International Journal of Law in Context (2019), **15**, 126–139
doi:10.1017/S1744552319000077

ARTICLE

Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice

Vincent Chiao*

Faculty of Law, University of Toronto
*Corresponding author. E-mail: vincent.chiao@utoronto.ca

Abstract

Over the last few years, legal scholars, policy-makers, activists and others have generated a vast and rapidly expanding literature concerning the ethical ramifications of using artificial intelligence, machine learning, big data and predictive software in criminal justice contexts. These concerns can be clustered under the headings of fairness, accountability and transparency. First, can we trust technology to be fair, especially given that the data on which the technology is based are biased in various ways? Second, whom can we blame if the technology goes wrong, as it inevitably will on occasion? Finally, does it matter if we do not know how an algorithm works or, relatedly, cannot understand how it reached its decision? I argue that, while these are serious concerns, they are not irresolvable. More importantly, the very same concerns of fairness, accountability and transparency apply, with even greater urgency, to existing modes of decision-making in criminal justice. The question, hence, is comparative: can algorithmic modes of decision-making improve upon the status quo in criminal justice? There is unlikely to be a categorical answer to this question, although there are some reasons for cautious optimism.

Keywords: algorithms; artificial intelligence; big data; fairness; accountability; transparency; criminal justice

1 Introduction

Over the last few years, legal scholars, policy-makers, activists and others have generated a vast and rapidly expanding literature concerning the ethical ramifications of using artificial intelligence, machine learning, big data and predictive software in criminal justice contexts. Although there are significant differences between them, in general, what is common to those technologies is that they recommend treatments – police interventions, bail determinations, sentences – on the basis of mathematically structured processing, typically over extensive datasets. They thus stand in comparison to more traditional means of making decisions in legal settings, which have traditionally relied upon the judgment of line officials, guided by a wide variety of legal rules and standards, but nevertheless largely intuitive in character. The objective of this paper is to respond to a variety of ethical concerns that have been raised in the legal literature surrounding these emerging technologies, specifically in the context of criminal justice.

These concerns can be clustered under the headings of fairness, accountability and transparency. First, can we trust technology to be fair, especially given that the data on which the technology is based are biased in various ways? Second, whom can we blame if the technology goes wrong, as it inevitably will on occasion? Unlike judges, juries, loan officers, social workers, etc., arguing with an algorithm might seem about as fruitful as arguing with your toaster. Finally, does it matter if we do not know how the algorithm works or, relatedly, cannot understand how it reached its decision?

In this paper, I sketch an account of how stage-wise modes of decision can be compared in terms of the unfairness/accuracy trade-off, argue that much of the concern about individualised tailoring is exaggerated and consider the degree to which traditional conceptions of due process are in tension with algorithmic decision-making. I conclude with a brief discussion of the now highly salient, but

rapidly evolving, discussion of 'intelligibility' in this context. For the most part, I conclude that these are serious, but resolvable, concerns. More importantly, the very same concerns arise with regard to existing modes of decision-making in criminal justice. Existing systems of criminal justice already have serious problems with fairness, accountability and transparency. The question, hence, is comparative: can algorithmic modes of decision-making improve the status quo ante in criminal justice?

2 Fairness first: self-fulfilling prophecies

Perhaps the most commonly discussed fairness-related objection that has been raised in connection with algorithmic decision-making is that predictive factors can be unfairly biased. For instance, suppose that prior arrests are predictive of failure to appear or further offending at bail, and suppose that group A has a higher mean arrest rate than group B. The resulting disparity in predicted failures for members of group A relative to group B might be deemed unfair, depending upon the explanation for the difference in mean arrest rates. If it is, then using the higher mean arrest rate for group A as a basis for further negative treatment – denial of bail, for instance – would simply exacerbate the unfairness.

Contrast two cases. In the first, group A has a higher arrest rate because members of group A tend to engage in more risky, aggressive behaviour than members of group B for reasons that do not themselves reflect bias or discrimination. In the second case, the disparity arises because members of group A are more persistently surveilled by the authorities rather than any difference in base rates of criminal offending. One might regard members of group A as bearing an unfair burden in the second case, but not the first. Consequently, one might regard the use of arrest rates as a predictor of future criminality as unfair in the second case, even while allowing that it would not be unfair in the first case, where the explanation for the higher arrest rate among group A as compared to group B is exogenous to police activity.

Perhaps some disparities in criminal justice resemble the first case. Gender disparities, for instance, seem to track genuine differences in participation in crime.[1] But other disparities seem to resemble the second case. A notorious example is drug crime. Survey data suggest similar rates of use of illicit drugs among Blacks and Whites, although Black drug users are more likely to have criminal justice contact than White drug users.[2] One explanation for this phenomenon is that police have prioritised enforcement actions on open-air drug markets, primarily used by African-Americans, rather than residential transactions favoured by Whites; another concerns racial bias in policing.[3] In cases like this, predictive variables that are neutral on their face turn out to exhibit racially skewed correlations that, on account of their aetiology, it would be unfair to entrench.

A number of fairness-related objections are versions of this type of problem. For example, some commentators have expressed concern that algorithmic risk assessment in bail will unfairly deny racial minorities bail because the predictors of failure to appear or offending while on release themselves reflect racial bias in earlier stages of criminal justice.[4] Even if an algorithm does not expressly rely on race, it is likely to positively weight factors such as prior arrests and criminal record that are strongly correlated with race. Insofar as those factors are themselves reflective of policy decisions of doubtful fairness, the algorithm will simply serve as a transmission belt, replicating discriminatory practices at the subsequent bail stage. Otherwise put, basing policy on flawed statistical inputs can

[1]Silvestri and Crowther-Dowey (2008) note that '[t]he overriding consensus within criminology remains that while women do commit a broad range of offences, they commit less crime than men and are less dangerous and violent than their male counterparts' (p. 25). See also Heidensohn and Silvestri (2012, p. 344).

