Marion Oswald and Angela Paul

11 Bias

Abstract: This chapter presents examples of biases that can occur in relation to the use of data-driven and algorithmic tools within criminal justice, and explores the different definitions of ‘bias’ in criminology, computer science, and law. We highlight scholarly discourse on analysis, as well as risk and mitigation from the perspectives of the above disciplines, and offer a new taxonomy to aid researchers. The chapter advocates for an interdisciplinary strategy for understanding and regulating data-driven approaches within criminal justice.

Keywords: bias, policing, algorithms, law, discrimination

The origin of the word ‘bias,’ although contested, is thought to have derived from the Old French word ‘biais,’ which means slant, slope, or against the grain (CEBM, 2018). The word may have also entered the English language through the game of bowls, which is played with ‘biased’ balls that travel obliquely because of the difference in weight on one side of the balls. The term ‘bias’ is still used in everyday language to refer typically to a psychological inclination—sometimes a reprehensible one—towards or against a certain opinion. However, the term takes on alternative or additional nuances when employed by law, computer science, and criminology/surveillance studies in the context of the use of algorithmic tools in criminal justice (see Algorithm by Leese). The daily operations of law enforcement and the legal system have been revolutionized by the incorporation of machine learning techniques, sometimes replacing human expertise with algorithmic evaluations (see Artifical Intelligence by Van Brakel).

This chapter explores definitions and understandings of the term ‘bias’ common to the disciplines of law, computer science, and criminology, which are crucial to understanding the significance of bias in digital criminology. By doing so, we highlight similarities and differences in such definitions and understandings. The chapter presents selected examples of biases that can occur in relation to the use of data-driven and algorithmic tools within criminal justice, highlighting scholarly discourse on analysis, as well as risk and mitigation from the perspectives of the above disciplines. The term bias is frequently used to refer to unlawful discrimination based on protected characteristics, such as race, which can commonly occur in the criminal justice system (Richardson et al., 2019). We offer a taxonomy of bias to help researchers looking at bias in these contexts. Although out of scope for this chapter, we note the overlap between issues of bias and other important rights, including those relating to privacy and data

Disclaimer: Angela Paul’s PhD studies were supported by Northumbria University.

protection, and the issue of collection and use of sensitive demographic data for the purposes of mitigating bias (CDEI, 2023).

Understandings of bias in the context of algorithmic criminal justice

It is perhaps in the legal context that we come closest to the term 'bias' reflecting common usage. This is because bias is a concept that is relevant to the governing of state power, and to the issue of unlawful discrimination. For example, in English common law, judicial review is the procedure whereby the courts supervise the lawfulness of the exercise of power by public bodies. A key ground of review (and element of natural justice) is the 'rule against bias.' That is, to ensure a decision-making process is fair, no proven or real danger of bias should be present, assessed from the point of view of a fair-minded and informed observer. What does bias mean in this context, however? Law recognizes that bias in human decision-making represents a mental inclination, even one unknown to the decision-maker, as per Lord Goff:

> bias is an insidious thing that, even though a person may in good faith believe he was acting impartially, his mind may unconsciously be affected by bias (*R v Gough*, 1993).

The issue of bias is linked closely to the need to ensure the fairness of any tribunal or decision-making process (Article 6, ECHR, emphasizes the right to a fair trial by an independent and impartial tribunal), as stated by Lord Hope:

> The word 'bias' is used as a convenient shorthand. But it would be a mistake to approach it in this context as if its only meaning were pejorative. The essence of it is captured in the Convention concept of impartiality. An interest in the outcome of the case or an indication of prejudice against a party to the case or his associates will, of course, be a ground for concluding that there was a real possibility that the tribunal or one of its members was biased … but the concept is wider than that. It includes an inclination or pre-disposition to decide the issue only one way, whatever the strength of the contrary argument. (*Davidson v Scottish Ministers*, 2004)

