The Missteps of the FIRST STEP Act: Algorithmic Bias in Criminal Justice Reform

Raghav Kohli[1]

I. Introduction

Contrary to his tough-on-crime rhetoric, Donald Trump in December 2018 signed the FIRST STEP Act[2] (the '**Act**') into law, a criminal justice reform legislation aimed at reducing recidivism and reforming prison and sentencing laws.[3] With a 87-12 vote in the Senate and a 358-36 vote in the House, a bitterly divided Congress approved the Act in a rare display of bipartisanship earlier that month.[4] Apart from triggering an awakening within Congress about the dire need to decarcerate, the Act unified an unusual coterie of proponents, including tycoons such as the Koch Brothers, and celebrities such as Kim Kardashian.[5]

Whilst hailed as historic and sweeping in some quarters,[6] the Act only affects the federal system, which houses a small fraction of the United States prison population. Out of approximately 2.1 million people imprisoned,[7] only 180,413 are federal inmates.[8]

Nonetheless, the Act aims to introduce several reforms. It mandates the Department of Justice to establish a 'risk and needs assessment system' to classify the recidivism risk of prisoners, and to incentivise participation in productive activities.

---

[1] Raghav Kohli is reading for a B.A., LL.B. (Hons.) at Gujarat National Law University, India, in his fourth year.

[2] The Formerly Incarcerated Reenter Society Transformed Safely Transitioning Every Person Act 2018 (US.)

[3] 'President Donald J. Trump Secures Landmark Legislation to Make Our Federal Justice System Fairer and Our Communities Safer' (*White House*, 21 December 2018) <https://www.whitehouse.gov/briefings-statements/president-donald-j-trump-secures-landmark-legislation-to-make-our-federal-justice-system-fairer-and-our-communities-safer/> accessed 6 February 2019.

[4] Van Jones and Jessica Jackson, '10 reasons to celebrate the First Step Act' *CNN* (21 December 2018) <https://edition.cnn.com/2018/12/21/opinions/ten-reasons-to-celebrate-first-step-act-jones-and-jackson/index.html> accessed 6 February 2019.

[5] Ryan Bort, 'Kim Kardashian, Alyssa Milano, Van Jones Among 50+ Celebrities Lobbying for Prison Reform Legislation' *Rolling Stone* (New York, 14 November 2018) <https://www.rollingstone.com/politics/politics-news/kardashian-prison-reform-755934/> accessed 6 February 2019.

[6] Nikki Schwab, 'House passes criminal justice reform bill' *New York Post* (New York, 20 December 2018) <https://nypost.com/2018/12/20/house-passes-criminal-justice-reform-bill/> accessed 6 February 2019.

[7] 'USA Data' (*World Prison Brief*) <http://www.prisonstudies.org/country/united-states-america> accessed 6 February 2019.

[8] 'Inmate Statistics' (*Federal Bureau of Prisons*, 7 February 2019) <https://www.bop.gov/about/statistics/population_statistics.jsp#pop_report_cont> accessed 6 February 2019.

For instance, it allows prisoners to earn 'time credits' through their participation and apply them towards early release to pre-release custody. Other proposed changes include retrospective modification of 'good time credit' computation, reduced sentences for drug-related offences, and a ban on shackling of pregnant women.

However, inmates do not benefit equally from these reforms. The risk and needs assessment system employs algorithms to classify each prisoner as having a minimum, low, medium, or high risk for recidivism. The Act only permits prisoners falling within the minimum and low risk brackets to apply for time credits towards pre-release custody.

This article seeks to critically examine the impact of such algorithmic decision-making in the criminal justice system. Analysing different instances of algorithmic bias and the recent Wisconsin Supreme Court decision of *State v. Loomis*,[9] it argues that opaque algorithmic decisions violate due process safeguards. In conclusion, the increasing use of such algorithms in the criminal justice system, including the FIRST STEP Act, is found to be undesirable, unless tempered with solutions which meaningfully improve their accuracy and transparency.

II. Algorithmic Bias in Decision-Making

Algorithmic decision-making has been demonstrated to perpetuate, at times even accentuate, gender and racial stereotypes within the criminal justice system.[10] In a study conducted in 2016, ProPublica examined the COMPAS algorithm, one of the most popular scores used in pre-trial and sentencing to assess a criminal defendant's likelihood of recidivism.[11] The analysis found that black defendants were twice as likely to be misclassified as higher risk compared to their white counterparts.[12] Furthermore, white defendants were mistakenly labelled low risk almost twice as often as black re-offenders.[13] Similarly, a recent report in 2019 by Liberty, a civil rights NGO, discovered that several UK police forces have used discriminatory algorithms to predict where crime will be committed, and by whom, based on factors such as racial profiling.[14]

---

[9] *State v Loomis* 881 NW 2d 749 (Wis 2016).

[10] Danielle Kehl, Priscilla Guo, and Samuel Kessler, 'Algorithms in the Criminal Justice System' (*Berkman Klein Center for Internet & Society, HLS*, 2017) <https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf> accessed 6 February 2019.