[2]See the Substance Abuse and Mental Health Services Administration (2014, p. 26) (finding that 8.8 percent of Hispanics, 9.5 percent of Whites and 10.5 percent of Blacks engaged in illicit drug use in 2013), available at https://www.samhsa.gov/data/sites/default/files/NSDUHresultsPDFWHTML2013/Web/NSDUHresults2013.pdf (accessed 19 February 2019).

[3]See Mitchell and Caudy (2017) (finding evidence that nature of drug offending explains Hispanic–White disparity in drug arrests, but not Black–White disparity).

[4]Harcourt (2006); Starr (2014); Hannah-Moffat (2012). For a more technical discussion, see Chouldechova (2017) as well as Berk *et al.* (2018). For a recent discussion in the context of American law, see Huq (forthcoming, 2019).

generate a self-fulfilling prophecy. If policy-makers expect *ex ante* to find more crime among group A than group B, then it is possible that they will find this expectation validated *ex post*, but only because they have spent more time looking for crime among members of group A than among members of group B.

While self-fulfilling prophecies are clearly problematic, they are not limited to computerised risk assessment. It is essentially built into the criminal process. The law of bail already requires officials to predict the likelihood of both failure to appear as well as further offending in deciding whether to release someone on bail.[5] In other words, the question is not *whether* to perform risk assessment, but *how*: through predictions based on statistical correlations across large datasets or through the clinical judgment and intuition of judges and prosecutors. Since it is rather plausible that human decision-makers also consider many of the same factors, such as a history of arrests and convictions, that correlate with racial identity, self-fulfilling prophecies are likely to be an inevitable problem in criminal justice. Hence, a better way of understanding the issue is comparative, rather than absolute: which form of risk assessment best manages the risk that bail will further entrench biases from upstream in the criminal process?

Because the criminal process comprises a series of sequentially ordered decision points – patrol, investigation, arrest, charge, bail, plea, trial, sentence, parole, pardon – decisions made earlier in the process are prone to influence decisions in subsequent stages of the criminal process. In light of this feature of the criminal process, a plausible starting point is: does this decision point *introduce* racial skew beyond what is already represented in its antecedent inputs? A completely racially neutral process might then be conceptualised as one in which no decision point introduces racial skew, whereas a racially biased process is one where one or more of the stages introduces racial skew. In other words, the impact of a decision method can be evaluated by reference to either an absolute baseline (a completely racially neutral process) or a comparative one (the racial skew present in antecedent inputs to the decision).

The impact of a mode of decision at a particular stage can be assessed in terms of its accuracy, along one or more dimensions (e.g. a weighted rate of false positives vs. false negatives). A metric of this kind would provide an apples-to-apples comparison between human decision-making and algorithmic decision-making at a decision node. We could, for instance, compare risk-assessment algorithms as against the clinical judgment of judges and prosecutors, in terms of (1) how much racial skew each form of decision-making introduces and (2) their comparative accuracy. The social choice would be greatly simplified if it turns out that one mode of decision-making is superior to the other on both counts, or even superior on just one count while no worse on the other. In a scenario like that, choosing between the two modes of decision-making would not involve a potentially controversial trade-off of one value as against the other.

There is evidence that, when assessed in this way, machine-learning algorithms can outperform human judgment.[6] Kleinberg and his co-authors trained a machine-learning algorithm on a dataset comprising bail decisions from New York City between 2008 and 2013. They found that the algorithm was able to more accurately predict crime than human judges, even when it was constrained to ensure that the racial composition of detainees tracked that of arrestees. '[I]it is possible,' they write, 'to reduce the share of the jail population that is minority – that is, reduce racial disparities within the current criminal justice system – while simultaneously reducing crime rates relative to the judge.'[7] If this result turns out to be robust, that would count in favour of preferring computerised risk assessment over clinical human judgment, for the former would strictly dominate the latter: it would show that

[5]This is what is referred to, in Canadian law, as 'secondary grounds' for pre-trial detention: *Criminal Code*, R.S.C. 1985, s. 515(10)(b).

[6]Kleinberg *et al.* (2018).

[7]*Ibid.*, at p. 32.

algorithmic assessment outperforms human judgment at both minimising introduced racial skew and improving substantive accuracy.

More generally, in evaluating the performance of risk-assessment algorithms, it would be desirable to compare various iterations of those algorithms against each other: for instance, those that are optimised to maximise substantive accuracy and those that are constrained to introduce no further racial skew (input populations constrain output populations). Comparisons of this kind would illustrate, in a particularly clear way, trade-offs between potentially competing values. How best to strike those trade-offs is a question for the moral, rather than technical, judgment of the relevant political community.

It is important to note that even a completely racially neutral process is no guarantee of racial fairness in criminal justice. Even if every stage, from reporting and investigation of crime onward, operates in a way that introduces no racial skew, structural economic, social and political disadvantage may still affect participation in crime. If those structural disadvantages are themselves unfair, that unfairness will not be remedied by even a perfectly neutral criminal process. Remedying disadvantage of that sort goes well beyond criminal justice. The question I am addressing here is considerably narrower. It is how to determine when a stage of the criminal process introduces unfair racial skew. My proposal is to compare the racial composition of the outputs of decisions made at that stage to the racial composition of its inputs, in order to clearly compare the degree to which that mode of decision introduces racial skew to the degree to which it improves accuracy against other modes of decision, including random selection.

It is worth recalling that line officials, particularly in the common-law world, have traditionally enjoyed significant discretion in the criminal process, from whether to question someone, investigate a crime, effectuate an arrest and lay charges, to seeking detention pending trial and how to sentence. Criminal justice in the US and Canada is to a large extent built out of the accretion of millions of such individualised human judgments. Both are marked by significant racial disparities in policing and incarceration.[8] Moreover, the criminal justice process is a sequentially ordered set of decisions, and it seems to generally be the case that upstream decisions exert significant influence over downstream ones, if for no other reason than that they determine the profile of the population over which downstream decisions operate. Compounded racial bias is a structural problem in criminal justice, regardless of whether we utilise structured algorithms or human judgment at a given stage of the process.

