Sentencing Disparity and Artificial Intelligence

Ryberg, Jesper

Published in:
The Journal of Value Inquiry

DOI:
10.1007/s10790-021-09835-9

Publication date:
2023

Document Version
Peer reviewed version

Citation for published version (APA):
Ryberg, J. (2023). Sentencing Disparity and Artificial Intelligence. *The Journal of Value Inquiry*, *57*(3), 447-462. https://doi.org/10.1007/s10790-021-09835-9

Sentencing Disparity and Artificial Intelligence

Jesper Ryberg

Abstract. The idea of using artificial intelligence as a support system in the sentencing process has attracted increasing attention. For instance, it has been suggested that machine learning algorithms may help in curbing problems concerning inter-judge sentencing disparity. The purpose of the present article is to examine the merits of this possibility. It is argued that, insofar as the unfairness of sentencing disparity is held to reflect a retributivist view of proportionality, it is not necessarily the case that increasing inter-judge uniformity in sentencing is desirable. More generally, it is shown that the idea of introducing machine learning algorithms, that produce sentencing predictions on the ground of a dataset that is built of previous sentencing decisions, faces serious problems if there exists a discrepancy between actual sentencing practice and the sentences that are ideally desirable.

Key words: artificial intelligence; disparity; machine learning; proportionality; sentencing.

In recent years, the idea of using computer-run artificial intelligence systems for sentencing decision support has received increasing theoretical attention. While several theorists have considered the challenges and risks involved in the application of artificial intelligence in sentencing – for instance, challenges concerning algorithmic transparency and trust – others have directed attention to some of the potential advantages of involving such technology in the sentencing process.[1] Not only has it been underlined that artificial intelligence may improve efficiency in sentencing – e.g. by saving time or resources (Stobbs et al. 2018) – it has also been suggested that such technologies carry the potentials of curbing more basic challenges concerning disparity in sentencing.

[1] For discussions of challenges concerning transparency and trust in relation to the use of artificial intelligence in the sentencing process, see e.g. Kehl et al. (2017); Roth (2016); Simmons (2018); Stobbs et al. (2018); Zerilli et al. (2018).

The question of sentencing disparity has received considerable theoretical attention and is still a dominant theme in discussions of current penal practice. This is not surprising. Not only is it the case that similar criminal cases are being treated differently in different jurisdictions, but there is strong evidence of the existence of widespread inter-judge disparity within the same jurisdictions. Though it is not a simple matter to conduct such research, several studies have managed to demonstrate substantial sentencing disparity. For instance, a study conducted in Nebraska over a four-year period showed that one judge sentenced drug offenders twice the number of months in prison than did a colleague in similar cases (Kopf 2012). Similar results have been found in studies from Massachusetts and California (Divine 2018; Mason and Bjerk 2013; Scott 2010). And there is evidence to the same effect in several other Western countries.[2] Moreover, there is even some evidence supporting the existence of intra-judge variability in sentencing.[3]

The existence of disparity in sentencing has formed the background for traditional discussions of what constitutes the most desirable way of structuring sentencing discretion – that is, whether one should apply statutory sentencing principles, numerical guidelines, or mandatory sentences – but it has also prompted considerations of various types of tools and techniques that might assist the judiciary in the sentencing process. For instance, since the 1990s a number of "non-intelligent" computer-run sentencing support systems have been developed and implemented in various jurisdictions.[4] The explicit purpose of these systems has been to enable judges to pass sentences of greater uniformity. In several cases, though, these systems have not been regarded as successful (see Chiao 2018; Schild 1998). However, the development of modern artificial intelligence

[2] For a recent study of sentencing disparity within the same jurisdiction in Australia, see Farmer et al. (2017).

[3] Schild anecdotally mentions an example of a judge who says that he has changed his approach to sentencing with time (1998, p. 153). See also Chiao (2018).

[4] For an overview of some of the early systems, see Schild (1998).

technology seems to offer new possibilities. In particular, it has been suggested that systems involving machine learning, are in this respect highly promising. Such systems are intelligent in the sense that they not only do not reach decisions that are in any simple way prespecified "by hand", rather can they learn from previous experiences and constantly improve their capabilities (Mittelstadt et al. 2016). As has been underlined by recent adherents, the application of such a system in the sentencing process may have several major advantages: it may provide quick access for judges to the severity levels of previous sentencing decisions for individual crimes; it will be genuinely self-updating when new sentencing decisions are fed into the system; and, first and foremost, it may help in reducing inter-judge sentencing disparity.

The purpose of this article is to subject this *prima facie* appealing proposal to critical scrutiny. More precisely, it will be argued that, given one of the standard explanations of why sentencing disparity is undesirable – namely, that such differential punitive treatment involves a violation of the sort of fairness captured in the idea of proportionality in sentencing – the contention that sentencing disparity constitutes a problem to which artificial intelligence *qua* machine learning may provide a desirable new countermeasure, becomes premature. The overall aim will be to identify a serious dilemma confronting the use of this type of artificial intelligence in sentencing. In order to substantiate these contentions, the article will proceed as follows. Section (1) provides a brief exposition of how machine learning techniques may be involved in the sentencing process, such as has been envisioned by a recent adherent of this scheme. In section (2) it is argued that, insofar as the desirability of parity in sentencing is regarded as an instantiation of the retributivist idea of proportionality, it is no longer clear that the introduction of machine learning in the sentencing process will constitute a desirable way of curbing disparity. Section (3) discusses possible attempts at answering this challenge. This is done, mainly, by considering alternative explanations of the wrongness of sentencing disparity. Section (4) extrapolates the lessons from the previous discussion

by engaging in more general considerations of the overall challenge facing the application of artificial intelligence systems in real-life sentencing practice. Finally, section (5) summarises and concludes.

