Artificial intelligence, risk assessment, and potential racial implications

Pamela Ugwudike from the University of Southampton examines the role of Artificial Intelligence in probation and potential racial implications.

https://doi.org/10.54006/PANK7644

'Artificial intelligence: Algorithms face scrutiny over potential bias' (BBC 2019)

'Rise of the racist robots – how AI is learning all our worst impulses' (The Guardian Newspaper 2017)

'How AI Could Reinforce Biases in The Criminal Justice System' (CNBC 2020)

'The Rise—and the Recurring Bias—of Risk Assessment Algorithms' (The Markup 2021)

Pamela Ugwudike
University of Southampton

These and similar headlines are increasingly drawing attention to the potential racial implications of risk assessment tools and other predictive technologies deployed in contemporary justice systems. The tools are sometimes described as Artificial Intelligence systems in line with the current usage of the term 'AI' which broadly refers to a machine or computer programme trained to perform tasks which rely on human intelligence. One such task is learning how to use information from the past to try and predict the future. AI is therefore, 'about machines which act intelligently – typically making predictions or decisions about multiple aspects of the world in which we live' (Weller 2021).

The deployment of AI for risk assessment in criminal justice systems primarily involves using algorithms (e.g. logistic regressions) to statistically analyse administrative and other datasets, in order to predict recidivism risks in individual cases. Perhaps for this reason, risk assessment tools are also now commonly referred to as 'risk assessment algorithms'.
Broadly defined, an algorithm is "a self-contained step-by-step set of operations that computers and other 'smart' devices carry out to perform calculation, data processing, and automated reasoning tasks," (Association for Computing Machinery (ACM) US Public Policy Council and ACM Europe Council (2017).

Some of the risk assessment algorithms that have been deployed in recent years possess machine learning capabilities in that they can be trained using advanced statistical techniques and training data, to identify patterns in new datasets and predict recidivism risks (Berk and Bleich 2013; Berk 2021). Examples of risk assessment algorithms include the HART (Harm Assessment Risk Tool) machine learning model which has been deployed in the UK (Oswald et al. 2018). Additional examples include the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm (Brennan et al. 2009) which is used by prison and probation services in some parts of the US and the Offender Assessment System (OASys) in the UK (see, Her Majesty's Inspectorate of Probation – HMIP 2020).

Alongside numerous media reports highlighting the capacity for risk assessment algorithms to foment racial discrimination, a fast-growing multidisciplinary scholarship on the problem now exists, emerging from areas such as criminology, law, and the broad field of Science and Technology Studies (STS) (Bao et al. 2021; Green and Chen 2019; Hannah-Moffatt 2018; Kehl et al. 2017; Lavorgna and Ugwudike 2021; Starr, 2014). This scholarship along with negative press releases may be penetrating public consciousness and undermining trust in the systems. Indeed, government bodies (e.g., CDEI 2019, 2020) civil society organisations (e.g., AI Now 2018) and others have acknowledged the problem of potential bias.

There are several variants of the algorithms in question but fundamentally, the generic tools that are used to assess most people coming into the justice system perform a predictive function. This involves identifying patterns in historical data to make generalisations about an individual's risk based on the characteristics (defined as risk predictors) they share in common with others, typically criminal justice populations. Commonly cited risk predictors include criminal history, educational attainment, employment history and family circumstances (see, Hamilton 2015). Though conceptualised by the developers of risk assessment algorithms as risk predictors, if viewed through a socially conscious lens these indicators could just as easily be understood as adverse outcomes which have their roots systemic problems such as racial discrimination and other forms of structural disadvantage.

The risk assessment process yields risk scores and categories that can inform degrees and types of penal intervention although variants of risk assessment algorithms known as structured tools do permit a degree of contextualised clinical assessment based on professional discretion in each case (HMIP 2020). Ultimately, algorithmically generated risk scores influence penal outcomes.

Origins of risk assessment: A brief overview

The practice of forecasting recidivism risks on the basis of historical factors and placing people in risk categories that determine levels of penal intervention is by no means novel. As far back as the 19th century, people coming into contact with the justice system were exposed to various forms of individualised or personalised assessment for transformational or reformative intervention (see, Vanstone 2004). These were clinical assessments based mainly on professional judgement although predictions of probable reoffending were also made as far back as the early 20th century to determine parole outcomes (e.g., Burgess 1928). Some argue that risk assessments have since shifted from individualised analysis of treatment needs to actuarial prediction. This technique is said to support the allocation of risk subjects to statistically defined categories or 'risk pools' for cost-effective and efficient risk management (e.g., Feeley and Simon 1992).