[11] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin, 'How We Analyzed the COMPAS Recidivism Algorithm' (*ProPublica*, 23 May 2016) <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> accessed 6 February 2019.

[12] *Ibid*.

[13] *Ibid*.

[14] 'Liberty Report Exposes Police Forces' Use Of Discriminatory Data To Predict Crime' (*Liberty*, 4 February 2019) <https://www.libertyhumanrights.org.uk/news/press-releases-and-statements/liberty-report-exposes-police-forces%E2%80%99-use-discriminatory-data-0> accessed 6 February 2019.

However, this cannot be conceived of as an entirely unanticipated phenomenon. Since machine learning depends on the data it processes, biases or inaccuracies in the sample size can easily be amplified.[15] Even where the data used to train machine learning algorithms is neutral, prejudices have been shown to creep in at multiple other stages of the deep-learning process.[16] The problem of bias is further compounded as algorithms are often protected intellectual property or are kept secret due to their proprietary nature. This prevents affected parties from challenging its decisions by permitting opaqueness in its decision-making process.[17]

The evidence strongly suggests that algorithmic bias may further entrench the already glaring structural inequalities in the US criminal justice system.[18] Given this risk, the US justice system seems to be at a constitutional, ethical, and technological crossroads.

III. The Threat to Due Process Rights and Accountability in Criminal Justice: Analysing the Pitfalls of *State v. Loomis*

Specifically, the use of such algorithms in risk assessments and sentencing raises grave concerns about due process safeguards and accountability in criminal justice. Recently, in *Loomis*, an individual facing a six-year imprisonment term challenged the sentencing judge's use of the COMPAS algorithm before the Supreme Court of Wisconsin, arguing that it violated his due process rights on three grounds. First, it violated the defendant's right to be sentenced based upon accurate information, as the proprietary nature of COMPAS prevented him from assessing its accuracy. Secondly, it violated a defendant's right to an individualised sentence. Finally, it improperly used gendered assessments in sentencing.

The Court rejected these arguments and held that the use of the COMPAS algorithm at sentencing, within narrow constraints, did not violate the defendant's due process rights.[19] It observed that while it could not be a determinative factor, a COMPAS risk assessment helps in deciding an individualised sentence alongside other supporting factors.[20] The Court also mandated that any Presentence Investigation Reports ('**PSI**'s) containing a COMPAS assessment must give an advisement cautioning the sentencing court about its limitations.[21] In spite of explicitly recognising

[15] David Danks and Alex John London, 'Algorithmic Bias in Autonomous Systems' (2017) Proceedings of the 26th IJC on AI, 2 <https://www.cmu.edu/dietrich/philosophy/docs/london/IJCAI17-AlgorithmicBias-Distrib.pdf> accessed 6 February 2019.
[16] Karen Hao, 'This is how AI bias really happens—and why it's so hard to fix' (*MIT Technology Review*, 4 February 2019) <https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/>accessed 15 March 2019.
[17] Anupam Chander, 'The Racist Algorithm?' (2017) 11 Michigan Law Review 1023.
[18] 'Report to the UN on Racial Disparities in the US Criminal Justice System' (*The Sentencing Project*, 19 April 2018) <https://www.sentencingproject.org/publications/un-report-on-racial-disparities/> accessed 6 February 2019.
[19] *Loomis* (n 7).
[20] *Ibid*.
[21] *Ibid*.

the problems that plague algorithmic decision-making, the Court's reasoning leaves much to be desired.

A. The Court misunderstood how algorithmic decision-making works

While addressing the defendant's inability to assess the accuracy of the algorithm due to its proprietary nature, the Court made fatally erroneous observations about the nature of algorithms. It held that even though the COMPAS report did not disclose how risk scores are determined or how the factors are weighted, the defendant could challenge the resulting risk scores set forth in the report attached to the PSI. This could be done based on the 2015 Practitioner's Guide to COMPAS prepared by Northpointe, the developer of the algorithm, which explains some of the variables considered in calculating the risk scores.

However, this approach misunderstands the different ways in which bias creeps into the algorithmic process. As discussed earlier,[22] even if the algorithm is facially neutral in the factors it considers, it may amplify biases and inaccuracies in the input data. Thus, in limiting a defendant's right to verify the results only to the extent of his answers to questions and publicly available data about his criminal history, the Court failed to account for the different stages of algorithmic decision-making that impact its final assessment. Further, by referring to the 2015 Guide prepared by the developer of the algorithm itself as an instrument for the defendant's benefit, the Court ignored the manifest conflict of interest involving Northpointe, a for-profit company with a $1,765,334 contract at stake in Wisconsin.[23]

B. The Court's proposed advisement is ineffective and inadequate

In an attempt to mitigate the defendant's concerns regarding racial bias and erosion of individualised sentencing, the Court mandated that a PSI using a COMPAS risk assessment must contain a four-part advisement that cautions sentencing courts about its limitations. This would include, *inter alia*, a warning stating that the proprietary nature of COMPAS has been invoked to prevent the disclosure of information relating to its functioning, and how some studies have raised questions about whether it discriminates against minority offenders.[24]

