Algorithmic risk assessment tools in criminal justice: the need for better data[i]

Thomas Douglas,[1] Benjamin Davies,[1] Jonathan Pugh[1], Rebecca Brown,[1] Binesh Hass,[1] Lisa Forsberg,[1] Abhishek Mishra,[1] Ilina Singh,[2] Julian Savulescu,[1] Seena Fazel[2,3]

1. Oxford Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, Suite 8, Littlegate House, St Ebbes Street, Oxford OX1 1PT, United Kingdom
2. Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford OX3 7JX, United Kingdom
3. Oxford Health NHS Foundation Trust, Warneford Hospital, Oxford OX3 7JX, United Kingdom

Introduction and Summary

[1] We are academics from the University of Oxford specialising in criminal justice ethics, the ethics of emerging technologies, developmental psychology, and forensic psychiatry.

[2] This evidence submission addresses several questions posed in the call for evidence, namely those on the aims and acceptable uses of new technologies (Question 2); on the reliability of outputs from new technologies (Question 3); on equality, and on trust in the rule and application of law and equality (Question 4), on weighing potential costs and benefits, and employing safeguards (Question 5); and on the evaluation and transparency of new technologies (Question 6).

[3] It builds on work previously published by some of the authors in *European Psychiatry.*[ii]

[4] Our evidence considers whether the accuracy of algorithmic risk assessment tools is sufficiently understood to forestall serious ethical problems. We argue that there are current deficiencies in the evidence base and that these hamper efforts to (a) interpret risk scores with appropriate caution, (b) match different algorithmic tools to different applications, and (c) avoid discrimination.

[5] We recommend that these deficiencies be remedied as soon as possible, through conducting high quality studies on samples that represent the population of arrestees, defendants and offenders in the United Kingdom. These studies should employ transparent methods, present all relevant measures of accuracy, and investigate the degree to which data

[i] Declaration of Interest: SF has published research on risk assessment, including as part of a team that has derived and validated one tool for prisoners with psychiatric disorders.

[ii] Douglas T, Pugh J, Singh I, Savulescu J, Fazel S, Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data, *European Psychiatr*y 2017;42:134–7.

types included in current tools (a) add incremental validity to tool performance and (b) track membership of disadvantaged groups.

[6] The government should introduce rigorous procedures for assessing these studies and other evidence on tool performance before employing tools in the criminal justice system. These assessments should include consideration of the broader social harms from the use of (potentially) discriminatory tools. They should compare risk assessment tools with alternative algorithmic tools, and alternative non-algorithmic forms of risk assessment (such as structured or unstructured professional judgment).

Algorithmic risk assessment tools in criminal justice

[7] More than 200 algorithmic tools are currently available for assessing risk of violence in criminal justice.[iii] These are widely deployed to inform initial sentencing, parole decisions, and decisions regarding post-release monitoring, management and rehabilitation. In addition, violence risk assessment tools are used outside of criminal justice, for example, to inform decisions regarding detention, discharge, and patient management in forensic and, increasingly, general psychiatry.

Dimensions of predictive accuracy

[8] For the use of a risk assessment tool to be ethically acceptable, the tool must be sufficiently accurate at predicting the outcome of interest—often recidivism.[iv] Both absolute accuracy and accuracy relative to alternative means of risk assessment, such as structured or unstructured professional judgment, are important.

[9] There are many different dimensions of accuracy, and different tools perform better on different dimensions. There are thus ethical trade-offs to be made between different types of accuracy.

[10] Four of the most important dimensions of accuracy are *sensitivity*, *specificity, positive predictive value* and *negative predictive value.* These relate to the tool's ability to detect *true positives* and *true negatives*, and avoid *false positives* and *false negatives*.

True positive	Tool correctly identifies the	E.g. Identifies as 'high risk' someone who would go on to

[iii] Singh JP, Desmarais SL, Hurducas C, Arbach-Lucioni K, Condemarin C, Dean K, et al., "nternational Perspectives on the Practical Application of Violence Risk Assessment: A Global Survey of 44 Countries, *International Journal of Forensic Mental Health* 2014;13:193–206.