Although bias in this context is concerned with bias (or reasonable suspicion of bias) of the individual decision-maker or tribunal, the insertion of an algorithm into the process may also create or exacerbate that bias. As noted by Oswald et al., "it could be said that a risk-averse algorithm, which we know over-estimates risk in order to maximise public protection but which generates a degree of 'false positives' of high-risk results to do this, might actually be creating a biased process (or tribunal of sorts)" (Oswald et al., 2018: 241.). For example, a facial recognition tool might have a low face-match threshold in order to minimize the risk of missing a wanted individual, but this then results in a higher risk of false positives occurring (see Facial Recognition by Fussey). Similarly, Cobbe points out that due to less favorable treatment or disadvantage reflected in training data, "ADM [automated decision-making] systems may be prone to making decisions which are systematically skewed in some way, rather than acting impartially." Objective judgment could be prevented by "the presence of an internal model which does not produce fair and consistent outputs (for example, a system could, without any intention to do so on the part of the public body, treat those from certain socioeconomic backgrounds less favourably than others)" (Cobbe, 2019).

The biases which are present in automated systems are referred to as 'statistical biases,' which are caused by using data that includes systematic errors which, in turn, skew results. These errors can arise from the algorithm itself and/or from faulty, incomplete, irrelevant, excluded, or biased data sets, thus creating consistent errors in a model, the error being the difference between the ground truth and the average model prediction (i. e., between the reality on the ground and what the model is telling you). Bias is also often used as a shorthand, for unlawful discrimination on the grounds of certain protected characteristics such as sex, gender, age, sexual orientation, disability, and race. These two understandings of bias are closely connected as we explore below.

Machine learning tools have been incorporated into the daily work and processes of police and the judicial system, for instance, to replace the expert opinions of psychiatrists or probation officers, which were used in the past to judge the risk of criminality (see Policing by Wilson). The main justification given for using risk assessment instruments can be to "reduce the noise inherent to human decision making … for example, some judges predict recidivism better than others" (Goel et al., 2018). However, the use of data analytics to make law enforcement decisions or predictions may mean that disparities become encoded into the datasets, thus feeding back into the system. In the context of the enforcement of marijuana violations in the US, Butcher et al. argue that "predictions made by these tools may reflect or even exacerbate past racially disparate enforcement" (Butcher et al., 2022: 137). Recidivism is sometimes predicted using data inputs that include certain personal characteristics, for instance, an individual's neighborhood or socioeconomic status. Although race may not be an explicit predictor, algorithms can in fact "contain an imperfect proxy for race" or other protected characteristics (Davies and Douglas, 2020), for instance the risk of using arrest in tools that aim to predict risk of re-offending (Fogliato et al., 2021).

Discussions of bias, in terms of unlawful discrimination that can result from it, are extremely relevant in the context of automated criminal justice processes. A type of statistical bias is 'sampling bias,' which can occur when systemic discrimination is reflected in training data (Borgesius, 2018). Systems built on data from periods of flawed, racially biased, and potentially unlawful practices ('dirty data') could result in flawed predictions and harmful feedback loops (Richardson et al., 2019). Birhane too calls for understanding of historical injustices and power asymmetries embedded within algorithmic systems (Birhane, 2022). In addition, a study which surveyed legal professionals in the UK revealed that racial bias plays a significant role in the judicial system, and 55.6 % of the legal professionals said that they have "witnessed one or more judges act with racial bias in their treatment of defendants" (Monteith et al., 2022). Furthermore, the Casey Review into the UK's Metropolitan Police found that:

> Black Londoners in particular remain over-policed. They are more likely to be stopped and searched, handcuffed, batoned and Tasered, are over-represented in many serious crimes, and when they are victims of crime, they are less satisfied with the service they receive than other Londoners. (Baroness Casey, 2023: 17)

These revelations around 'dirty data,' and the witting or unwitting institutional biases that are already embedded in the criminal justice systems, may correlate to the biases that can be present in the outputs of digital analytical processes.