However, the efficacy of such a warning mechanism remains doubtful for two reasons. First, while the advisements recognise some limitations of such risk assessments, they do not provide any guidance to judges on how much they should discount such assessments.[25] This is significant as judges still have no means of verifying the accuracy of such tools, but are expected to factor them into their decisions. Secondly, the advisement is likely to merely pay lip service due to

[22] *Supra* (n14).
[23] Katherine Freeman, 'Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis' (2016) 18 *NCJOLT* 75, 92.
[24] *Loomis* (n 7).
[25] 'Wisconsin Supreme Court Requires Warning Before Use Of Algorithmic Risk Assessments In Sentencing' (2017) 130 *HLR* 1530, 1534.

widespread recognition of what is today known as 'automation bias',[26] or the 'technology effect',[27] or the 'anchoring effect'[28] – the phenomenon whereby judges tend to be submissive to scientifically generated results due to a cognitive bias supporting data reliance.

C. The Court's decision upholding the use of COMPAS as a non-determinative factor to arrive at an individualised sentence is flawed

Rejecting the defendant's argument that a COMPAS risk assessment amounts to sentencing based on group data, the Court held that considering such a risk assessment as non-determinative along with other supporting factors is helpful in arriving at an individualised sentence.[29] However, this observation does not consider the fact that in the absence of any means to ascertain the accuracy of the risk assessment, the Court cannot determine the appropriate degree of reliance to be placed on this factor. Given the grave concerns regarding the high possibility of bias, the Court failed to offer a satisfactory justification for its continued use even as a mere 'non-determinative factor' in risk assessment and sentencing.

IV. The Way Forward

The decision in *Loomis* aptly demonstrates the challenges of algorithmic decision-making in the criminal justice system. While the FIRST STEP Act does not concern itself with the use of risk assessment tools in sentencing, reservations about due process continue to plague other determinations for which the Act employs algorithms, such as deciding which prisoners are entitled to pre-release custody. Since these determinations have a significant impact on the lives of prisoners, the continuing use of algorithms, however minor, is both disconcerting and arguably illegitimate.

It is therefore imperative to prevent the perpetuation of discrimination as a consequence of criminal justice reform. Even if the use of algorithms is considered indispensable, the Attorney General, along with the Bureau of Prisons, must actively take steps to avoid algorithmic bias under the Act.

For instance, promoting greater transparency regarding, *inter alia*, how an algorithm was developed, what assumptions were considered in its design, how its factors are weighted, what data was used to train it, and how frequently it is updated, will improve its credibility in two significant ways.[30] First, such information will enable affected individuals to meaningfully challenge algorithmic decisions. For instance, Europe's new General Data Protection Regulation provides rights to 'meaningful

[26] Liberty (n 12).
[27] Freeman (n 21) at 97.
[28] HLR (n 23) at 1536; Han-Wei Liu, Ching-Fu Lin, and Yu-Jie Chen, 'Beyond State v. Loomis: Artificial Intelligence, Government Algorithmization, and Accountability' (2019) 27 *International Journal of Law and Information Technology* 122, 133.
[29] *Loomis* (n 7).
[30] Kehl *et al* (n 8) at 32.

information about the logic involved' in automated decisions, commonly known as the 'right to explanation'.[31] While it is a welcome development, the debate on the scope and applicability of this right remains contentious.[32] Second, it will facilitate audits by external reviewers. Regular audits by independent researchers to check the design and real-world impact of algorithms are expected to go a long way in improving accountability.[33]

Further, algorithmic bias may be actively countered by a combination of manual and automated steps. For instance, human oversight of algorithmic decisions, and manual reviews of algorithmic correlations for identification of stereotypes can temper algorithmic decisions with optimal human intervention.[34] This may be coupled with 'algorithmic affirmative action', which seeks to train algorithms against biases and incorporate safeguards in its design to reflect equal opportunity.[35] Such an approach would require designers to foresee how algorithms function in the real-world and rectify problematic results accordingly.[36] In fact, such algorithms are increasingly being used by giants such as Facebook and Google.[37]

As we continue to redefine our relationship with technology, we must ensure that new pervasive technologies in our lives are informed by ethical considerations. Perhaps, adopting the aforementioned measures could prevent the FIRST STEP Act from becoming a misstep in criminal justice reform.

[31] Selbst and Powles, 'Meaningful information and the right to explanation' (2017) 7(4) International Data Protection Law 233-242.
[32] *Ibid*.
[33] 'Algorithmic Impact Assessments' (*AI Now Institute*, April 2018) <https://ainowinstitute.org/aiareport2018.pdf> accessed 10 June 2019.
[34] Gideon Mann and Cathy O'Neil, 'Hiring Algorithms Are Not Neutral' (*Harvard Business Review*, December 2016) <https://hbr.org/2016/12/hiring-algorithms-are-not-neutral> accessed 6 February 2019.
[35] 'Big Data' (*The White House*, May 2016) <https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf> accessed 10 June 2019.
[36] Chander (n 15) at 1043-1044.
[37] *Ibid.*