[iv] European Commission, 'REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL: LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS', April 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206; Lo Piano S, Ethical Principles in Machine Learning and Artificial Intelligence: Cases from the Field and Possible Ways Forward, Humanities and Social Sciences Communications 2020;7(1):1–7, https://doi.org/10.1057/s41599-020-0501-9; The Institute of Electrical and Electronics Engineers Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically Aligned Design: Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, 2019.

	relevant feature as present	recidivate
False positive	Tool incorrectly identifies the relevant feature as present	E.g. Identifies as 'high risk' someone who would not go on to recidivate
True negative	Tool correctly identifies the relevant feature as absent	E.g. Identifies as 'low risk' someone who would not go on to recidivate
False negative	Tool incorrectly identifies the relevant feature as absent	E.g. Identifies as 'low risk' someone who would go on to recidivate

[11] Suppose a tool is intended to predict recidivism. The *sensitivity* of the tool is the proportion of individuals who will *in fact* recidivate that the tool predicts will recidivate (typically, this would mean that they are classified as 'high risk'). Other things being equal, the more sensitive a tool, the less likely it is to incorrectly deliver a negative result (a 'false negative'). A false negative here would be a 'low risk' result assigned to an individual who in fact recidivates.

[12] The tool's *specificity* is the proportion of individuals who will not recidivate who are correctly predicted not to recidivate (e.g., classified as 'low risk'). Other things being equal, the more specific the tool is, the less likely it is to incorrectly deliver a positive result (a 'false positive', e.g. a 'high risk' result assigned to an individual who will not in fact recidivate).

[13] *Sensitivity* and *specificity* are measures of how good the tool is at picking out those in a population who will recidivate or not recidivate. By contrast, the predictive value of a tool is a measure of how likely its predictions are to prove true. The predictive value of the tool depends on the baseline rates of re-offending in the tested population. A high risk classification from even a highly specific and sensitive system may have low predictive value if there is a very low rate of re-offending in the population under assessment.

[14] Two types of predictive value are of interest: positive and negative. In the present context, the positive predictive value is the proportion of individuals classified as high risk who will in fact recidivate. A high positive predictive value implies that few 'high risk' classifications will be false positives. The negative predictive value is the proportion of individuals who are classified as 'low risk', and who will not recidivate. A high negative predictive value implies that few 'low risk' classifications will be false negatives.

[15] An additional layer of accuracy that is typically ignored and not reported is calibration, which is a measure of how close the predicted rates of recidivism (based on a risk assessment tool) are to the actual rates of recidivism. This is different to the above measures, which use risk categories such as 'high' or 'low'. A tool may be able to perfectly differentiate between low and high risk persons, but still be systematically off target. For example, it may classify

all high risk persons as having a greater than 80% chance of reoffending within two years, when in fact it the rate is 50%. These persons are still high risk in the sense that they reoffend at higher rates than the 'low' risk group, but their probability score is incorrect. This accuracy measure—calibration—needs to be reported in addition to the measures described above. In assessing accuracy, it is crucial to compare algorithmic tools with the alternatives. In many cases, the alternative will be an unstructured professional judgment by, for example, a judge or psychiatrist. Such judgements are known to be affected by biases and inaccuracies that are typically greater than those which afflict algorithmic tools.[v]

Evidence on accuracy

[16] Most risk assessment tools have poor positive predictive value in most applications. Typically, positive predictive value is less than 0.5, meaning that more than half of individuals classified by tools as high risk are false positives—they will not go on to offend.[vi] These persons may be detained unnecessarily.

[17] Negative predictive value is normally higher, often over 0.9. Nevertheless, in typical cases around 9% of those classed as low risk will go on to offend (corresponding to a negative predictive value of 0.91 or 91%).[vii] These individuals may be released too early, given the risk they in fact pose. Such false negatives are frequently associated with significant controversy and outrage, as reaction to high profile cases demonstrates.[viii]

[18] In addition to being poor, the predictive value of most tools is difficult to assess in advance, making it difficult to appropriately account for the risk of false positives and false negatives in interpreting risk scores. Published validation findings for the most widely used tools frequently present a misleading picture.[ix] First, most tools have not been evaluated outside their derivation sample.[x] Consequently, it is unclear how far reported accuracy findings can be extrapolated to new settings and populations.[xi] Second, there is strong

---

[v] Desmarais SL, The Role of Risk Assessment in the Criminal Justice System: Moving Beyond a Return to the Status Quo, *Harvard Data Science Review* 2020;2, https://doi.org/10.1162/99608f92.181cd09f.

[vi] Fazel S, Singh JP, Doll H, Grann M,. Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24 827 People: Systematic Review and Meta-Analysis, *BMJ* 2012;345:e4692.

[vii] Fazel S, Singh JP, Doll H, Grann M, Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24 827 People: Systematic Review and Meta-Analysis, *BMJ* 2012;345:e4692.

[viii] E.g., Parry H, Rapist Released Halfway through Sentence Went on to Attack Three More While on Parole Including a Schoolgirl Who Was Raped in Front of Her Boyfriend, *Daily Mail* 2015, http://www.dailymail.co.uk/news/article-3131895/Rapist-released-halfway-sentence-went-attack-three-parole-including-two-schoolgirls-raped-boys-with.html.