(1) Machine learning and the sentencing process

How would artificial intelligence in the form of machine learning algorithms be able to support the sentencing process? An outline of how this might work has recently been presented and defended by Vincent Chiao (2018). The overall idea is simple.

The first step of the proposal is to build up a dataset of a sufficiently large number of cases in which judges have determined what constitute the appropriate sentences for different crimes. Chiao acknowledges that there may be practical hurdles in arriving at a sufficiently rich dataset; however, for the moment, we can leave such problems aside. The next step, then, is that a judge may provide the algorithm with input concerning the relevant sentencing factors of a present case. On the ground of this information, the algorithm will deliver an output in the form of a prediction of a sentence based on the judiciary's own sense of what have previously constituted appropriate punishments. In Chiao's words, the algorithm "would provide sentencing judges, with a particularized snapshot of the central tendency of how they and their colleagues have been treating similar cases" (2018, p. 246) This information may be used in somewhat different ways in relation to a present case. Following Chiao's proposal, the prediction made by the algorithm should not be binding on the sentencing judge. The prediction is merely meant to inform the judge of what has constituted the sentencing level in previous similar cases. The judge, however, should be free to make his or her own final decision. In fact, Chiao reasonably suggests that judges should be encouraged to set aside the algorithm's decision if a case is deemed highly unusual. Moreover, it is underlined that sentencing reasons should be presented by the judge in order to ensure that the sentencing process

does not merely become "an exercise in rubberstamping" (2018, p. 246). Finally, when the judge has passed the sentence, the decision should be used as new input that will enrich the algorithm's predictive capacity in the future. Thus, what would then follow if the judiciary starts drawing on algorithms in the outlined way?

What Chiao envisions is that the algorithmic prediction will influence sentencing decisions in a simple manner. Consider first a situation in which judges determine the punishments for similar crimes without the use of the suggested machine learning algorithm. In this case it is highly likely that there will be some inter-judge variation in the sentencing decisions. Even if the judges all agree on what constitutes the relevant sentencing factors, the fact that there does not exist any strict metric for the relative weighing of these factors implies that the sentences are likely to be clustered around a mean sentence of a certain degree of severity μ. This is illustrated by the curve in Figure 1. Suppose, alternatively, that the machine learning algorithm is fully implemented. According to Chiao,

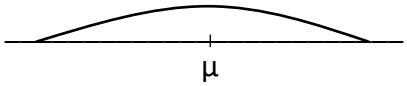

Fig. 1

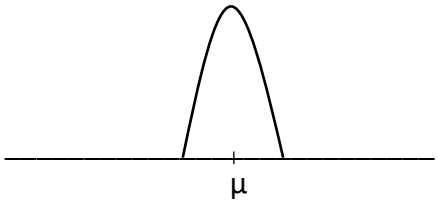

Fig. 2

what we should expect is still some variation, but the sentences will be more narrowly distributed around the mean μ as illustrated in Figure 2. In Chiao's view, this procedure is likely to satisfy both critics and defenders of discretionary sentencing. It will satisfy critics, because the system will reduce what has been regarded as arbitrariness or even "lawlessness" in sentencing (Frankel 1972). Sentences will be determined on the ground of relevant sentencing factors and the decisions in similar individual cases will be pulled closer to a mean sentence. Moreover, it will satisfy defenders by not *forcing* any artificial restrictions on the discretion of judges. The individual judge will still be free to

depart from the sentence predicted by the algorithm in cases where this is regarded as appropriate. The crucial advantages of the procedure, namely, that it will reduce sentencing disparity – i.e. that the system will initiate a transition from the sentencing distribution showed in Figure 1 to the one depicted in Figure 2 – is of course a purely empirical claim. However, it seems reasonable to believe that this will be the case. For instance, one of the explanations that has been given of inter-judge disparity is that judges are not always fully aware of the sentences that have been passed by their colleagues in previous cases. If this is correct, then it seems likely that an algorithmic prediction reflecting the judiciary's own sense of what constitutes an appropriate sentence, will in fact influence the sentencing decisions of individual judges. However, for the present, there is no need to consider this question any further. Rather, for the sake of argument, it will in the following be assumed that the system does in fact result in the suggested convergence in inter-judge sentencing decisions. The question then arises: Is this way of reducing sentencing disparity ultimately desirable?

(2) Sentencing disparity and proportionality

In order to prepare the ground for the discussion – to which we will turn in the following section – of what I believe constitutes the overall challenge facing the idea of using machine learning algorithms as an instrument for curbing sentencing disparity, we must start to consider an example of an introduction of this procedure in real-life penal practice. More precisely, suppose that we have a jurisdiction in which there exists wide-spread inter-judge sentencing disparity. Suppose, further, that the machine learning algorithm is implemented in the following way. First, every judge is asked to register all the cases they are confronted with in their daily work by specifying the relevant factors of each individual case and the sentencing decisions they have reached. Suppose, further, that after some time a sufficiently rich dataset has been built up and that the algorithm is now implemented in

sentencing practice in the way outlined above. Finally, suppose that the introduction of this procedure results in an increased uniformity in the sentencing decisions passed by judges in similar cases. Now, would this change constitute an improvement of sentencing practice in the jurisdiction?