There is of course a high risk that the use of a biased algorithm could result in unlawful direct or indirect discrimination on the grounds of protected characteristics, or breach of other equality duties, contrary to legislation such as the UK's Equality Act 2010. We refer to this concept as 'statistics-based discrimination,' where actions or decisions taken based on or guided by the biased output contribute to prohibited conduct. We would need however to consider the link between the biased algorithm and the unlawful discrimination (or positive duties to have due regard to eliminating such discrimination: s149(1) Equality Act 2010). For example, Allen and Masters give an example of AI-driven online advertising resulting in women being shown a job advert less frequently than men. This is likely to be unjustified indirect discrimination contrary to s19 of the Equality Act (Allen and Masters, 2021). Another example can be seen from the famous analysis by journalists at ProPublica who concluded that the risk scores generated by a system called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), used to inform sentencing decisions, were nearly twice as likely to predict falsely that black defendants would commit crime in the future when compared to white defendants (Angwin et al., 2016). Allen and Masters describe this as creating "a new stereotype" (Allen and Masters, 2021: 44).

Furthermore, an evaluation undertaken by the UK's National Physical Laboratory of the facial recognition system used by the Metropolitan Police concluded that at a face-match threshold of 0.6, the system operated equitably between gender and ethnicity. However, if a lower threshold was used, then ethnic disparities increased. At the 0.56 setting, of the 33 people falsely identified in the trial, 22 were black, 8 were Asian, and 3 were white, 18 being in the 21–30 age group (Mansfield, 2023: 7). The bias in the system revealed at this setting would have direct implications for the justifiability, reliability, and lawfulness of decisions informed by the probabilistic output, including stop and search, and even arrest. Cases from the US suggest that despite guidance being issued that a 'match' can never be a sole ground for arrest, officers tend to rely on the outputs of facial recognition and wrongful, biased arrests have occurred (Magnet, 2011: 150; Johnson, 2022; Bhuiyan, 2023).

We have focused so far on data bias connected to discrimination on the grounds of protected characteristics. However, bias as it relates to the deployment of algorithms within criminal justice may impact wider decision-making. Algorithmic decision-making and data analytics may create new categories or groups on the basis of criteria that are not necessarily linked to protected characteristics, but may still cause unfairness, for example due to social inequalities (Gerards and Borgesius, 2022). Wachter, Mittelstadt, and Russell point to certain fairness metrics, such as equalized odds, classification parity, or false positive error rate equality, that may in fact preserve bias rather than reducing it, because these methods do nothing to address underlying causes of inequality (Wachter et al., 2021).

This concern could arguably be illustrated by the assessment of equitability carried out on the facial recognition system used by the Metropolitan Police as mentioned above (Mansfield, 2023: 7) which defines equitability between demographics as requiring that, “in the operational setting, the outcomes for the subjects (i. e., recognition rates and false alert rates) should be broadly equivalent for demographics considered.” False alerts may still occur however which may raise broader fairness issues. Computational methods of determining fairness or bias could be insufficient to incorporate the interpretative and contextual legal tests of proportionality and legitimacy (Sanchez-Monedero et al., 2020). In respect of probabilistic classifications (i. e., a statistical determination that someone/thing might meet a group’s characteristics), it may be impossible to satisfy the conditions of competing notions of fairness simultaneously (Kleinberg et al., 2017).

Babuta and Oswald argue that focusing on ‘data’ bias may distract attention away from the issue of whether algorithmic techniques are appropriate at all for particular criminal justice contexts (Babuta and Oswald, 2019). Bias may occur at all stages of the project lifecycle, from bias in favor of data-driven solutions, moving to bias in training data and misleading accuracy rates. Issues of ‘model fit’ may occur if an algorithm is trained and validated only on a limited sample, and then applied to a more diverse dataset, for example if an algorithm is trained only on older males, and then applied to a dataset that includes females and younger adults. Automation bias may be an issue, a form of cognitive bias in which individuals are partial towards the decisions of automated systems, leading to the individual deferring to automated decisions over human judgment (Citron, 2008; Rieke and Bogen, 2018: 9). Oversight itself could be biased if limited to data science expertise or where there is a lack of independence (see figure 1, Risk of bias in implementation and oversight of police algorithms in Babuta and Oswald, 2019).