Consequently, that algorithmic risk assessment is likely to carry forward racial discrimination is not, on its own, a sufficient reason to reject such tools. *Any* mode of decision is likely to do the same. The question is simply whether algorithmic risk assessment can strike a more appealing trade-off between enhanced accuracy in the legitimate aims of criminal justice and the introduction of further racial skew at a given point in the criminal process. While there is evidence that algorithmic risk assessment may be able to do this in some contexts, I have not claimed to show that this will inevitably be the case. The claim is only that there is reason to prefer whichever mode of decision-making, whether by human judgment or predictive algorithm, strikes the most appealing trade-off between accuracy and bias.[9]

The scale and depth of the malfunction in criminal justice, particularly in the US – conviction and incarceration have now become normalised as part of the ordinary life course for Black men – give us powerful reason to resist retreating to the familiar simply because it is familiar. This is particularly so when there is reason to believe that a more structured, data-driven approach can be both more accurate and fairer than case-by-case clinical judgment has proven to be.

[8]See e.g. Owusu-Bempah and Wortley (2014, Table 10.1) (indicating that Blacks compose 2.5 percent of Canada's population but 8.4 percent of incarcerated persons and that Aboriginals compose 3.8 percent of the population but 18.5 percent of incarcerated persons). According to data from the Federal Bureau of Prisons, approximately 37 percent of incarcerated persons in the US are Black; see https://www.bop.gov/about/statistics/statistics_inmate_race.jsp (accessed 19 February 2019). For state-level data from the US, see Nellis (2016).

[9]See Grove and Meehl (1996) (meta-analysis reporting that formal, algorithmic decision-making procedures performed equivalently or better than clinical methods in 128 out of 136 studies across a wide range of subject areas).

3 Fairness again: individualised tailoring

The next set of challenges to the new technologies of criminal justice is in one respect the opposite of the previous challenge. The first fairness challenge amounts to the charge that computerised algorithms are not as objective as they seem, because they encode the discriminatory preferences of humans. The second fairness challenge – that of individualised tailoring – amounts to the claim that computerised algorithms are *too* objective, precisely because they do not give sufficient room to human decision-making. The idea here is that any computerised algorithm, no matter how sophisticated, and no matter the size and composition of the dataset on which it was trained, will inevitably yield recommendations that are too crude, and, in particular, cruder than what could be achievable by human judgment. Human judgment is responsive to an indefinitely large range of relevant factors and hence is suited to addressing decision-making contexts in which each case is unique from every other case. In short, either computerised algorithms give too much play to human discretion and judgment (as in the previous objection) or they give it too little (as in the current one).

Perhaps the clearest examples are mandatory minimums. Mandatory minimums are, in effect, very simple algorithms: if convicted of crime C, impose punishment of P or greater. One might think that this is too crude, since there may be cases where someone is convicted of C but who reasonably should receive a punishment of less than P. Yet, even if something like the one-size-fits-all objection has traction against very simple rules of this kind, it is far from clear why it should have purchase against machine-learning algorithms that are trained on a feature space with hundreds or thousands of dimensions.[10] That would be like inferring that, because a four-by-four grid of pixels cannot adequately represent the Mona Lisa, no digital image, of any resolution, could ever hope to do so. That is a pretty uncompelling inference.

More plausibly, one might understand individualised tailoring as amounting to the claim that each case is distinct, perhaps in general, but especially in the criminal context. There is certainly a sense in which, trivially, any two cases are distinct from each other. For individualised tailoring to be a substantive objection, however, it must amount to the claim that either:

1 it is not the case that like cases should be treated alike; or
2 although like cases should be treated alike, no two cases are similar along all relevant dimensions; or
3 although like cases should be treated alike, and although cases may be similar along all relevant dimensions, computerised algorithms are not sensitive to all relevant dimensions.

I start by considering the strongest interpretation of the claim. How might one try to motivate interpretation (1)? If it is interpreted as a categorical rejection of the principle that like cases should be treated alike, it amounts to a challenge to the rule of law. If treating people fairly, in the sense of judgment tailored to each person's case, means the rejection of the formal equality before the law, then of course not only are computerised algorithms called into question, but so too is much of the existing criminal process. After all, although criminal procedure is notoriously discretionary, still it would be an exaggeration to say it is *entirely* discretionary and *entirely* unconstrained by legal rules and principles.

Whether through statutory law, case precedent or local court procedure, judges, prosecutors and defence counsel are expected to be faithful to a wide range of legal rules of general applicability. Hence, if this version of the individualised tailoring objection is accepted, it will be disabling of *both* algorithmic decision-making and clinical judgment by lawyers, for the law requires precisely what this interpretation says is impermissible, namely to treat people equivalently on the basis of a discrete range of considerations as indicated by law.

More plausibly, and more sensibly, one might interpret (1) by observing that we can often identify some range of equally justifiable outcomes for a given case. Within that range, no one outcome is

[10]This cannot be a categorical objection to mandatory minima, as the degree to which it is plausible is contingent on the actual value assigned to P.

demonstrably superior to others. Hence, while like cases should be treated *roughly* alike, it is not required that they be treated *precisely* alike. This might be because of epistemic or cognitive limitations (we are unable to employ rules of sufficient granularity) or it might be because of vagueness in the legal concepts or values that we are applying (proportionality might be consistent with a range of sentences) or for some other reason. Whatever the cause, one might conclude that even cases that are similar along all relevant dimensions should not all be forced to conform to the same precise outcome.

Indeed, something like this seems very plausible. Part of the reason line officials in criminal justice tend to have as much as discretion as they do is no doubt because the kinds of decisions they are called upon to make – whether to stop someone for questioning, where to patrol, whether to order someone detained pending trial, how to sentence and so forth – do not always have unique correct answers, at least relative to the evidence available at the time of decision. Seen in this way, the concern might be that new technologies try to give an artificial air of quantitative precision to what are at bottom qualitative judgments that permissibly vary from each other.