[ix] Shepherd SM, Sullivan D, Covert and Implicit Influences on the Interpretation of Violence Risk Instruments, *Psychiatry, Psychology and Law* 2016; doi: 10.1080/13218719.2016.1197817.

[x] Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA, External Validation of New Risk Prediction Models Is Infrequent and Reveals Worse Prognostic Discrimination, *Journal of Clinical Epidemiology* 2015;68:25–34. Siontis and collaborators found that only 16% of new prediction models are validated by different authors within five years of their first use, and when external validation of tools did occur, predictive accuracy was generally reduced.

[xi] A recent Chinese review found that instruments developed in the West had significantly lower predictive accuracy when used in China compared to that reported for Western populations. See Zhou

evidence that conflicts of interest are often not disclosed, and some evidence of publication and authorship bias (research on tools being published by the authors of those tools).[xii] It is notable that the tool most commonly used in the UK, Offender Assessment System or OASys, has not been the subject of independent evaluation to our knowledge.

[19] As well as hindering attempts to interpret risk scores in light of their likely accuracy, this limited and skewed evidence base generates two further problems: it thwarts attempts to match risk assessment tools to different contexts of application, and it exacerbates the risk of an objectionable form of demographic profiling. We explain these problems in the subsequent two sections.

The right tool for the context

[20] Selecting the optimal risk assessment tool for a given application requires trade-offs to be made between false negatives and false positives; attempts to reduce the number of false positives will increase the number of false negatives.[xiii] Tools with a low rate of false negatives will be most effective at preventing additional violence and crime, and may garner most political support, while tools with a low rate of false positives (due to high specificity) will best protect the rights and interests of defendants, convicted offenders, and prisoners.

[21] The appropriate balance between false positives and false negatives depends on the stage in the criminal justice process at which the tool is deployed. For instance, a risk assessment tool may be used to inform decisions about whether to impose additional 'post-sentence detention' on an individual who has already served out a sentence that was proportionate to the seriousness of their crime. Since in this context, further detention will be disproportionate, special care should be taken to avoid false positives—which here would mean imposing post-sentence detention on an underserving individual, and there are grounds to select a tool with a very low false positive rate.

[22] The situation is different when a tool informs parole decisions. In this context, false positives—i.e. incorrect 'high risk' classifications—may lead to refusal of parole and an unnecessarily long period of incarceration from the perspective of public protection. Yet if the initial sentences are themselves proportionate, then the overall period of detention for 'false positive' individuals will remain within the limits required by proportionality. In this context it may be more important to avoid false negatives.

[23] Matching risk assessment tools to different contexts of application thus requires trade-offs between false positives and false negatives. For each context, we must first decide which type of error to prioritise to which degree, and then select a tool reflecting this priority. Unfortunately, in the absence of reliable data on predictive accuracy, it is not possible to make the latter decision confidently.

J et al., Violence Risk Assessment in Psychiatric Patients in China: A Systematic Review, *Australian and New Zealand Journal of Psychiatry* 2015;50:33-45.

[xii] Singh JP, Grann M, Fazel S, Authorship Bias in Violence Risk Assessment? A Systematic Review and Meta-Analysis, *PLoS ONE* 2013;8:e72484.

[xiii] Walker N. Dangerousness and Mental Disorder. *Royal Institute of Philosophy Supplement* 1994;37:179-190, doi: 10.1017/S1358246100010055, at 182.

Discrimination

[24] Singling out individuals for unfavourable treatment on the basis of protected characteristics, such as race and sex, amounts to discrimination. In the UK, the Equality Act 2010 distinguishes direct and indirect discrimination.[xiv] Direct discrimination occurs when:

> because of a protected characteristic, A treats B less favourably than A treats or would treat others.[xv]

[25] Indirect discrimination, on the other hand, occurs when, though there is no direct discrimination:

> A applies to B a provision, criterion or practice which is discriminatory in relation to a relevant protected characteristic of B's.[xvi]

Typically, a practice that is not directly discriminatory will nevertheless qualify as indirectly discriminatory when its burdens fall disproportionately on bearers of a particular protected characteristic.

[26] Where risk assessment tools are used to justify unfavourable treatment, such as refusal of bail or parole, they will be directly discriminatory when they employ protected characteristics as explicit predictors.