Furthermore, Kaufmann argues that the concept of bias itself could be misleading or superfluous: “In understanding information as relational, context-specific and lively, bias becomes a superfluous concept, because it is everywhere, in every dataset. The only way to engage with ‘bias’ is then to identify and reflect about the specificities of information and how to engage with them” (Kaufmann, 2023: 16).

Conclusions and lessons for research

The deployment of digital, algorithmic, and AI technologies within criminal justice brings with it a risk of bias. Yet researching such ‘bias’ also brings the risk of misunderstandings between disciplines if the term and the type(s) of bias in play or to be investigated are not clearly defined from the outset. We set out below a classification of the types of bias explored in this chapter which can be deployed by researchers in this field at the outset of their projects in order to minimize the likelihood of working at cross-purposes:

Table 1: Bias classification

***Type* of bias**	***Elements* and *location* of this bias**
Cognitive or decision-making	Inclination to decide one way due to partiality—human decision-maker or systematically skewed algorithm. Legal bias—The rule against bias in English common law through judicial review i. e. to ensure a tribunal or decision-making process is impartial and fair. Automation bias—human decision-maker deferring to the algorithmic output.
Statistical	The overarching term for statistically driven decision-making that can be based on faulty, incomplete, irrelevant, excluded. or biased data sets. Sampling bias—systemic discrimination is reflected in automated decision-making due to biased training data, incomplete data or proxy variables. Model fit—where an algorithm is trained and validated only on a limited sample, and then applied to a more diverse or different dataset. Statistics-based discrimination—where actions or decisions based on, guided by, or linked to the biased output contribute to prohibited/unlawful conduct.
Institutional, historical or systemic	These are witting or unwitting patterns of bias that are deeply embedded in institutions and systems of society, including the criminal justice system. Examples include racial and socioeconomic biases or disparity, which can lead to discrimination. It could also include datasets that reflect imperfect or skewed enforcement or investigation practices.

In digital criminology, when biased automated decision-making systems lead to discriminatory outcomes, the types of biases mentioned above become interlinked. However, it is also important to dissect the different biases separately to understand the wider implications, and this can be done through an interdisciplinary approach to bias. It is only by taking an interdisciplinary approach that ongoing attempts to understand, oversee, and regulate data-driven approaches within criminal justice can hope to achieve success.

Suggested reading

Babuta, A., & Oswald, M. (2019). *Data Analytics and Algorithmic Bias in Policing.* RUSI Briefing Paper. Available at https://rusi.org/explore-our-research/publications/briefing-papers/data-analytics-and-algorithmic-bias-policing.

Borgesius, F. (2018). *Discrimination, Artificial Intelligence and Algorithmic Decision-Making.* The Council of Europe. Available at https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73.

Butcher, B., Robinson, C., Zilka, M., Riccardo, F., Ashurst, C., & Weller, A. (2022). Racial disparities in the enforcement of marijuana violations in the US. *AI Ethics and Society (AIES) '22 Conference, Oxford, United Kingdom, 1–3 August.* Available at https://doi.org/10.1145/3514094.3534184.

Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *N.Y.U. L. REV. ONLINE*, 94, 192–233.

References

Allen, R., & Masters, D. (2021). *Technology Managing People – The Legal Implications.* Trade Union Congress. Available at https://www.tuc.org.uk/research-analysis/reports/technology-managing-people-legal-implications (Accessed 3 July 2023).

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias, *ProPublica*, 23 May. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed 3 July 2023).

Babuta, A., & Oswald, M. (2019). *Data Analytics and Algorithmic Bias in Policing.* RUSI Briefing Paper. Available at https://rusi.org/explore-our-research/publications/briefing-papers/data-analytics-and-algorithmic-bias-policing (Accessed 5 July 2023).

Baroness Casey of Blackstock. (2023) *Baroness Casey Review: An independent review into the standards of behaviour and internal culture of the Metropolitan Police Service.* Available at: https://www.met.police.uk/police-forces/metropolitan-police/areas/about-us/about-the-met/bcr/baroness-casey-review/ (Accessed 3 July 2023), p.17.

Birhane, A. (2022). The limits of fairness. *AI Ethics and Society (AIES) '22 Conference*, Oxford United Kingdom, 1–3 August. Available at: https://doi.org/10.1145/3514094.3539568 (Accessed 5 July 2023).