At first glance the answer would seem to be in the affirmative. Given the comprehensive criticism that has been directed against disparity in current penal systems, it certainly seems appealing to hold that the change will constitute an improvement. However, on closer scrutiny the answer is not so simple. In order to see this, we will first have to consider the question as to why sentencing disparity should be considered a moral problem. Though it is not always the case that an answer to this question is fully spelled-out by those who object to sentencing disparity, there is one answer that immediately comes to mind; namely, that parity in sentencing reflects the kind of fairness that is captured in the idea of principle of proportionality. Conversely put, sentencing disparity is basically a question of unfairness or injustice violating the idea of proportionality in sentencing. Several theorists have underlined this basic unfairness in disparity.[5] For instance, Chiao holds that disparity raises "obvious questions about fairness" (Chiao 2018, p. 239). And Andrew Ashworth contends that "disparity is a manifest form of injustice" (Ashworth 1998, p. 236). Moreover, by explaining such unfairness in terms of disproportionality, the objection to disparity will be placed on firm ground. The principle of proportionality has been defended on the basis of various versions of retributivism that have dominated penal theory over the last four decades (see e.g. Davis 1992; Sheid 1997; Tonry 2019; von Hirsch 1993; von Hirsch and Ashworth 2005). Though the principle can be given different interpretations, the idea that "like case should be treated alike" has repeatedly been emphasized as a basic implication of proportionality. However, as we will now see, this way of underpinning the idea of using machine learning to reduce disparity faces a serious challenge.

[5] For instance, von Hirsch holds that "A sentencing system should seek to be just – or at least, to be as little unjust at possible. Claims about fairness … underline the requirements of proportionality". (von Hirsch 1993, p. 103).

The idea of proportionality in sentencing, as noted, can be given different interpretations. On a minimal account the principle concerns the relative ordering of punishments for different crimes. A more serious crime should be punished more severely than one that is less serious. And, correspondingly, equally serious crimes should be responded to with equally severe punishments. This is usually referred to as "ordinal" proportionality (von Hirsch 1993). However, though many theorists subscribe to ordinal proportionality, it is obvious that this idea cannot provide a *sufficient* answer as to how different crimes should be punished. First, ordinal proportionality does not *per se* say anything about how severe a punishment should be for a particular crime. This account only concerns how the crime should be punished relative to other crimes. Second, if this minimal idea of proportionality were nevertheless interpreted as a sufficient constraint on the distribution of punishment, it is obvious that one would open up various highly unacceptable implications. For instance, the principle would be observed by a system that punishes minor theft with 25 years in prison, as long as more serious crimes are more severely punished. Likewise, the principle would be observed by punishing murder with a 100$ fine, as long as less serious crimes are fined by minor amounts.

Therefore, in order to block such implications, proportionalists have developed theories of how severely different offences should be punished. More precisely, the idea of ordinal proportionality has been supplied by further theoretical considerations in order to provide a full and morally satisfactory theory of penal distribution. This has been done in various ways. Some theorists have suggested that one should start by ranking all crimes in seriousness, all punishments in severity, finally anchoring the two scales to each other at the end points, that is, by punishing the most serious crime with the most severe punishment and the least serious crime with the most lenient punishment (Kleinig 1973; Scheid 1997). Other theorists have defended different approaches (Davis 1992; Lippke

2012; von Hirsch 1993).[6] For now, there is no need to provide a more precise outline of the different accounts. The important thing is that these theories attempt to provide more precise answers as to what constitute the appropriate punishments for different types of crime and, ultimately, the appropriate punishments in individual cases of crime.[7] However, once this is realized, it also becomes clear that the idea of curbing sentencing disparity by the use of machine learning in order to satisfy proportionality considerations becomes premature. This is easily seen.

Suppose, as assumed, that the dataset has been built up on the ground of previous sentencing decisions, that the machine learning algorithm has been implemented, and that the new practice has led to a narrower distribution around a mean degree of severity μ. However, suppose furthermore that, given the most plausible fully developed theory of proportionality, the proportionate sentence for this crime has a degree of severity Ω which is not identical to μ. Then, as illustrated in Figure 3, the distribution will differ from the wider distribution that would have occurred had the

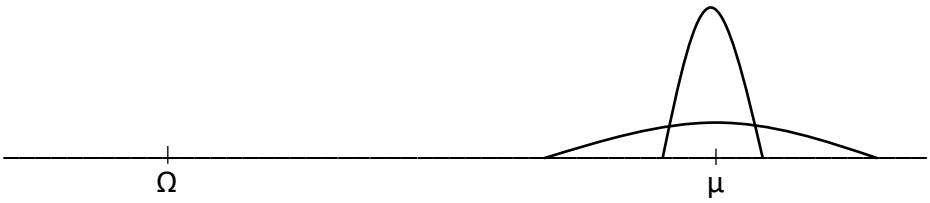

Fig. 3

[6] For an overview and critical discussion of various ways of linking the scales of crimes and punishments, see for instance Ryberg (2004) and (2010).