[27] Risk assessment tools could theoretically exclude such characteristics and thus avoid direct discrimination. However, they may still indirectly discriminate, if bearers of certain protected characteristics are disproportionately negatively impacted by use of the tool. This risk is particularly significant for machine learning-based ('ML') tools, as these may 'learn' associations between protected and unprotected characteristics and so functionally mimic a tool which explicitly employed protected characteristics. For instance, where there is some correlation between race and reported recidivism, an ML tool that has race excluded from its training data may still disproportionately impact particular racial groups, and perhaps to the same degree as a tool that explicitly employs race as a predictor.[xvii]

[28] That a tool has a disproportionate negative impact upon bearers of a protected characteristic is not always a decisive ethical reason against it, especially when it will also do considerable good. However, where burdens fall disproportionately on already disadvantaged groups, including exacerbating existing burdens such as racial discrimination, this is a considerable ethical cost that makes them very unlikely to be justified.

[29] Moreover, when the workings of a risk assessment tool are opaque to the public, and the tool is observed to assign higher risk scores to bearers of certain protected characteristics,

---

[xiv] Equality Act 2010, pt 2, ch 1, s 4. Protected characteristics are: age; disability; gender reassignment; marriage and civil partnership; pregnancy and maternity leave; race; religion or belief; sex; and sexual orientation.

[xv] ibid pt 2, ch 2, s 13.

[xvi] ibid pt 2, ch 2, s 19.

[xvii] Davies B, Douglas T. Learning to Discriminate: The Perfect Proxy Problem in Artificially Intelligent Criminal Sentencing, forthcoming in Roberts J, Ryberg J (eds), *Principled Sentencing and Artificial Intelligence* (Forthcoming, Oxford University Press).

they may create the *impression* of direct discrimination. Given the pervasiveness of kinds of discrimination in society (e.g. racial discrimination), these impressions may be reasonable even if inaccurate, and can create distrust of the state and law enforcement, and social and political alienation.[xviii] In the UK, any policy whose outcome is that only persons of colour are targets of screening for criminality is liable to be reasonably perceived as racist.

[30] To mitigate the risk of indirect discrimination, and of the reasonable impression of direct discrimination, risk assessment tools should be designed and used in a way that avoids imposing burdens that disproportionately affect already disadvantaged or oppressed groups unless this confers benefits, in the form of crime prevention, that are considerably greater than those conferred by alternative approaches, and the disadvantaged group shares equitably in that benefit. This could be achieved by either (a) ensuring that individuals assigned high risk scores are not subjected to greater burdens, overall, than those assigned lower risk scores, for example by fully compensating individuals who are subjected to additional liberty-restricting measures on the basis of their high risk classification, (b) employing de-biasing methods to ensure that the use of the tool does not disproportionately negatively impact members of already disadvantaged or oppressed groups unless this can be justified by a very large payoff in crime prevention,[xix] or (c) taking separate measures to address the oppression and disadvantage, which might include, for example, training and rehabilitation programmes targeted at the disadvantaged group, and intensive involvement of the group in the design of criminal justice policies and risk assessment tools. Unfortunately, lack of data currently makes it impossible to perform the balancing required for (b). What is required is data on the degree to which inclusion of a particular data type in a predictive model will (i) incrementally improve the predictive value of the model in the population to which it is applied, and (ii) incrementally increase the degree to which risk scores track membership of disadvantaged groups. Such data is not currently available. Acquiring it will be particularly challenging for ML tools, since, in these tools, the incremental impact of different data types is often difficult to interpret using standard statistical techniques.[xx]

Conclusion and recommendations

[31] The government should introduce rigorous procedures for assessing the evidence on tools before employing them in the criminal justice system.

[32] To improve the evidence base available to inform these assessments, the government should support high quality research into the performance of risk assessment tools. This research should assess (a) all relevant measures of accuracy, and (b) the degree to which data types included in tools add incremental validity to tool performance and track membership of disadvantaged groups. Studies should employ transparent methods samples that represent the population of the United Kingdom.

[xviii] Hosein AO. Racial Profiling and a Reasonable Sense of Inferior Political Status. *Journal of Political Philosophy* 2018;26: e1-e20.

[xix] De-biasing methods might include, for example, employing different risk thresholds for restricting liberty in different racial groups.

[xx] Rudin C, Wang C, Coker B, The Age of Secrecy and Unfairness in Recidivism Prediction, *Harvard Data Science Review* 2020;2, https://doi.org/10.1162/99608f92.6ed64b30.

[33] Assessments of algorithmic tools should include consideration of the broader social harms from the use of (potentially) discriminatory tools, including stigmatisation of racial minorities, and reasonable perceptions by members of racial minorities that they are being unfairly singled out.

[34] Since decisions on tool adoption require ethical trade-offs, for example, between different types of accuracy, and assessments of broader social impacts, for example, on disadvantaged groups, assessors should include individuals with ethical and social scientific expertise and members of the public, in addition to legal, clinical, forensic and statistical experts.

[35] Tools proposed for use should be assessed against the best available alternative means (algorithmic or otherwise) for assessing risk.