real-life consequences would be serious. It might be objected that in penal practice, it is not very likely that algorithmic decisions will supplant human decision-making. In reality, it is much more likely that such tools will be implemented to inform and support human decisions and that the discussion of performance of algorithms versus performance of humans therefore is not very urgent after all. Now, obviously this answer cannot be given with regard to risk assessment algorithms. When it comes to such assessments, the replacement of the judgements of medical professionals with those of algorithms has already taken place. However, when it comes to the use of sentencing algorithms, which of course constitutes a much more radical step in the involvement of technology in the criminal courts, it is unlikely that fully automated decision-making will be introduced. Rather, what we should expect is precisely what has already happened: namely, that algorithmic tools will be used to provide sentence recommendations which judges can draw on when they mete out sentences. But the sentencing decisions will remain human. Therefore, it might be held, the previous considerations could be seen as directing attention to a theoretical challenge when it comes to the comparison of the performance of algorithms and humans, but it is not a challenge that is likely to have ramifications in penal practice. Is this objection convincing?

The question of whether it is likely that algorithms will in future be implemented in lieu of human judges at sentencing, and if so to what extent and regarding which types of crime, is of course an empirical question. It is not easy make precise predictions of what we can expect. If it turns out—which does not seem unlikely—that fully automated sentencing decisions, at least where less complicated types of crime are concerned, will constitute a procedure that can save time and resources, then it does not seem far-fetched to imagine that there will be pressure on decision-makers to introduce such tools in practice. In fact, this is precisely what has happened with regard to the use of risk assessment algorithms.[15] However, I shall not attempt to qualify these predictions here. In fact, there is no reason to do so. Even if it is correct that fully automated sentencing decisions constitute a hypothetical scenario, and that algorithms are much more likely to be used in the future as instruments that merely inform judicial decision-making, this will still not suffice to circumvent the challenge that has been presented. What we have seen is that we currently lack the penal ethical background for making comparisons between the performance of sentencing algorithms, on the one hand, and sentencing decisions by human judges on the other. The reason is that we do not possess the theoretical resources to determine what is ethically preferable if there is a divergence between the sentencing decisions reached in either of the compared methods. However, this comparative challenge is of course not only pertinent in the comparison of algorithms and humans. If what we are comparing is either a state in which human judges determine sentencing without drawing on the recommendations of algorithms, or a state in which human sentencing decisions are supplemented by the use of such algorithms, then we are faced with the very same problem of being able to tell which decisions will be preferable when the two scenarios lead to different sentences. Therefore, the lack of proper ethical criteria for the comparative assessment of sentencing scenarios will be urgent even if what we are considering is not fully automated sentencing decisions but only the more moderate step of implementing sentencing algorithms as an aid in the ultimately human work of meting out appropriate sentences to offenders.

Thus, what we have seen is that the implications of a lack of proper criteria for the assessment of the performance of algorithms and humans are comprehensive. Not only does the lack of such criteria have direct implications with regard to prospective considerations of the implementation of such technology in sentencing practice, they also have implications regarding the possibility of reevaluating sentencing practice once such tools have been implemented. And the challenge of making justified comparisons remains intact even if one contends that the most realistic future scenarios will involve considerations of the advantage of introducing algorithms in a sentencing process where the ultimate decisions remain in the hands of human judges.

4 Conclusion

The question of when algorithmic tools work sufficiently well to take over the work of humans is currently arising in multiple contexts. Within medical treatment, algorithms have already supplanted the work of radiologists when it comes to the analysis of various types of scan images. And the time when self-driving vehicles become sufficiently safe to replace humans behind the wheels in ordinary traffic does not seem far ahead. Many other examples of replacement of humans with artificial intelligence could be mentioned. In the present article, the focus has been on the replacement of humans at sentencing. More precisely, what has been considered is the use of algorithmic risk assessment tools and of sentencing algorithms in lieu of human judgements and decisions. While increasing philosophical attention has been directed to the use of algorithms in a criminal justice context, the focus so far has been on issues concerning due process and on various types of dubious collateral consequences of the use of such technological tools (see e.g. [47–50]). Very little attention has been paid to the question of when such algorithms should be regarded as performing well enough to supplant humans.

As suggested, a possible explanation of this lacuna in the current debate might be that the answer seems uncontroversial. It seems obvious to consider a risk assessment algorithm to be performing better than psychologists and medical doctors if it produces more accurate predictions. Equally obvious is the idea that a sentencing algorithm should replace human judges only if it is able to determine the same sentences that have been given by judges. However, what has been argued, firstly, is that on closer inspection, neither of these assessment criteria are ethically plausible. Secondly, it was shown that plausible assessment criteria would presuppose answers to how severely different crimes should be punished; that is, to aspects of the ethics of punishment which have not yet been sufficiently developed. Thirdly, it was argued that this lack of workable assessment criteria would have consequences with regard to both the question of the justified implementation of algorithms at sentencing and the possibility of reevaluating the use of such algorithms once they have been introduced. Finally, it was suggested that even though the current discussion, for reasons of ease in exposition, has been framed as a comparison between algorithms and humans, the lack of proper assessment criteria would also have direct implications with regard to the possibility of assessing scenarios in which algorithms are implemented as instruments to inform rather than supplant human decision-making. Thus, the previous discussion adds a further aspect to the ethical challenges associated with the use of algorithmic tools in criminal justice practice. It also underlines the importance of engaging in more basic