[7] Some theorists in the retributivist tradition would hold that desert theory does not prescribe precisely how different crimes should be punished. Rather, considerations of desert only place upper constraints on how severely an offender may be punished. In the following, I will not discuss this distinction between positive and negative versions of retributivism. The important thing is that the argument presented here is equally relevant for those who subscribe to a negative retributivist interpretation of proportionality.

algorithm not been introduced as a sentencing guide, but it is no longer obvious that the new narrower distribution is morally preferable.[8] The two alternatives constitute different distribution patterns around a sentencing level that is not the one that is ideally desirable. In other words, in order for the narrower distribution to be morally preferable, it seems that one would have to presuppose that the mean level of severity around which the sentences converge coincides with the sentence that should have been given for this crime according to the full-blown theory of proportionate sentencing. Are there reasons to believe that ideal penal theory and actual penal practice will coincide in this way? Or should we rather believe that in actual penal practice we will end up in a situation such as the one depicted in Figure 3?

If one takes a look at the penal distributional considerations that have been presented within the modern retributivist era then, even though these considerations differ in various respects, they generally share the view that there exists a major discrepancy between what is theoretically desirable and what is going on in actual penal practice. More precisely, what has repeatedly been underlined is that the sentences for many crimes are much too severe. For instance, George Singer has held that confinement should be reserved only for the most serious crime and that, even then, its duration should be relatively short. In his view, it is a misconception of retributive theory to think of it as a derivative of a "throw away the key" approach to punishment (Singer 1979). Along the same

[8] If the algorithm is introduced, there will be some offenders (at the right side of μ in Fig. 3) who will be punished less severely than they would have been had the algorithm not been introduced. However, there will also be other offenders (at the left side of μ in Fig. 3) who will be punished more severely than they would have been had it not been introduced. Thus, one distribution does not seem preferable to the other. It might perhaps be suggested that there would be a difference if the disvalue of over-punishment increases in such a way that one unit of over-punishment (say, a day in prison) is morally worse the more severely an offender is being over-punished. However, it is hard to see how this view can be buttressed and worked out in theoretical detail, and no-one – to my knowledge – has ever defended such a view.

lines, Jeffrie Murphy has suggested that, if retributive theory were to be followed consistently, one would punish less and in more decent ways than one actually does (1979). Several more recent theorists in the retributivist tradition have made similar claims. In fact, Andreas von Hirsch – one of the theorists who has contributed the most to the development of the modern conception of proportionality – has held that terms of imprisonment even for the most serious crimes, should seldom exceed five years (von Hirsch 1993; see also von Hirsch and Ashworth 2005). Thus, there is little theoretical support of the view that penal practice coincides with what is morally desirable.[9]

What then does this imply? Given these considerations, the conclusion seems straightforward. If the implementation of a support system in the form of a machine learning algorithm implies that sentencing in similar cases will become more uniform, that is, that inter-judge disparity will be reduced then, insofar as a penal distribution should be governed by a full-fledged theory of proportionality, it does not follow that the increased uniformity in sentencing is morally desirable. This would presuppose that sentences converge around the penal level which, in fact, is morally appropriate. However, this is a premise which proportionalists seem almost univocally to reject. Proportionalists generally believe there to be a significant mismatch between actual and ideal sentencing. Therefore, unless it can be shown that parity in sentencing is desirable even when offenders receive sentences that deviate significantly from the ones they should have had – a possibility that will be considered in the next section – the introduction of the machine learning algorithm does not seem to improve sentencing practice. It only implies a narrower distribution around a punishment level that is not morally desirable.

[9] Obviously, there are major differences between the penal level in different countries. For instance, in the US a great number of crimes are punished much more harshly than in many other Western countries. However, if von Hirsch's recommendation is taken for granted, then it becomes pretty clear that even those countries that use imprisonment more sparingly will be punishing more severely than what is ideally desirable.

(3) Possible objections: The significance of parity

Even though there is widespread agreement on the view that current criminal justice systems – not only in the US but also in many other countries – do not live up to what is ethically desirable, this discrepancy between penal theory and penal practice has been generally ignored in the traditional discussion of sentencing support systems as well as in more recent considerations on the use of artificial intelligence in sentencing. This is remarkable in light of the fact that the discrepancy, as argued, raises an obvious challenge to the idea of introducing such technological tools in the sentencing process. However, the question is whether the contention that there is such a challenge is somehow premature. In the following, two potential objections, that may have struck the reader while going through the reasoning of the previous section, will be considered. It is argued that not one of these objections succeeds in meeting the challenge satisfactorily.

