Philosophy & Technology (2024) 37:31
https://doi.org/10.1007/s13347-024-00726-y

COMMENT

Artificial Intelligence and the Assessment of Sentencing Algorithms: a Reply to Douglas

Jesper Ryberg[1]

Received: 16 February 2024 / Accepted: 20 February 2024 / Published online: 22 February 2024

The question as to how one should assess the performance of sentencing algorithms has hitherto received almost no attention in the discussion of the use of artificial intelligence in criminal justice practice. In an attempt to initiate a dialogue on this topic, I have argued that an ethical assessment criterion is contingent on penal theory, and consequently, the fact that such ethical theories have not been sufficiently developed to determine the severity of sentences means that in many cases it will not be possible to assess the performance of sentencing algorithms.

Thomas Douglas, in his very thoughtful reply, has suggested a candidate for an assessment criterion which is designed to overcome some of the challenges outlined in my paper. What he contends is that such an assessment should be based on the mean (or median) judgement of what constitutes the optimal sentencing length as determined by a range of experts holding different penological views. This will provide what he calls the "target sentences". The performances of sentencing algorithms should then be assessed on the grounds of:

> "*Expert Agreement Criterion:* α is preferable to β in relation to some range of cases R if and only if, across R, α recommends sentences that are overall closer to the target sentence pattern R than are those recommended by β" (Douglas, 2024).

This criterion is not meant as a criterion of rightness. Rather, it is suggested as a "heuristic" which, in Douglas' view, makes it "very likely" that we will reach the right answer as to which algorithm is ethically preferable. The idea of introducing a heuristic when one is navigating a field dominated by theoretical disagreement, seems attractive. This is also partly what I had in mind when I considered the "over-punishment criterion" according to which α would be preferable to β if and only if α determines sentences that are more lenient than β (Ryberg, 2024). However, in comparison, Douglas' criterion has the major advantage that it considers sentences across a range of cases, which means that it helps to avoid

✉ Jesper Ryberg
ryberg@ruc.dk

[1] Roskilde University, 4000 Roskilde, Denmark

the problem facing the over-punishment criterion; namely, that this criterion cannot be applied when α recommends more lenient sentences in some case while β recommends more lenient sentences in other. Thus, though I am very sympathetic to Douglas' proposal it still seems to me that the main concern of my paper – namely, that the current theoretical deficiencies within the ethics of punishment will often block the possibility of assessing the performance of sentencing algorithms – remains unresolved, even if we are considering a heuristic such as the Expect Agreement Criterion. The three following examples underline this worry.

First, the current theoretical disagreement amongst experts within the ethics of punishment is not only about the severity of sentences, such as the length of prison terms. There is also disagreement about the types of punishment that should be imposed. For instance, while some theorists believe that imprisonment should be maintained as a punitive measure, others are very skeptical. But if there is expert disagreement about the appropriate types of punishment – that is, if one expert recommends imprisonment in some cases, while another recommends alternative sanctions – then it is not sufficient to refer to the "mean or median view" of sentence severity. It is not clear that the Expert Agreement Criterion will be of much help in the face of expert disagreement of this nature.

Second, suppose that we are only considering sentences that vary across one dimension; that is, for instance, the length of prison terms (or the magnitude of fines). Even in this case, it is not clear in what sense the Expert Agreement Criterion will bring us closer to what constitutes the ethically right sentences. Simply put, suppose that one penal expert is a consequentialist who believes that it follows from her theory that the right punishment for a particular crime is one year in prison. Suppose, further, that another expert is a retributivist who believes that her theory implies that the appropriate punishment is three years in prison. In the light of this sort of disagreement, would a sentencing algorithm that recommends the mean sentence (two years behind bars) have brought us any closer to what constitutes the ethically right punishment (as compared to one that recommends one or three years)? In one sense, it is obvious that the two-year sentence can be regarded as a compromise between the two competing views. However, one could also argue that the recommended punishment would be ethically wrong both from a consequentialist and a retributivist point of view, and therefore the sentence would ultimately lack any sound ethical foundation. Thus, it is not clear to me in what sense the Expert Agreement Criterion is "very likely" to bring us closer to the right answer.

Third, and perhaps most importantly, the kind of deficiency that currently exists within the ethics of punishment is not only due to there being different penal theories that provide different answers on optimal sentencing severity. As pointed out in my discussion of the Penal Ethical Criterion, the main challenge is that none of these different penal theories have succeeded in providing answers as to how severely various crimes should be punished. Broadly speaking, consequentialist theories are often empirically underdetermined, while retributivist theories have not yet succeeded in developing the theoretical framework required to determine the severity of appropriate punishments (see also Ryberg, 2004, 2020). This means that there will be cases where it is currently not possible to determine the "mean or median"

sentences on the grounds of consequentialist or retributivist expert views, and where the Expert Agreement Criterion therefore will not be applicable.

In summary, I am very sympathetic to Douglas' brilliant idea of developing a heuristic that can help us to navigate this theoretical field. However, as illustrated in the three above examples, I still believe that the current theoretical deficiencies within the ethics of punishment constitute a major obstacle to the possibility of assessing the ethical performance of sentencing algorithms, and that this challenge has not been sufficiently met by invoking the Expert Agreement Criterion.

Authors Contributions Not applicable.

Funding The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability Not applicable.

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

Douglas. T (2024). Criteria for assessing the performance of AI-based sentencing algorithms: a reply to Ryberg. *Philosophy & Technology* 37 (forthcoming).

Ryberg, J. (2004). *The ethics of proportionate punishment*. Kluwer Academic Publishers.

Ryberg, J. (2020). Proportionality and the seriousness of crimes. In M. Tonry (Ed.), *Of one-eyed and toothless miscreants* (pp. 51–75). Oxford University Press.

Ryberg, J. (2024). Criminal justice and artificial intelligence: How should we assess the performance of sentencing algorithms? *Philosophy & Technology* 37 (forthcoming).