Does this interpretation of the individualised tailoring objection provide a good reason for objecting to algorithmic decision-making in criminal justice? It does not, for the very simple reason that there is no reason that an algorithm must yield an extremely granular outcome. For instance, the Federal Sentencing Guidelines in the US (a paper-based sentencing algorithm) defined a range of permissible sentences within each cell of the sentencing grid.[11] But nothing in the nature of sentencing guidelines requires a maximally narrow sentencing band. Similarly, a computerised risk assessment need not set some artificial bright line rule that, for instance, everyone over a certain threshold determination of risk must be detained. There might be some cases, even a substantial number of cases, where the proper disposition is unclear *ex ante*, even with the benefit of an empirically validated risk assessment. It might well be defensible to allow different officials to make different judgment calls in those cases. This would not be inconsistent with requiring those officials to first consult a risk-assessment tool before they make a final decision.

Equally important, that there are not uniquely correct outcomes for some cases does not show that there are not plenty of *in*correct outcomes for those cases. A substantial part of the appeal of an empirically validated risk assessment is its ability to screen out incorrect outcomes more reliably than clinical human judgment. It may be, for instance, that popular views among judges and prosecutors about risk factors for further offending – such as employment status or drug use – turn out, upon analysis of the evidence, to be simply incorrect or of trifling significance. Furthermore, while these types of tools are not immune to other forms of racial bias, they are not prone to the kinds of unconscious or implicit biases that might be at work in lawyers and judges making bail or sentencing determinations. Algorithmic decision-making can therefore facilitate an individually tailored outcome by screening out factors that are known to be empirically or morally irrelevant. Consequently, while relying on algorithmic decision-making may cause a distribution of outcomes to be more tightly centred on the mean, this is not because the ultimate objective is to converge upon a single correct answer in all cases. Rather, a reduction in statistical variance may be defended on grounds of limiting the influence of empirically or morally irrelevant factors.[12]

Turning to interpretation (2), let us define 'all relevant dimensions' as a finite set of dimensions referred to (or implied by) a law: $\{d_1, d_2 \ldots d_n\}$. For instance, the law of bail in Canada refers to attendance in court, protection of the public and confidence in the administration of justice with regard to 'all the circumstances' as grounds for pre-trial detention.[13] As detention may be justified on any one of these grounds, interpretation (2) thus states that no two cases are similar in terms of likelihood of

[11]'If a sentence specified by the guidelines includes a term of imprisonment, the maximum of the range established for such a term shall not exceed the minimum of that range by more than the greater of 25 percent or 6 months, except that, if the minimum term of the range is 30 years or more, the maximum may be life imprisonment'. 28 U.S.C. §994(b)(2).

[12]I discuss this issue further in Chiao (2018).

[13]*Criminal Code*, R.S.C. 1985, s. 515(10).

attending court, or in terms of the risk they pose to the public, or in terms of their impact on the public's confidence in the administration of justice.

Of course, it might well be that no two cases are *exactly* alike in *all* particulars, to the n^{th} degree, even with regard to the finite list of factors, $\{d_1, d_2 \ldots d_n\}$, that are legally relevant. But that is hardly an objection to algorithmic decision-making in criminal justice. In many cases, what is required is a threshold determination about an individual, rather than a judgment about how that individual compares to others. For instance, what the law of bail requires is that people who pose an unreasonable risk to the community be detained pending trial. It is neither here nor there whether two people pose precisely the same degree of risk. All that is required is that they both pose a level of risk that the law deems unreasonable. The concern that no two cases are similar across all legally relevant criteria seems stronger in cases where, for instance, someone's detention is sought on the ground of bolstering public confidence in the administration of justice, as the law there explicitly states that 'all the circumstances' are relevant, creating a much more open-ended type of inquiry than in the other two grounds for detention. But, again, all the law of bail requires is a threshold determination about an individual case, not a comparison across cases. Hence, even if no two individuals are exactly alike in terms of whether their release pending trial would traumatise a community, what the law deems relevant is only whether the release of any of them, considered on its own, would do so.

One might also wonder whether it is really plausible that no two cases are alike in terms of the *ex ante* risk of a failure to appear or risk to the public. It is worth distinguishing between the level of risk that someone poses and the reasons why that person poses that risk. It could well be that different people are risky for different reasons (one person because of the circumstances of the crime he was charged with, another because of his criminal history), but not only are those reasons presumably finite, but, in any case, what the law demands is only that people posing equivalent risks are treated equivalently, even if the reason they pose those risks differs. In short, to vindicate this interpretation of the individual tailoring objection would require showing that no two individuals could ever pose equivalent risks. Offhand, it is unclear how one would show this.

To be sure, there are contexts in which the law is more sensitive to comparative concerns, such as fairness in sentencing, But, just as treating cases 'alike' need not mean alike in every possible respect (since there can be a margin for permissible variation), regarding cases as similar does not entail regarding them as *ultimately* similar – that is, similar in every least respect. There is a limit to the amount of detail and nuance to which the law is, and perhaps should be, sensitive.

That leaves interpretation (3), the weakest interpretation of the individualised tailoring objection. What is clear about this interpretation is that it is an empirical claim about the capabilities of algorithmic decision-making. As such, it can only be answered by looking at each particular application in its specific context. This would include consideration of whether the humans currently making those types of decisions are themselves responding in an appropriately sensitive manner to all the morally salient features of the cases before them. It might well turn out to be the case that no computerised algorithm will be appropriately sensitive to all the factors that human actors (police, prosecutors, judges) are. This is essentially an empirical claim and should be assessed based on the available evidence, rather than simply taken on faith. The rapid progress of predictive algorithms, from driverless cars to predicting tastes in music and film, should caution against overly bold speculation along these lines.

More importantly, the appropriate comparison here is not to the very best that human judgment can be, but to the proven track record of human judgment in our existing institutions, or at least reasonably likely versions of those institutions. Potentially life-changing decisions about arrest, bail, plea and sentence are often made rapidly, on a limited informational basis, by people who suffer from all the usual cognitive biases, imperfect heuristics and unconscious influences with which we are familiar. There is by now a large literature on the biases that affect judges, from the time of day a case is heard to the appearance of victim to the race of the accused/victim to proximity to an election.[14] This suggests that human

[14]See e.g. Danziger *et al.* (2011); Berdejó and Yuchtman (2013); Englich (2009); Goodman-Delahunty and Sporer (2010).

decision-makers themselves are far from perfect in providing sufficiently individualised judgment. Hence, while interpretation (3) is a serious concern, it is also one whose import is rather ambiguous.