The problem of algorithmic bias

More advanced risk assessment technologies including machine learning variants have since emerged, provoking new concerns. Commonly cited problems include predictive accuracy, bias, and limited transparency (Berk 2021). Here, I focus on the issue of potential racial bias which can occur when the algorithms rely on certain types of data such as administrative datasets from some law enforcement services. This problem has been brought to light by several studies. To cite an example, in 2016, ProPublica (an organisation that specialises in investigative journalism), conducted a study of the COMPAS risk assessment algorithm. Their analysis identified evidence of racial disparities in the form of over-prediction (high rates of false positives) in cases involving Black people. Other studies have shown that this potentiates more punitive penal intervention (e.g., Lowder et al. 2019).

Such artificial inflation of risk can occur because Black people have worse criminal justice outcomes (e.g., arrests) and biased decision making cannot be ruled out as a possible cause (Shiner et al. 2018). More individuals in the group would therefore be vulnerable to false positives since they belong to a group with qualities (e.g., higher arrests rates) that algorithms have been programmed to interpret as risk predictors (see also, Hao and Stray 2019). In other words, Black people will have greater odds of being misclassified by the algorithm as higher risk than they are because of the racial group to which they belong. Commenting on this problem, Vincent and Viljoen (2020: 1576) note that, 'if some groups get apprehended more, those groups will score higher on non-biased, well-validated instruments derived to maximize prediction of recidivism because of mathematics'. Therefore, the problem of risk inflation occurs even if the algorithm attains predictive parity in the sense that it predicts risks with the same level of accuracy across all subgroups, and most of those predicted to reoffend do so regardless of protected or sensitive attributes such as race.

This reveals the potential for administrative datasets to foment racially biased algorithmic outcomes. But apart from criminal history, some of the other predictors on which commonly used, generic, risk assessment algorithms rely, can provoke similar outcomes. Consider for example, the predictors 'employment and education'. Black people can be more disadvantaged by these predictors than other groups. As official statistics in the UK for example reveal, they are more vulnerable to expulsion from school (Department of Education 2016) and stable employment (Office for National Statistics 2011). Thus, along with criminal history predictors, socioeconomic predictors can operate as proxies for race.

It is also worth noting that socioeconomically marginal groups in general can be disadvantaged if the algorithms are programmed or trained to interpret their adverse circumstances as individual deficiencies warranting high risk scores and penal intervention, instead of structural problems requiring social welfare provision. As van Ejick (2016) notes in an analysis of commonly used risk assessment algorithms, predictors based on socioeconomic circumstances can foment the discriminatory criminalisation of poverty and disadvantage people from deprived communities.

Black people can be further disadvantaged where the predictor 'family circumstances' is operationalised as parental involvement in the justice system. Given their aforementioned over-representation in criminal justice statistics, such a predictor can constitute a proxy for race, exposing them to more punitive intervention because they belong to a specific racial group (see also, Harcourt 2015).

Conclusion

This paper has drawn attention to several ethical issues that touch upon the racial and broader social implications of deploying risk assessment algorithms in justice systems. As debates and studies focusing on the use of algorithms in probation and across justice systems continue to expand, a growing consensus seems to be that remedial strategies are required to address the potential for the algorithms to reproduce historical forms of discrimination. In response, some have developed mitigating techniques, which for example, attempt to debias datasets and limit their capacity to operate as proxies for race (e.g., Skeem and Lowenkamp 2021). Additionally, a multidisciplinary field of AI ethics has emerged to highlight the importance of internal and independent audits for identifying and mitigating biases whilst embedding ethical principles into algorithm design and application (e.g., Raji et al. 2020). Some contend that a robust legal framework is also urgently needed to regulate AI design and deployment (e.g., Favaretto et al. 2019).

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed June 2018).

ACM US Public Policy Council and ACM Europe Council (2017) Statement on Algorithmic Transparency and Accountability https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf (accessed March 2019).