(A) The first objection that may come to mind might be to suggest that the account of proportionalist penal theory that has been employed in the discussion so far is not plausible after all. As we have seen, the idea has been that parity in sentencing could be regarded as an instantiation of the idea of proportionality, but also that a full-fledged theory of proportionate sentencing would imply that there are certain sentences that are proportionate to particular crimes in a non-relative sense. This means that the idea that similar cases should be treated alike has been interpreted as an implication of the view that there are specific appropriate sentences for particular crimes. For instance, if the proper sentence for a particular burglary is 1 year in prison, then it follows that another offender who has committed a burglary of precisely the same degree of gravity, should also receive 1 year in prison. In other words, the purely relative idea of proportionality has been held to follow from an absolute

view of proportionality.[10] However, it might be held that this way of interpreting proportionality is premature. More precisely, it might be suggested that relative proportionality *in itself* is morally valuable; this is not merely a principle that is required in order to develop a complete theory of punishment distribution. In fact, this is a view that has significant intuitive appeal (see e.g. von Hirsch 1993). Several desert theorists have underlined that relative desert has value independently of absolute desert considerations (see e.g. Duus-Otterström 2019). But if this is the case, then it could also be held that a decrease in inter-judge sentencing disparity is at least in one respect morally desirable; that is, that the narrower distribution of sentences around μ will *ceteris paribus* be morally preferable to a wider distribution.

The problem with this objection, however, is that it rests on an insufficient interpretation of the contents of a theory of proportionality. As we have already seen, there are no penal theorists who would hold that all that matters in the distribution of punishment is the idea of justice as expressed in the relative concept of proportionality. As noted, this would commit one to the view that 20 years in prison for minor theft is acceptable as long as the relative ordering of other sentences is observed. What this means is that a non-relative or absolute idea of proportionality cannot be abandoned.[11] But this opens up the intricate question as to how these two conceptions of justice –

[10] In his comprehensive analysis of desert, Shelly Kagan notes that "when noncomparative desert is perfectly satisfied, comparative desert is perfectly satisfied as well" (Kagan 2012, p. 352). See also Duus-Otterström 2019.

[11] Some retributivists may prefer to say that a fully developed theory of punishment – that is, one that provides the answer to how specific crimes should be punished – need not be based on absolute proportionality if this implies that it is only considerations of justice that provide this complete theory. It might be held that other types of consideration – say consequentialist consideration – will have to be involved within a proportionalist scheme in order to reach a complete theory. However, the important thing here is that theorists would subscribe to the view that a complete theory should provide answers as to how specific crimes should be punished (and that this is so, independently of whether one talks of absolute proportionality or simple punishments that are morally appropriate for particular crimes).

relative and absolute proportionality – should be worked together into a coherent overall theory. From the outset there are three possible ways in which this can be done.

The first possibility would be to contend that the idea of justice captured in relative proportionality always has primacy to the one captured in considerations of absolute proportionality. Another way of putting this view – well-known in traditional discussions of value pluralism – is to say that there is a lexical ordering between the two types of proportionality, such that relative proportionality dominates absolute proportionality. However, this interpretation is vulnerable to almost the same objection as the one directed against the view that only relative proportionality matters. If one offender has been given 20 years in prison for minor theft, and if another offender has committed the same crime, then on this account he or she should also be given 20 years behind bars. And this is so, even if the proportionate sentence absolutely speaking would be, say, two months of probation. I believe that if one finds the punishment of the first offender unacceptable, one will also be inclined to regard the sentence of the second offender as unacceptable, despite relative proportionality is being observed. Therefore, this possibility does not seem plausible.

A second possibility would be to hold that, even though both the ideas of justice captured by respectively the relative and the absolute account of proportionality should count morally, it is always the case that absolute proportionality lexically dominates relative proportionality. At first sight this answer may seem more promising. For instance, on this account it could be admitted that, even if there is a discrepancy between penal theory and penal practice, such that the morally appropriate sentence in Figure 3 would be Ω, it is nevertheless the case that if in practice the penal system tends to sentence around μ for this sort of offence, then it would be morally preferable if sentences are more narrowly distributed around μ. This would be the case because, if the narrower and the wider distribution of sentences around μ are equally disproportionate in absolute terms then,

given the fact that relative proportionality has a lexically subordinated value, it follows that the narrower distribution would be preferable.

However promising this approach may seem, it nevertheless also faces problems of its own. Briefly, the problem is that any tiny improvement in terms of absolute proportionality would be preferable, even when this implies a significant loss in terms of relative proportionality. This can be illustrated by the use of Figure 3. Suppose we have a situation, as indicated in this figure, where the ideal sentence for a particular offence is Ω, but where the actual sentences are spread narrowly around μ. Suppose, further, that it is possible to introduce a new practice that would imply that the mean sentence becomes marginally more lenient, that is, it would be pushed slightly closer to Ω. In that case, the latter alternative would be preferable to the one indicated in the figure, even if this new practice also meant that the sentences of this sort of offence would be spread very widely around the new mean sentence; that is, even if this would result in significant inter-judge disparity in sentencing. I believe that if one really subscribes to the view that relative proportionality has value in itself – which is the core of the objection we are here considering – then it is hard to believe that one would find such an implication morally acceptable.