It is worth noting here an ironic consequence of insisting too strongly upon individualised tailoring. Denying that cases can be meaningfully compared to each other not only makes it hard to see how that bias could be corrected; it makes it hard to even understand it *as* bias. After all, if no two individuals are identically situated for purposes of sentence, then it follows that no two individuals of different races are identically situated either. Racial fairness requires comparing how different cases are treated relative to some common standard. A robust interpretation of individualised tailoring denies that any such comparison is possible. This suggests that a robust interpretation of individualised tailoring will make it difficult, if not impossible, to come to terms with the challenge of racial fairness.[15] This is not a fanciful concern; empirical studies of sentencing in the US have shown that the demise of the Federal Sentencing Guidelines has been correlated with a rise in racial disparity in sentencing.[16]

Where does this leave us? There is, I think, a sensible interpretation of individualised tailoring that survives these objections, but it is both limited and contingent. This is the concern that existing statistical and predictive models might be too crude to reliably capture those features of legal cases that we would expect of them. For instance, one might argue that, although A and B might share a similar background, causing a predictive instrument to assign similar recidivism scores, they might nevertheless differ in a way that affects their future riskiness; and that a human judge might be alert to these differences in a way that a purely statistical instrument is not.[17] This is an eminently sensible concern, particularly in light of the relatively untested character of some of the predictive devices in use or being developed. However, it is a concern that should become less pressing over time, as those technologies mature. It is also limited because, rather than a categorical objection to the use of, say, predictive policing or sentencing algorithms, it supports only the thought that such devices should only be employed in ways that ensure that the ultimate decision is appropriately sensitive to case-specific context.

More ambitious interpretations of the individualised tailoring objection turn out, on closer inspection, to be unpersuasive. One might suspect that they are largely rationalisations for a felt discomfort with unfamiliar and new technologies. Be that as it may, it hardly seems unreasonable to suggest that, like new technologies in high-stakes areas generally, the use of algorithmic decision-making in criminal justice contexts merits close supervision, in part to ensure that they do not miss reliable signals. As I have noted, however, this is true not just of computerised algorithms, but of human decision-making as well. Even if human judgment at its best is capable of very finely nuanced and calibrated judgment, it is far from clear that this is generally true of our systems of 'mass justice' – particularly in the context of relatively low-level offences that do not garner significant investment of time and attention.[18] In my view, the greatest promise of algorithmic decision-making in criminal justice lies in improving the accuracy of decision-making in the context of everyday, routine cases, rather than in the rarer instances of high-stakes, and intensely litigated, cases.

4 Accountability: the due-process model

An intuitively appealing feature of dealing with humans is that you are often able to argue with them if something has gone wrong. Part of the appeal, no doubt, is that sometimes this even works. But that is

[15]Consider the sentencing of Indigenous offenders in Canada. In a landmark case, *R v. Gladue*, [1999] 1 SCR 688, the Supreme Court of Canada took notice of the Indigenous overrepresentation in Canadian jails and prisons. Rather than promulgating any objective benchmarks for sentencing judges, the Supreme Court merely engaged in a moral exhortation to judges to think especially carefully about an Indigenous person's background – often, though not necessarily, in the context of a special '*Gladue* court' – before imposing sentence. Yet, in the years since *Gladue*, Indigenous overrepresentation has only increased, even as overall admissions were declining. See *R. v. Ipeelee* [2012] 1 SCR 433, at para. [62].

[16]Yang (2015).

[17]Grove and Meehl (1996).

[18]Natapoff (2012); Kohler-Hausmann (2018).

only part of the appeal because, even when challenging a decision does not change it, still it can be satisfying to have a venue in which to air one's grievances and concerns. Something like this thought may motivate the concern that computerised algorithms are not accountable in the way that human decision-makers are, since the algorithm either scores you (for instance) as high-risk or not. Even if it is still up to a judge or other human decision-maker to decide what to do with that score, the score itself has a sense of finality about it, even in cases where it is (or might be) mistaken.

A traditional way of expressing this concern is in the language of due process. Insofar as how an algorithm classifies you is impactful, particularly on traditional liberty interests, then you are entitled to a range of procedural rights – an oral hearing, to call witnesses, challenge the evidence against you, cross-examine and so forth. But algorithms deny you that process. They classify you based on information fed into it by a database – information that may potentially be erroneous or explained away. Consequently, due process requires giving you an opportunity to challenge how you are classified by an algorithm. This may include not only challenging the data on which the classification is based, but also challenging the design of the algorithm itself.

This is a powerful intuition. That said, the analogy has its limits. Consider how a similar analysis plays out in more traditional settings. Consider, once again, the bail context. In Canada, the law of bail requires prosecutors and jurists to gauge an arrestee's level of risk, but is silent as to how they discharge that task. In making an initial recommendation to a judge on bail, prosecutors are likely to rely on information provided by the police about the current allegations, as well as a criminal history, record of prior arrests or interactions with the complainant, supports available in the community, known substance abuse or mental health issues and so forth. If any of this information is believed to be erroneous, the accused, either personally or through counsel, is entitled to bring that potential error to the attention of the jurist, who ultimately decides whether or not to grant release. So, in this respect at least, it would not seem unusual to extend a similar procedural right if, for instance, jurists and prosecutors began relying on algorithmic risk assessments in making bail recommendations.

But the analogy only extends so far. The law of bail is, as I have noted, silent as to *how* prosecutors or jurists determine whether an accused presents a risk of reoffending or failing to attend court. One jurist might give substantial weight to a documented history of failing to appear or a lengthy criminal history, whereas another might attend more to whether an accused has community supports, housing and employment. One jurist might regard certain types of offence as more intolerable, and hence have a lower risk threshold, than another. Outside extreme and obvious idiosyncrasy, none of these currently raises due-process concerns. Indeed, given the broad discretion with which jurists make these decisions, and the limited documentation on how they make them, we essentially have no systematic knowledge about how Canadian jurists operationalise the bail statute's directive to gauge risk. As a result, the analogy to existing practices suggests that, while an accused may have a procedural right to challenge the reliability of the evidence upon which a risk-assessment device is based, he does not have a similar right to challenge the design parameters of the algorithm itself.