[15] For instance, on the grounds of their studies, Brayne and Christin have concluded that one of the justifications for using predictive algorithms in policing and criminal courts was an "efficiency argument, which described predictive algorithms as a cost-cutting device at a time of funding and budgetary constraints" [46], p. 8).

considerations of the ethics of punishment. One can only hope that research will keep up with the pace at which artificial intelligence is currently infiltrating all parts of critical societal infrastructure, including the criminal courts.

Funding Open access funding provided by Roskilde University.

Declarations

Conflict of interest The authors have no competing interests to declare that have relevance to the content of this article.

References

1. Khazanah Research Institute: #NetworkedNation: Navigating Challenges, Realising Oppurtunities of Digital Transformation. Khazanah Research Institute, Kula Lumpur (2021)
2. Shi, J.: Artificial intelligence, algorithms and sentencing in Chinese criminal justice: problems and solutions. Crim. Law Forum (2022). https://doi.org/10.1007/s10609-022-09437-5
3. Schwarze, M., Roberts, J.V.: Reconciling artificial and human intelligence: supplementing not supplanting the sentencing judge. In: Ryberg, J., Roberts, J.V. (eds.) Sentencing and Artificial Intelligence, pp. 207–231. Oxford University Press, Oxford (2022)
4. Wingerden, S., Plesnicar, M.: Artificial intelligence and sentencing: humans against machines. In: Ryberg, J., Roberts, J.V. (eds.) Sentencing and Artificial Intelligence, pp. 230–251. Oxford University Press, Oxford (2022)
5. Chiao, V.: Transparency at sentencing: are human judges more transparent than algorithms? In: Ryberg, J., Roberts, J.V. (eds.) Sentencing and Artificial Intelligence, pp. 34–56. Oxford University Press, New York (2022)
6. Ryberg, J.: Sentencing and algorithmic transparency. In: Ryberg, J., Roberts, J.V. (eds.) Sentencing and Artificial Intelligence, pp. 13–33. Oxford University Press, Oxford (2022)
7. Ryberg, J.: Sentencing and algorithmic decision-making: when would it be justified to replace a human judge with a robojudge? In: Castro-Toledo, F.J. (ed.) La Transformation algoritmica del sistema de justicia penal. Aranzadi, Thomson Reuters (2022)
8. Wisser, L.: Pandoras algorithmic black box: the challenge of using algorithmic risk assessments in sentencing. Am. Crim. Law Rev. **56**, 1811–1832 (2019)
9. Davies, B., Douglas, T.: Learning to discriminate: the perfect proxy problem in artificially intelligent sentencing. In: Ryberg, J., Roberts, J.V. (eds.) Sentencing and Artificial Intelligence, pp. 97–120. Oxford University Press, Oxford (2022)
10. Lippert-Rasmussen, K.: Algorithmic-based sentencing and discrimination. In: Ryberg, J., Roberts, J.V. (eds.) Sentencing and Artificial Intelligence, pp. 74–96. Oxford University Press, Oxford (2022)
11. Hunter, D., et al.: A framework for the efficient and ethical use of artificial intelligence in the criminal justice system. Fla. Univ. State Law Rev. **47**, 749–800 (2020)
12. O'Neil, C.: Weapons of Math Destruction. Broadway Books, New York (2016)
13. Ryberg, J.: Risk and retribution: on the possibility of reconciling considerations of dangerousness and desert. In: de Keijser, J., Robert, J., Ryberg, J. (eds.) Predictive Sentencing, pp. 51–68. Hart Publishing, Oxford (2019)

14 Fazel, S.: The scientific validity of current approaches to violence and criminal risk assessment". In: de Keijser, J., et al. (eds.) Predictive Sentencing: Normative and Empirical Perspectives. Hart Publishing, Oxford (2019)

15. Tonry, M.: Sentencing and prediction: old wine in old bottles. In: de Keijser, J., et al. (eds.) Predictive Sentencing: Normative and Empirical Perspectives. Hart Publishing, Oxford (2019)
16. Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. Sci. Adv. **4**, 1–5 (2018)
17. Yang, M., et al.: The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. Psychol. Bull. **136**, 740–767 (2010)
18. Ryberg, J.: Risk assessment and algorithmic accuracy. Ethical Theor. Moral Pract. **23**, 251–263 (2020)
19. Hester, R.: Risk assessment at sentencing: the Pennsylvania experience. In: de Keijser, J., et al. (eds.) Predictive Sentencing: Normative and Empirical Perspectives. Hart Publishing, Oxford (2019)
20. Husak, D.: Hybrid theories. In: Ryberg, J. (ed.) The Oxford Handbook of the Philosophy of Punishment. Oxford University Press, Oxford (2024)
21. Ryberg, J.: The Ethics of Proportionate Punishment: A Critical Investigation. Kluwer Academic Publishers, Dordrecht (2004)