What the previous considerations indicate is that, if one holds that both absolute and relative proportionality have value, then the most appealing approach seems to be to reject the idea of one type of justice lexically dominating the other. Rather what one would have to hold is that these two values should be weighed against each other. Though the question of the relation between relative and absolute proportionality has as yet received very modest theoretical attention, the few theorists who have engaged in such considerations have reached the same conclusion (see Duus-Otterström 2019; Kagan 2012). However, this leaves a significant theoretical challenge; namely, how should these two types of proportionality be weighed against each other? The problem is that it is very hard to imagine a plausible answer. The first question that arises is whether relative proportionality and

absolute proportionality are at all commensurable values. And, even if this is the case, the next question is how differences in relative and absolute proportionality would contribute in the overall weighing. To my knowledge, no one has even tentatively managed to give an answer to these questions. Put in the words of Duus-Otterström, one of the few theorists to have addressed the issue: "… weighing relative and absolute proportionality should be added to an already long list of fascinating and vexing problems facing retributivist thought" (2019, p. 48). The problem is, of course, that in the absence of some sort of weighing theory, the suggested answer will fail to provide any genuine guidance with regard to how different offences should be sentenced. It might perhaps, and with some plausibility, be held that even though a weighing theory is required in order to deliver genuine action guidance, the lack of such a theory does not establish that this sort of value pluralism is mistaken. However, though there is certainly something to this answer, the problem in the present context is that the lack of such a theory would, in many cases, leave it unclear whether the introduction of machine learning and a subsequent change in the sentencing pattern would actually be preferable. For instance, in a situation as the one just outlined, in which there is a minor gain in absolute proportionality but a more significant loss in relative proportionality, it would remain unclear whether this constitutes an improvement and, hence, whether the machine learning support system ought to be implemented. Therefore, this third possibility does not really seem capable of providing the theoretical framework for an assessment of changes in the sentencing practice that might follow from introducing machine learning as an instrument to reduce sentencing disparity.

In summary, what these considerations show is that the attempt to block the challenge I have presented against the initially sketched idea of introducing machine learning in penal practice, by insisting that not only absolute proportionality but also relative proportionality is intrinsically valuable, is confronted with comprehensive theoretical challenges to which it is very difficult to imagine satisfactory answers.

(B) The second answer to which we will now turn follows nicely in the wake of the theoretical considerations we have just been through. The attempt to reject the challenge that has been advanced by invoking further epicycles of proportionality resulted in serious theoretical intricacies. Thus, it might be suggested that there is another way to go that is theoretically much less demanding; namely, to hold that the narrower distribution of sentences around a mean sentence is preferable for consequentialist reasons. As has been argued by several theorists, it seems reasonable to believe that any plausible theory of punishment will have to take consequences into account (see e.g. Husak 2019; Ryberg 2019). Accordingly, it might be held that even if retributivist considerations imply that the appropriate sentence for a certain crime is Ω (see again Figure 3), then under non-ideal conditions where the alternatives are either a wide or a narrow distribution around μ, the latter will be preferable for consequentialist reasons. This answer might be sustained by considerations of the fact that unequal treatment of offenders for similar crimes may well be *regarded* as highly unfair, that this might evoke disrespect for the courts, and perhaps more generally undermine confidence in the criminal justice system; all highly undesirable consequences. In short, since the two possible distributions of sentences around μ are equally unjust seen from a retributivist perspective – both depart equally from the appropriate sentence level Ω – and, since the narrower distribution is preferable for consequentialist reasons, it seems reasonable to hold that everything considered the narrower distribution is preferable. In other words, from such a theoretical perspective that combines retributive and consequentialist considerations there would be good reasons to introduce machine learning as a support system in sentencing.

While this answer has immediate attractions in bypassing the theoretical challenges associated with the previous answer, it is, however, still not clear that it succeeds in providing a sufficient justification. The suggestion faces two challenges. First, in order to provide adequate guidance this suggestion may have to be supplied by further theoretical considerations. For instance,

if we reconsider the above example, where a change in sentencing practice implies that the mean sentence is marginally less severe than μ – that is, it is moved slightly closer to Ω – then it is not clear what follows. Either, one will have to hold that justice considerations trump respect to consequences, and that this minor decrease in injustice makes this alternative morally preferable even if it were to be followed by a very wide distribution of sentences for the same crimes, that is, if disparity increases. Or, one will have to hold that one should somehow balance considerations of justice and consequences in order to assess this possibility. But it is very difficult to explain how this should be done and, in the absence of such a theory, one will not possess the sufficient theoretical capacities required to assess possible alternative scenarios related to the introduction of new AI-technologies in the sentencing process. Second, and more importantly, if we follow this way of justifying the introduction of machine learning in sentencing, then it would have to be established that the narrower sentencing distribution is in fact preferable. This is basically an empirical question. As indicated, there may in some cases be strong reasons to hold that disparity will be regarded by offenders and the public as highly unfair. For instance, this is likely to be the case if the disparity follows an easily recognizable pattern – say, that members of a minority group are systematically sentenced more harshly. However, if there is no such simple pattern in the disparity then it may be less detectable for offenders or the public at large and, therefore, less likely to cause undesirable consequences. Furthermore, from a consequence-based perspective, there may even be some reasons in favour of disparity. Even if two crimes are similar with regard to all factors that count from a retributivist point of view – i.e. the harm of the crime and the culpability of the offender – there could be some consequentialist reasons in favour of passing different sentences on the two offenders. For instance, there may be reasons concerning differences in dangerousness, how the sentence will influence the lives of the offenders, or similar considerations, that would count in favour of differential treatment. What this shows is obviously not there are general reasons against the introduction of machine

learning in sentencing. Rather the point is whether this technology will constitute an everything-considered improvement, when seen through the lenses of a theory that combines retributive and consequentialist considerations in the suggested way, that still needs to be shown. And such an assessment may vary from one context to another. A general conclusion is not warranted.