It is true that someone who is detained has the right to a review of the initial decision and that he can use that review as a basis for challenging errors in the original decision to order detention.[19] But those appeals are limited to errors of law, 'clearly inappropriate' bail judgments and cases in which there has been a material and relevant change in circumstances.[20] Those appeals do not serve to enforce substantive criteria about how jurists must structure their thinking about the various factors that they might deem relevant to risk. So, while a person who is classified as high-risk by a risk-assessment device would, presumably, continue to have the right to challenge the use that the jurist made of it, traditional notions of due process do not support the view that he has the right to contest the design parameters of the algorithm itself. This is because, within broad limits (having to do, for instance, with racial discrimination), neither statutory nor constitutional law mandates any particular means of assessing an accused person's risk.

[19] *Criminal Code*, R.S.C. 1985, s. 520.

[20] *R. v. St-Cloud*, 2015 SCC 27, [2015] 2 S.C.R. 328, at paras [120]–[121].

In any case, focusing on a due-process right to contest an algorithm's output asks too little of algorithms. The appeal of algorithmic decision-making in criminal justice contexts rests to a large degree on their promise to enhance outcomes by disciplining the intuition, experience and feeling of judges and lawyers by rigorous empirical methods. It would thus be a disappointment if all we could say about risk-assessment algorithms is that they are no worse than human judges. Rather, we should expect them to be substantially more reliable than prosecutors and judges acting on their own.

In contexts where decisions are left to the relatively unstructured discretion of a human decision-maker, there is some sense to providing an opportunity for adversarial disputation. However, we should not assume that adversarial disputation will continue to be equally valuable in contexts where predictive algorithms turn out to be substantially more reliable than human decision-makers. In those contexts, public accountability, in the sense of ensuring that decisions are as likely to be correct as we can manage, is probably not best fostered by having individual accused challenge the technical details of an algorithm in the course of their criminal proceedings. Neither lawyers nor judges, after all, are likely to have the requisite technical expertise to evaluate the statistical reliability of a risk-assessment instrument, nor is relying on the vagaries of the litigation and bargaining process a good means for securing consistent reliability in technical instruments. Accountability – in the sense of ensuring the ongoing reliability of an algorithmic instrument – is likely better served by placing this task in the hands of a specialised regulatory body with a specific mandate and the requisite technical expertise. In areas requiring highly technical expertise, regulatory oversight may prove to be a more satisfactory means of ensuring accountability than litigation by private parties before non-specialist courts.

5 Transparency as intelligibility

The final set of concerns I will address centre on the idea of intelligibility. I will consider two distinct concerns under this heading.

The first is an unease with adopting algorithmic decision-making in high-stakes areas like criminal justice because of how poorly understood the technologies in question are.

Once again, a few distinctions are helpful in getting a grip on the underlying intuition. Consider that many technologies, from aeroplanes to pharmaceuticals, involve processes that most people do not understand, although their lives depend upon those technologies doing what they are supposed to do. Thus, it cannot be a categorical objection to algorithmic decision-making that most people do not understand how they work. Nor can it be an objection to algorithmic decision-making that criminal justice is a high-stakes environment. A great deal of the technology we rely on in high-stakes environments are complex and far beyond the ken of most people. Few people who undergo heart surgery could, I suspect, really explain the fundamental principles of anatomy, biology, chemistry and medicine that such a procedure inevitably relies upon. Intelligibility must mean something different than intelligibility to most people, or even to those who are the subject of the technology in question. More plausibly, intelligibility is a matter of intelligibility to a range of experts – or, we might say, intelligibility in principle: in principle, if someone became an expert in this or that technology, she *would* understand its operation.

Does this interpretation provide a basis for objecting to algorithmic decision-making? Not categorically. While much of the attention surrounding algorithmic decision-making has to do with 'machine-learning' techniques such as neural networks, which do present questions of in-principle intelligibility, some algorithmic devices in use today in criminal justice contexts do not rely on machine-learning techniques. For instance, the risk-assessment tool developed by the Arnold Foundation is based on relatively straightforward regression models.[21] Hence, even if most people,

[21]For an overview of the Arnold Foundation's bail assessment tool, see *Public Safety Assessment: Risk Factors and Formula* (2016), available at https://craftmediabucket.s3.amazonaws.com/uploads/PDFs/PSA-Risk-Factors-and-Formula.pdf (accessed 6 March 2019).

including most judges, lawyers and accused persons, do not understand how regressions work, that is no objection to relying on a risk-assessment device of this kind.

Machine-learning techniques, neural networks in particular, raise a distinct set of concerns. Machine learning is 'atheoretical', in that a machine-learning algorithm 'learns' on its own to draw correlations between outcomes and inputs, including inputs that would not make much sense to a human.[22] In the case of aeroplanes, bridges and pharmaceuticals, even if lay persons do not understand how they work, still *experts* do. There are technicians, engineers and chemists who understand the mechanism. In contrast, in the case of a machine-learning algorithm, it may be the case that no one really understands the basis upon which it is drawing its correlations. Those correlations might be quite reliable, but it might be that no one is in a position to articulate quite *why* they are reliable and this surely does raise distinctive concerns about intelligibility.

Saying just what that concern is, in a way that does not ultimately boil down to more familiar concerns such as reliability and accuracy, is a bit trickier. Nevertheless, I do not deny that there is something unsettling about a technology that, while reliable, works on correlations that are potentially meaningless to humans. Insofar as this is an ethical objection to neural networks for high-stakes settings such as criminal justice, however, it is once again worth asking: compared to what? Even if we do not fully understand the correlations upon which a machine-learning algorithm bases its predictions, we should perhaps wonder how well we understand the reasons why people – including judges, prosecutors, police and others – make the predictions *they* do. To be sure, you can ask a judge to explain why she granted bail in this case but ordered detention in that one. But it would seem naive to think that the reasons people give publicly always overlap with the reasons that actually explain why they do what they do. Indeed, this would be not just naive, but ill-informed as well, as the disciplines of social psychology, behavioural economics and neuroscience are providing ever more sophisticated accounts of the gap between what people say, believe or experience and what they actually do. In other words, our understanding of why a human decision-maker decided a close case this way rather than that way is itself far from perfect, particularly once we discount that person's own subjective reports – even reports that take the form of *ex post* legal rationalisations.