22 Tonry, M.: Doing Justice, Preventing Crime. Oxford University Press, Oxford (2020)

23. Davis, M.: To Make the Punishment Fit the Crime. Westview Press, Boulder (1992)
24. Scheid, D.E.: Constructing a theory of punishment, desert, and the distribution of punishments. Can. J. Law Jurisprud. **10**, 441–506 (1997)
25. von Hirsch, A.: Censure and Sanctions. Clarendon Press, Oxford (1993)
26. Hellman, D.: Measuring algorithmic fairness. Va. Law Rev. **106**, 811–867 (2020)
27. Long, R.: Fairness in machine learning: against false positive rate equality as a measure of fairness. J Moral Philos. **19**, 49–78 (2021)
28. Taylor, I.: Justice by algorithm: the limits of AI in criminal sentencing. Crim Justice Ethics **42**, 193–213 (2023)
29. Ryberg, J.: Sentencing disparity and artificial intelligence. J. Value Inq. **57**, 447–462 (2023)

30 Ryberg, J.: Criminal justice and artificial intelligence: how should we assess the performance of sentencing algorithms? Philos. Technol. (2024). https://doi.org/10.1007/s13347-024-00694-3

31. Ryberg, J.: Punishment and artificial intelligence. In: Ryberg, J. (ed.) The Oxford Handbook of the Philosophy of Punishment. Oxford University Press, Oxford (2024)
32. Chiao, V.: Predicting proportionality: the case for algorithmic sentencing. Crim. Justice Ethics **37**, 238–261 (2018)
33. Englich, B., et al.: Playing dice with criminal sentences. Pers. Soc. Psychol. Bull. **32**, 188–200 (2006)
34. Harley, E.M.: Hindsight bias in legal decision making. Soc. Cogn. **25**, 48–63 (2007)
35. Lassiter, G.D., et al.: Evidence of the camara perspective effect bias in authentic videotaped interrogations: implications for

emerging reform in the criminal justice system. Leg. Criminol. Psychol.. Criminol. Psychol. **14**, 157–170 (2009)

36. Murphy, J.G.: Retribution, Justice, and Therapy. Kluwer Academic Publishers, Dordrecht (1979)
37. Singer, R.G.: Just Deserts. Ballenger Publishing Company, Pensacola (1979)
38. Simlansky, S.: Overpunishment and the punishment of the innocent. Anal. Philos. **4**, 232–244 (2022)
39. Tonry, M.: Making American sentencing just, humane, and effective. Crime Justice **46**, 441–504 (2016)
40. Gerdes, A., Øhrstrøm, P.: Issues in robot ethics seen through the lens of a moral turing test. J. Inf. Commun. Ethics Soc.Commun. Ethics Soc. **13**, 98–109 (2015)

41 Altman, M. (ed.): The Palgrave Handbook of the Philosophy of Punishment. Palgrave Macmillan, London (2023)

42. Ryberg, J. (ed.): The Oxford Handbook of the Philosophy of Punishment. Oxford University Press, Oxford (2024)
43. von Hirsch, A.: Proportionality in the Philosophy of punishment: form “How punish?” to “How much?” Israel Law Rev. **25**, 549–580 (1991)
44. Duus-Otterström, G.: Retributivism and severity. In: Ryberg, J. (ed.) The Oxford Handbook of the Philosophy of Punishment. Oxford University Press, Oxford (2024)
45. Bagaric, M.: Consequentialism and severity. In: Ryberg, J. (ed.) The Oxford Handbook of the Philosophy of Punishment. Oxford University Press, Oxford (2024)

46 Brayne, S., Christin, A.: Technologies of crime prediction: the reception of algorithms in policing and criminal courts. Soc. Probl. (2020). https://doi.org/10.1093/socpro/spaa004

47. Ryberg, J., Petersen, T.S.: Sentencing and the conflict between algorithmic accuracy and transparency. In: Ryberg, J., Roberts, J.V. (eds.) Sentencing and Artificial Intelligence, pp. 57–73. Oxford University Press, Oxford (2022)

48 Ryberg, J., Roberts, J.V.: Sentencing and artificial intelligence: setting the stage. In: Ryberg, J., Roberts, J.V. (eds.) Sentencing and Artificial Intelligence, pp. 1–13. Oxford University Press, Oxford (2022)

49 Ryberg, J., Roberts, J.V. (eds.): Sentencing and Artificial Intelligence. Oxford University Press, Oxford (2022)

50. Thomsen, F.K.: Iudicium ex machinae: the ethics challenges of automated decision-making at sentencing. In: Ryberg, J., Roberts, J.V. (eds.) Sentencing and Artificial Intelligence, pp. 254–278. Oxford University Press, Oxford (2022)