AI Now (2018). Algorithmic Accountability Policy Toolkit: https://ainowinstitute.org/aap-toolkit.pdf (accessed May 2019).

Bao, M. et al. (2021), It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks https://arxiv.org/abs/2106.05498 (accessed October 2021).

Berk R. A. (2021) Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement. Annu. Rev. Criminol. 4:209–37.

Berk, R. A. and Bleich, J. 2013 Statistical Procedures for Forecasting Criminal Behaviour: A Comparative Assessment. Criminology & Public Policy. 12 (3) 513-544

Brennan T, Dieterich W and Ehret B (2009) Evaluating the predictive validity of the COMPAS risk and needs assessment system. Criminal Justice and Behaviour, 36, 21-40.

Burgess, E. M. (1928) Factors determining success or failure on parole. In A. A. Bruce, A. J. Harno, E. W. Burgess, E. W. Landesco (eds) *The workings of the indeterminate-sentence law and the parole system in Illinois. Springfield: Illinois State Board of Parole.* Pp.205-249.

CDEI (2019) Centre for Data Ethics and Innovation's approach to the governance of data- driven technology. https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovations-approach-to-the-governance-of-data-driven-technology

CDEI (2020) Review into bias in algorithmic decision-making. https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making

Department of Education (2016) Permanent and fixed period exclusions in England: 2014 2015. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/539704_SFR_26_2016_text.pdf#page=6 (accessed 25th July 2018).

Favaretto M, De Clercq E and Elger BS (2019) Big Data and discrimination: Perils, promises and solutions. A systematic review. Journal of Big Data 6(1): 1-27.).

Feeley, M. and Simon, J. (1992) The new penology: Notes on the emerging strategy of corrections and its implications. Criminology 30, 449–74.

Green, B. Chen, Y. (2019) Disparate interactions: An algorithm-in-the- loop analysis of fairness in risk assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 90–99.

Hamilton M (2015) Risk-Needs Assessment: Constitutional and Ethical Challenges, American Criminal Law Review 231, 236–39.

Hannah-Moffat K (2018) Algorithmic Risk Governance: Big Data Analytics, Race and Information Activism in Criminal Justice Debates. Theoretical Criminology 1-18.

Hao K and Stray J (2019) Can you make AI fairer than a judge? Play our courtroom algorithm game. MIT Technology Review.

Harcourt, B. E. (2015) Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, 27(4):237-243.

HMIP (2020) Assessment - Criminal Justice Inspectorates https://www.justiceinspectorates.gov.uk/hmiprobation/research/the-evidence-base-probation/supervision-of-service-users/assessment/

Kehl D, Guo P and Kessler S (2017) *Algorithms in the criminal justice system: Assessing the Use of Risk Assessments in Sentencing. Responsive Communities Initiative,* Berkman Klein Centre for Internet & Society: Harvard Law School.

Office for National Statistics (2011) 2011 Census analysis: Ethnicity and the Labour Market, England and Wales. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/ethnicityandthelabourmarket2011censusenglandandwales/2014-11-13 (accessed 16 May 2018).

Oswald, M., Grace, J., Urwin, S. and Barnes, G.C. (2018) 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and "Experimental Proportionality"', Information and Communications Technology Law, 27:223-250.

Raji, I. D. Smart, A. et al. (2020) Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA.

Shiner, M., Carre, Z., Delsol, R. Eastwood, N. (2018) The Colour of Injustice: 'Race', drugs and law enforcement in England and Wales. https://www.lse.ac.uk/united-states/Assets/Documents/The-Colour-of-Injustice.pdf Accessed February 2019.

Skeem, J. and Lowenkamp, C. (2020) Using algorithms to address trade-offs inherent in predicting recidivism. Behav. Sci Law 38(3):259-278.

van Eijk, G. (2016). Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality. Punishment and Society 19463-481.

Vanstone M (2004) Supervising Offenders in the Community: A History of Probation Theory and Practice. Aldershot: Ashgate.

Vincent, G. M., & Viljoen, J. L. (2020). Racist Algorithms or Systemic Problems? Risk Assessments and Racial Disparities. Criminal Justice and Behaviour, 47(12), 1576-1584. https://doi.org/10.1177/0093854820954501

Weller, A. (2021) What does AI mean for the Turing? https://www.turing.ac.uk/news/what-does-ai-mean-turing