In this connection, it is worth noting that machine-learning algorithms might be designed so as to ensure their continuing intelligibility to human operators, for instance by providing an account of its decisions at each layer of the network. The EU's General Data Protection Regulation includes, for instance, a right on behalf of someone affected by a decision based 'solely on automated processing' to 'obtain an explanation of the decision reached after such assessment and to challenge the decision'.[23] What specifically this right entails remains to be seen. It is possible that the explanations that are given might still be unintelligible to the vast majority of people, including the officials who rely on the algorithms and the people whose conduct they predict. But that would be unproblematic in much the same way as it is unproblematic for an airline passenger to be unable to explain the physics of flight or for a headache sufferer to be unable to explain the biochemistry of aspirin.

The second intelligibility-related concern has to do with the meaning of punishment. In particular, those who are drawn to communicative theories of punishment – most prominently defended by Antony Duff – might object that machine-learning algorithms cannot possibly serve the end of communicating to a wrongdoer the nature of her wrong.[24] Surely, communicating with a wrongdoer – expressing censure for violation of a public wrong, in the expectation (or at least hope) that the wrongdoer will understand and internalise the message – is an essentially human activity. As such, it would undermine the intelligibility of such a practice to rely in any significant way upon computerised algorithms in the criminal process.

There are at least three responses to this concern. First, even if the justification of punishment rests upon punishment's communicative function, punishment is far from the only function of the criminal

[22]Berk and Hyatt (2015).

[23]Council Regulation (EC) 2016/679 (GDPR) of 25 May 2018 on General Data Protection Regulation, Recital 71.

[24]See e.g. Duff (2007; 2018). This objection was suggested by a referee for the *International Journal of Law in Context*.

law. For instance, one of the most prominent areas in which algorithmic risk assessment has been deployed is in bail. As I have noted, bail is *already* an exercise in risk assessment. Similarly, predictive analytics are increasingly being used in policing, from predicting where crime will occur to predicting the identity of victims and perpetrators. Even if punishment is about communicating censure, policing and bail decisions cannot be analysed wholly in those terms, for decisions made at these stages of the criminal process occur prior to a finding of guilt for which anyone is properly held to account. Consequently, the question is not *whether* criminal justice officials should predict risk, but *how*. Unless risk assessment per se impairs the communicative potential of punishment, it is just not clear why technological innovation in how that risk assessment is performed should fundamentally alter the meaning of the criminal process.

Second, use of a predictive algorithm does not necessarily impair the communicative value of punishment. True, an algorithm that predicts, say, the likelihood of reoffending within a given time period may not further the communicative purpose of conveying to the accused the nature of the public wrong he had committed. However, not furthering the communicative purpose should not be conflated with frustrating it. Recall that Duff does not regard preventive rationales for punishment (such as deterrence) as per se incompatible with his communicative account.[25] Indeed, Duff goes further by conceding that a 'declarative criminal law … cannot plausibly be *merely* declarative' and must in addition accommodate 'forward-looking' reasons for punishment.[26] Thus, even predictive algorithms were defended on a purely preventive rationale that would not necessarily contravene a communicative theory of punishment.

Third, it is in any case not clear that a predictive algorithm must be defended on a purely preventive rationale. For instance, a predictive algorithm could be used not to predict whether an accused will reoffend, but instead to predict what a typical judge would regard as a proportionate sentence for a given case.[27] Since virtually all major theories of punishment, including communicative theories, endorse proportionality, an algorithm of this kind should hold broad appeal.[28] The function of an algorithm of this kind is akin to sentencing guidelines (whether mandatory, presumptive or advisory), in that it helps to minimise unwarranted arbitrariness. Minimising arbitrariness is not inconsistent with communicating a message to the accused. Indeed, the former may well serve to clarify the latter, by minimising the noisiness of the signal (i.e. arbitrary variations in sentence) as well as the possibility that the signal combines both legitimate messages of censure with illegitimate messages of invidious discrimination.

6 Conclusion

This may well come across as simply an apologia for relying upon computerised algorithms to substitute for human judgment at every possible juncture in the criminal process. That said, I believe there are serious ethical concerns regarding the use of algorithm decision-making in criminal justice. First among them is that it is far from clear that the private actors who are driving the development of most of these devices, from facial recognition to risk assessment to predictive policing software, have incentives that can be trusted to ensure that they reliably operate in the public interest. Criminal justice institutions largely interact with poorer and more marginalised sections of the population. This makes it harder to be sanguine about the incentives of market-based actors in societies where market power is quite unevenly distributed. While public policy in the criminal justice arena is subject to a wide range of familiar pathologies, one might worry that unregulated private development will be subject to its own pathologies. Rather than retreat from deploying algorithmic decision-making in criminal justice, this concern counsels public investment, either in developing the tools publicly or, perhaps more

[25]Duff (2018), pp. 22–23.
[26]*Ibid.*, at p. 208, emphasis in original.
[27]See Chiao (2018).
[28]Duff (2018), p. 38.

realistically, in some form of regulatory body to monitor the design, implementation and ongoing reliability of algorithmic devices.

Second, as I have been emphasising, improving upon the status quo in criminal justice – particularly in the US – is a low bar. This is because the status quo in criminal justice is deplorable. Criminal justice is perennially underfunded, its institutions are often reflexively resistant to change, subject to emotionally charged populist campaigns, and are shot through with unfair bias of all kinds. Although algorithmic decision-making holds out the promise of improving upon the status quo, one might reasonably expect more. Even if, say, a machine-learning algorithm could help us make sentencing less biased or bail decisions more accurate, still one might regard it as a missed opportunity if we could instead use similar means to intervene earlier, and in more constructive ways, so as to lessen the degree to which our social problems end up becoming criminal justice problems in the first place. Interventions that are much more targeted are also much more tractable, in the sense that it might become possible to provide a reasonable level of service (housing, counselling, substance-abuse management, separation from volatile social settings and so forth) when the population in question is defined more precisely. In this respect, it would be a grave ethical failing if our institutions treat technological innovation simply as a way of doing traditional criminal justice more effectively. Technological innovation should help us to minimise the adverse social impact of criminal justice, not simply make it more efficient.