What we have seen, therefore, is that none of the considered objections managed to rebut the challenge directed against the use of machine learning algorithms as a supplement in the sentencing process. To hold that considerations of relative proportionality would sustain a system that would reduce disparity, opened up a number of problems. An answer along these lines was provided at the cost of one being led into a theoretical wilderness that seemed to almost drain the theory of general sentencing guidance and would, at least in some cases, leave it unclear whether a change in the sentencing pattern caused by the use of machine learning tools would be morally desirable. Furthermore, an approach that combines retributivist and consequentialist considerations to sentencing disparity did not seem to justify any general conclusions on the desirability of introducing machine learning in sentencing. However, if these conclusions are valid, what do they ultimately establish?

(4) The dilemma of AI-based sentencing

Machine learning algorithms, as we have seen, work by making output predictions on the ground of a database. However, this raises the overall question as to how such a database should be constructed in order for an algorithm to contribute to a genuine improvement of sentencing practice. What led to the problem discussed in the previous sections was that the most obvious answer to this question – namely, to build up the database on the ground of the actual sentencing decisions of judges – was confronted with the challenge that sentencing practice may well deviate significantly from what is

ideally desirable. As we have seen, many adherents of proportionality regard current sentencing levels as disproportionate. But in that case, it was no longer clear that an increase in sentencing uniformity between judges would be desirable. This problem reflects the well-known fact that the output predictions of an algorithm are no better than the data on the ground of which predictions are made. For instance, this feature of algorithms has repeatedly been emphasized in relation to the use of machine learning to predict an offender's risk of falling back into crime. Insofar as such algorithms are based on previous decisions that are biased against certain minority groups of the society, this will be replicated in the current predictions.[12] This constitutes a major challenge to the use of algorithmic predictions. However, the problem that has been addressed here goes even further. Even if it is assumed that previous sentences were not in any way biased, it is still not clear that there would be any advantage in using such algorithms if there existed a mismatch between how offenders had so far been sentenced and how they ideally ought to have been sentenced. But if this is the case, are there any *practical* ways of overcoming this problem?

An answer that easily comes to mind might be to abandon the idea of building up the database on the ground of actual sentences. The obvious alternative would be to construct the database on the grounds what would have constituted the ethically appropriate sentences in previous cases. In this case, the gap between practice and what is ideally desirable would be closed – in terms of the figures, μ would coincide with Ω – and the narrower distribution of sentences that would follow from the use of the algorithm as a guide in similar criminal cases would now become preferable to a less uniform distribution. However, this solution faces obvious challenges.

First, building up such a database of sentences is complicated. If the database is too small, then there is an obvious risk that this will undermine or seriously threaten the quality of the

[12] This has led to a recent discussion of the concept of algorithmic fairness; see e.g. Berk et al. (2018); Kehl et al. (2017); Roth (2016)

output sentencing predictions. For instance, as Uri Shild has underlined in his discussion of the problems facing some of the early sentencing support systems, "data for violent 'professional' bank robbers would perhaps be considered together with the data for purse snatchers, giving a misleading picture of the fine-structure of the sentences" (Schild 1998, p. 166). However, if one, on the other hand, aims at constructing a large database, then this will be a highly demanding task. It will require that one has specified in advance what would constitute the ethically desirable sentencing levels for different crimes and that judges, on this background, have considered what would constitute the appropriate sentences in various individual cases. Indeed, a rather comprehensive work.

Second, even if one imagines that judges had taken part in this extraordinary task, the resulting algorithm would still be of little practical use. If the predictions produced by the algorithms differ significantly from the sentences which society has decided to impose on offenders and which have been implemented in the sentencing framework set by the legislature, then it would clearly not be accepted to suddenly introduce an algorithmic sentencing guideline that would revolutionize the existing sentencing order. Therefore, the construction of a database of what would have constituted appropriate sentences in various cases, will often not constitute a viable option.

These considerations show what I believe to constitute the basic problem in the use algorithmic sentences that base predictions on the ground of previous cases. The problem can be summarized in what might be called *the dilemma of AI-based sentencing*: An algorithm either works on the ground of a dataset that is built up on the basis of the actual sentencing decisions of the judiciary within a jurisdiction. In this case, the system may be practically workable but, as we have seen, it is not clear that this would result in an improvement of sentencing practice. Or, the database is built up on the ground of hypothetical assessments of what would have constituted the appropriate sentences in various cases. In this case, the system would in principle provide an ethically proper guideline for individual judges, but would hardly be workable under real-life circumstances where there is a

significant gap between penal theory and penal practice. If this dilemma is accurate, then it constitutes a serious challenge to the whole idea of introducing machine learning as an instrument for curbing sentencing disparity.

