3.1. Inaccurate or Discriminatory Rulings Due to Algorithmic Bias

Artificial Intelligence and the Law in Canada, 1st Ed.

Florian Martin-Bariteau, Teresa Scassa

Chapter 13 AI and Judicial Decision-Making

3. Risks and Opportunities

3.1. Inaccurate or Discriminatory Rulings Due to Algorithmic Bias

All human beings — even the most well-meaning — have implicit biases that can surface when making decisions. Algorithms are instead sometimes presented as fair and unbiased decision-making agents. But it is well documented that algorithmic impartiality is a myth: there exists a plethora of examples and documented cases in which decision-making algorithms have produced biased outcomes.[1] As AI increasingly assists judges, understanding algorithmic bias is important because it can lead to inaccurate and discriminatory results.

The promise of fair and unbiased algorithmic decisions is a myth not because algorithms are not useful to help humans make decisions (they are). It is a myth because algorithmic bias is as real as it is inevitable — and it must be regulated differently than human bias.[2] There are three types of bias in AI that can lead to inaccurate and discriminatory results: bias in the process of building the algorithmic model, bias in the sample that is used to train the algorithm, and societal biases captured and amplified by the algorithm.

The first type of bias is a biased process. This is a bias in how an algorithm processes information.[3] Biases in an algorithmic process often exist because human biases are translated into the system.[4] Even if no human chooses the outcome directly, there is always human involvement in how that outcome is arrived at: humans frame the problem and make a choice about what the algorithm should predict before any data are processed. Once that is decided, there is human involvement through gathering data to train the algorithm and selecting the variables that the algorithm should consider (the *features*). There are always, at some level, human decision-makers that influence the process.

The second type is a biased data sample. An algorithm's predictive power is only as good as the data that it is fed. If an algorithm mines in a section of a dataset that, for any reason, is unrepresentative of the population, it will produce non-representative outputs (*i.e.*, inaccurate and potentially discriminatory individual decisions).[5] Individual records, for example, may suffer from quality problems due to partial or incorrect data. The entire dataset might also have quality problems at higher rates for an entire protected group compared to others or might be unrepresentative of the general population.[6]

The third type of algorithmic bias is data that reflects societal biases. A machine-learning algorithm's training data may reflect prior systemic discrimination.[7] An AI can thus produce a disparate impact or indirect discrimination even when correctly trained with representative data.[8] The difference between biased sample data and this type of bias is that, here, the data is representative of the population, but this representative data still produces a disparate impact outcome because of embedded social inequalities.[9]

Examples of these biases exist in practically every area of decision-making where AI is used, but perhaps the most widely known among them is the use of the COMPAS software for risk assessment in criminal procedure. COMPAS

aims to predict the likelihood that an accused will recidivate if granted parole.[10] A few years ago, a ProPublica investigation accused COMPAS of producing racially biased results for both high-risk and low-risk classifications.[11] The investigation found that almost twice as many Black defendants than white defendants were incorrectly classified as high-risk by COMPAS, and white defendants were also more likely to be incorrectly classified as low-risk than were Black defendants.[12] COMPAS is still widely used by judges in the U.S. for parole hearings, bail hearings and sometimes sentencing.

Other risk assessment instruments employed in criminal justice exhibit similar patterns of bias in their outcomes. For example, a tool used at the federal level in Canada to make probation decisions (Post Conviction Risk Assessment) was found to give Black offenders higher average post-conviction risk assessment scores than white offenders.[13] Using criminal history as a predictor captures societal biases (the third type of algorithmic bias mentioned above) because criminal history captures the relationship between race and arrest, where Black individuals are more likely to be arrested than white individuals for the same level of criminal activity.[14]

If left unchecked, these biases can lead judges to inadvertently endorse inaccurate or discriminatory outcomes in their rulings.

Footnote(s)

1 See, *e.g.*, Julia Angwin & Jeff Larson, “Bias in Criminal Risk Scores is Mathematically Inevitable, Researchers Say” (December 30, 2016), online: *ProPublica* www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say; Jeffrey Dastin, “Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women” (October 9, 2018), online: *Reuters* https://reut.rs/2Od9fPr.

2 Kate Crawford & Jason Schultz, “Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms” (2014) 55:1 Boston College L. Rev. 93 at 124-28.

3 Ignacio N. Cofone, “Algorithmic Discrimination is an Information Problem” (2019) 70:6 Hastings L.J. 1389 at 1399-1402.

4 Batya Friedman & Helen Nissenbaum, “Bias in Computer Systems” (1996) 14:3 A.C.M. Trans. Inf. Syst. 330 at 333-36 (explaining the difference between preexisting bias and technical bias).

5 Ignacio N. Cofone, “Algorithmic Discrimination is an Information Problem” (2019) 70:6 Hastings L.J. 1389 at 1402-1404.

6 Solon Barocas & Andrew D. Selbst, “Big Data’s Disparate Impact” (2016) 104:3 Cal. L. Rev. 671 at 680-81, 684-87.

7 Ignacio N. Cofone, “Algorithmic Discrimination is an Information Problem” (2019) 70:6 Hastings L.J. 1389 at 1404-1406.

8 Solon Barocas & Andrew D. Selbst, “Big Data’s Disparate Impact” (2016) 104:3 Cal. L. Rev. 671 at 673-74, 691.

9 Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, “Semantics Derived Automatically from Language Corpora Contain Human-Like Biases” (2017) 356:6334 Science 183; Daniel Rosenberg, “Data Before Fact” in Lisa Gitelman, ed., *“Raw Data” Is an Oxymoron* (Cambridge, MA: M.I.T. Press, 2013), 15.

10 Timm Brennan, William Dieterich & Beate Ehret, “Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System” (2009) 36:1 Crim. Justice Behav. 21 at 22-24.

11 Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks” (May 23, 2016), online: *ProPublica* www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Jeff Larsen, Surya Mattu, Lauren Kirchner & Julia Angwin, “How We Analyzed the COMPAS Recidivism Algorithm” (May 23, 2016), online: *ProPublica* www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

12 See Alexandra Chouldechova, “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments” (2016) 5:2 Big Data 153 at 153, 156; Anupam Chander, “The Racist Algorithm?” (2017) 115:6 Mich. L. Rev. 1023 at 1033; Fiona Doherty, “Obey All Laws and Be Good: Probation and the Meaning of Recidivism” (2016) 104:2 Geo. L.J. 291 at 352-53; Jessica M. Eaglin, “Constructing Recidivism Risk” (2017) 67 Emory L.J. 59 at 96; Melissa Hamilton, “Risk-Needs Assessment: Constitutional and Ethical Challenges” (2015) 52 Am. Crim. L. Rev. 231 at

239-41; Kelly Hannah-Moffat, “Algorithmic Risk Governance: Big Data Analytics, Race and Information Activism in Criminal Justice Debates” (2019) 23:4 Theor. Criminol. 453 at 461.

13 Jennifer L. Skeem & Christopher T. Lowenkamp, “Risk, Race, and Recidivism: Predictive Bias and Disparate Impact” (2016) 54:4 Criminology 680 at 685-700.

14 Jennifer L. Skeem & Christopher T. Lowenkamp, “Risk, Race, and Recidivism: Predictive Bias and Disparate Impact” (2016) 54:4 Criminology 680 at 700.

End of Document