One final note: the criminal law is often viewed as the area of law most directly tied to everyday moral reasoning, as evidenced by the significance of conceptions of blame, desert, responsibility, excuse and so forth in criminal courts. One might worry that the increased use of predictive algorithms, no matter how accurate, reliable and fair they become, would amount to turning criminal law and criminal justice over to technocrats and experts. One might regard this as a loss, for it would transform criminal law from the public re-enactment of a society's moral habitus into the coldly calculating work of minimising net social harm. For reasons sketched in this paper, I suspect that concerns that technological innovation will make criminal law unaccountable and unintelligible are exaggerated. However, technological innovation could make criminal law and criminal justice less sensitive to popular emotion and more sensitive to expertise and evidence.[29] In reflecting upon this possibility, it is worth recalling that our existing institutions are built out of line officials exercising their discretion in a largely uncoordinated manner, typically in ways more reliant on experience and intuition than evidence and analysis. Our existing criminal justice institutions also frequently operate in dysfunctional and counter-productive ways. This should weigh heavily against being too optimistic that the moral judgment of lawyers and judges in an unstructured, case-by-case manner will bring about meaningful reform. To the contrary, it may well be that the refusal to discipline the judgment of individual actors by systematic and empirically tested criteria is itself a significant source of injustice and arbitrariness in the criminal law.

Acknowledgements. I am grateful to Anne Marshall for research assistance with this paper.

References

Berdejó C and Yuchtman N (2013) Crime, punishment and politics: an analysis of political cycles in criminal sentencing. *Review of Economics and Statistics* **95**, 741–756.

Berk R and Hyatt J (2015) Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter* **27**, 223.

Berk R *et al.* (2018) Fairness in criminal justice risk assessments: the state of the art. *Sociological Methods and Research*, July. Available at https://doi.org/10.1177/0049124118782533 (accessed 8 March 2019).

Chiao V (2018) *Criminal Law in the Age of the Administrative State*. Oxford: Oxford University Press.

Chiao V (2018) Predicting proportionality: the case for algorithmic sentencing. *Criminal Justice Ethics* **37**, 238–261.

Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* **5**. Available at https://doi.org/10.1089/big.2016.0047 (accessed 19 February 2019).

[29]I do not think, however, that this is necessarily to make criminal law less democratic. See Chiao (2018), Chapter 3.

Danziger S *et al.* (2011) Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 6889–6892.
Duff A (2007) *Answering for Crime*. Oxford: Oxford University Press.
Duff A (2018) *The Realm of Criminal Law*. Oxford: Oxford University Press.
Englich B (2009) Heuristic strategies and persistent biases in sentencing decisions. In Oswald ME, Bieneck S and Hupfeld-Heinemann J (eds), *Social Psychology of Punishment of Crime*. Hoboken: Wiley-Blackwell, pp. 295–314.
Goodman-Delahunty J and Sporer SL (2010) Unconscious influences in sentencing decisions: a research review of psychological sources of disparity. *Australian Journal of Forensic Sciences* **42**, 19–36.
Grove W and Meehl P (1996) Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical-statistical controversy. *Psychology, Public Policy, and Law* **2**, 293–323.
Hannah-Moffat K (2012) Actuarial sentencing: an 'unsettled' proposition. *Justice Quarterly* **30**, 1–27.
Harcourt B (2006) *Against Prediction: Profiling, Policing, and Punishing in An Actuarial Age*. Chicago: University of Chicago Press.
Heidensohn F and Silvestri M (2012) Gender and crime. In Maguire M, Morgan R and Reiner R (eds), *The Oxford Handbook of Criminology*. Oxford: Oxford University Press.
Huq A (2019) Racial equity in algorithmic criminal justice. *Duke Law Journal* **68**, forthcoming.
Kleinberg J *et al.* (2018) Human decisions and machine predictions. *Quarterly Journal of Economics* **133**, 237–293.
Kohler-Hausmann I (2018) *Misdemeanorland: Criminal Courts and Social Control in an Age of Broken Windows Policing*. Princeton: Princeton University Press.
Mitchell O and Caudy MS (2017) Race differences in drug offending and drug distribution arrests. *Crime & Delinquency* **63**, 91–112.
Natapoff A (2012) Misdemeanors. *Southern California Law Review* **85**, 101–163.
Nellis A (2016) The color of justice: racial and ethnic disparity in state prisons. The Sentencing Project, Washington, DC. Available at http://www.sentencingproject.org/wp-content/uploads/2016/06/The-Color-of-Justice-Racial-and-Ethnic-Disparity-in-State-Prisons.pdf (accessed 19 February 2019).
Owusu-Bempah A and Wortley S (2014) Race, crime, and criminal justice in Canada. In Bucerius S and Tonry M (eds), *Oxford Handbook of Ethnicity, Crime, and Immigration online*. Oxford: Oxford University Press.
Silvestri M and Crowther-Dowey C (2008) *Key Approaches to Criminology: Gender & Crime*. Los Angeles: SAGE Publications.
Starr SB (2014) Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review* **66**, 803–872.
Substance Abuse and Mental Health Services Administration (2014) *Results from the 2013 National Survey on Drug Use and Health: Summary of National Findings*, NSDUH Series H-48, HHS Publication No. (SMA) 14–4863. Rockville, MD: Substance Abuse and Mental Health Services Administration.
Yang C (2015) Free at last? Judicial discretion and racial disparities in federal sentencing. *Journal of Legal Studies* **44**, 75–111.

Cite this article: Chiao V (2019). Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice. *International Journal of Law in Context* **15**, 126–139. https://doi.org/10.1017/S1744552319000077