(5) Conclusion

The time has come to sum up the considerations of the previous discussion. The impetus behind the discussion was the *prima facie* appealing idea that the introduction of artificial intelligence in the form of machine learning algorithms in the sentencing process might, on the one hand, lead to a reduction of inter-judge sentencing disparity while, on the other, not placing any direct restrictions on the judicial discretion in individual cases. However, on closer scrutiny it turned out that this attraction was more apparent than real. If the problem of disparity is explained in terms of a retributivist idea of proportionality in sentencing – which is what several theorists contend – and if, as several retributivists also hold, such a theory implies that there are certain sentences that are appropriate for certain offences – that is, that disparity is undesirable because it means that at least one offender has not received the sentence he or she deserves – then it is not clear that reduced disparity would be preferable within a sentencing system that punishes offenders too harshly.[13] And, as we have seen, this is precisely what many modern retributivists believe characterises current penal practice. Furthermore, it was argued that attempts at meeting this challenge by invoking relative proportionality as something valuable in itself opened up serious theoretical problems. Moreover, we have seen that an attempt to justify the introduction of machine learning algorithms, on the ground of a view that combines retributive and consequentialist considerations, did not seem to warrant any

[13] Needless to say, this is not tantamount to holding that reduced disparity would always be wrong.

general conclusion. The upshot of these considerations, therefore, is that the question of whether it would constitute an improvement to implement artificial intelligence in the sentencing process is far more complicated than has so far been suggested. In particular, as we have seen, it is important to consider whether such a system is proposed within the framework of the actual penal order in which sentencing practice may deviate significantly from what is ideally desirable.

Bibliography

Chiao, V. (2018). "Predicting Proportionality: The Case for Algorithmic Sentencing". Criminal Justice Ethics 37:

Berk, R. et al. (2018). "Fairness in Criminal Justice Risk Assessments: The State of the Art". *Sociological Methods and Research* (online first).

Davis, M. (1992). Making the Punishment Fit the Crime. United States of America: Westview Press.

Divine, J. M. (2018). "Booker Disparity and Data-Driven Sentencing". *Hastings Law Review* 69:

Duus-Otterström, G. (2019). "Weighing Relative and Absolute Proportionality in Punishment". In M. Tonry (ed.), *On One-eyed and Toothless Miscreants: Making the Punishment Fit the Crime?,* New York: Oxford University Press (forthcoming).

Farmer, C. et al. (2017). "Sentencing Inconsistencies: A Case Study", *Australian Law Journal Reports* 92: 517-528.

Husak, D. (2019). "Why Philosophers (Including Retributivists) Should be Less Resistant to Risk-Based Sentencing". In J. de Keijer et al. (eds.) (2019), *Predictive Sentencing*, Oxford: Hart Publishing.

Kagan, S. (2012). *The Geometry of Desert*. Oxford: Oxford University Press.

Kehl, D. et al. (2017). “Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing”. Responsive Communities. Available at: https://cyber.harvard.edu/publications/2017/07/Algorithms.

Kleinig, J. (1973). *Punishment and Desert*. The Hague: Martinus Nijhoff.

Kopf, R. G. (2012). “Judge-specific Sentencing Data for the District of Nebraska”. *Federal Sentencing Report* 25: 50-52.

Lippke, R. (2012). “Anchoring the Sentencing Scale: A Modest Proposal”. *Theoretical Criminology* 16: 463-480.

Mason. C. and D. Bjerk (2013). “Inter-judge Sentencing Disparity on the Federal Bench: An Examination of Drug Smuggling Cases in the Southern District of California”. *Federal Sentencing Report* 25: 190-193.

Mittelstadt, B. D. et al. (2016). “The Ethics of Algorithms: Mapping the Debate”. *Big Data and Society* 16: 1-21.

Murphy, J. G. (1979). *Retribution, Justice, and Therapy*. Dordrecht: Kluwer Academic Publishers.

Roth, A. (2016). “Trial by Machine”, *The Georgetown Law Journal* 104: 1245-1305.

Ryberg, J. (2004). *The Ethics of Proportionate Punishment. A Critical Investigation*. Dordrecht: Kluwer Academic Publishers.

Ryberg, J. (2010). “Punishment and the Measurement of Severity”. In J. Ryberg and A. Corlett (eds.) (2010), *Punishment and Ethics. New Waves*. Basingstoke: Palgrave Macmillan.

Ryberg, J. (2019). “Risk and Retribution: On the Possibility of Reconciling Considerations of Dangerousness and Desert”. In J. de Keijer et al. (eds.) (2019), *Predictive Sentencing*, Oxford: Hart Publishing.

Ryberg, J. and J. Roberts (eds.) (2021). *Sentencing and Artificial Intelligence*. New York: Oxford University Press (forthcoming).

Scheid, D.E. (1997). "Constructing a Theory of Punishment, Desert, and the Distribution of Punishment?". *The Canadian Journal of Law and Jurisprudence* 10: 441-506.

Schild, U. J. (1998). "Criminal Sentencing and Intelligent Decision Support". *Artificial Intelligence and Law* 6: 151-202.

Scott, M. (2010). "Inter-judge Sentencing Disparity After Booker: A First Look". *Stanford Law Review* 63: 1-66.

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

Singer, M., *Just Deserts*, United States of America: Ballenger Publishing Company.

Stobbs, N. et al. (2017), "Can Sentencing be Enhanced by the Use of Ethical Intelligence?", *Criminal Law Journal* 41 (5): 261-277.

Tonry, M. (2019). *Of One-eyed and Toothless Miscreants: Making the Punishment Fit the Crime?* New York: Oxford University Press.

von Hirsch, A. (1993). *Censure and Sanctions*. Oxford: Clarendon Press.

von Hirsch, A. and A. Ashworth (2005). *Proportionate Sentencing*. Oxford: Oxford University Press.

Zerilli, J. et al. (2018). "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?", *Philosophy and Technology* (online first).