Bhuiyan, J. (2023). First man wrongfully arrested because of facial recognition testifies as California weighs new bills. *The Guardian*, 27 April. Available at: https://www.theguardian.com/us-news/2023/apr/27/california-police-facial-recognition-software (Accessed 3 July 2023).

Borgesius, F. (2018) *Discrimination, Artificial Intelligence and Algorithmic Decision-Making.* The Council of Europe. Available at: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 (Accessed 5 July 2023).

Butcher, B., Robinson, C., Zilka, M., Riccardo, F., Ashurst, C., & Weller, A. (2022). Racial disparities in the enforcement of marijuana violations in the US. *AI Ethics and Society (AIES) '22 Conference*, Oxford, United Kingdom, 1–3 August (pp. 130–143). Available at: https://doi.org/10.1145/3514094.3534184 (Accessed 5 July 2023).

Centre for Data Ethics and Innovation (CEDI). (2023) *Enabling responsible access to demographic data to make AI systems fairer*, 14 June. Available at: https://www.gov.uk/government/publications/enabling-responsible-access-to-demographic-data-to-make-ai-systems-fairer (Accessed: 5 July 2023).

Centre for Evidence-Based Medicine (CEBM) (2018) *Bias – Etymology and Usage.* Available at: https://catalogofbias.org/2018/04/10/a-word-about-evidence-4-bias-etymology-and-usag/ (Accessed: 5 July 2023).

Citron, D. (2008). Technological due process. *Washington University Law Review*, 85(6), 1271–1272.

Cobbe, J. (2019) Administrative law and the machines of government: Judicial review of automated public-sector decision-making. *Legal Studies*, 39, 653–654.

Davidson v Scottish Ministers [2004] UKHL 34.

Davies, B., & Douglas, T. (2022). Learning to discriminate: The perfect proxy problem in artificially intelligent criminal sentencing. In J. Ryberg and J. V. Roberts (eds.), *Sentencing and Artificial Intelligence* (pp. 97–121). Oxford: Oxford University Press.

Fogliato, F., Xiang, A., Lipton, Z., Nagin, D., & Chouldechova, A. (2021). On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. *AI Ethics and Society (AIES) '21 Conference*, Virtual Event USA, 19–21 May.

Gerards, J., & Borgesius, F. (2022). Protected grounds and the system of non-discrimination law in the context of algorithmic decision-making and artificial intelligence. *Colorado Technology Law Journal, 20(1), 1–55.*

Goel, S., Shroff, R., Skeem, J., & Slobogin, C. (2019). The accuracy, equity, and jurisprudence of criminal risk assessment. In R. Vogl (ed.), Research Handbook on Big Data Law (pp. 9–28). Cheltenham: Edward Elgar Publishing.

Johnson, K. (2022). How wrongful arrests based on AI derailed 3 men's lives. *Wired*, 7 March. Available at: https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/ (Accessed 5 July 2023).

Kaufmann, M. (2023). *Making Information Matter.* Bristol: University of Bristol Press.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores *Innovation in Theoretical Computer Science (ITCS) 2017 Conference, Berkeley, California, USA, 9–11 January.* Available at: https://drops.dagstuhl.de/opus/volltexte/2017/8156/pdf/LIPIcs-ITCS-2017-43.pdf (accessed 5 July 2023).

Magnet, S. A. (2011). *When Biometrics Fail: Gender, Race, and the Technology of Identity.* Durham, NC: Duke University Press.

Mansfield, T. (2023). *Facial Recognition Technology in Law Enforcement Equitability Study: Final Report.* National Physical Laboratory. Available at: https://science.police.uk/site/assets/files/3396/frt-equitability-study_mar2023.pdf (Accessed 5 July 2023).

Monteith, K., Quinn, E., Joseph-Salisbury, R., Dennis, A., Kane, E., Addo, F., & Mcgourlay, C. (2022). *Racial Bias and the Bench: A Response to the Judicial Diversity and Inclusion Strategy.* University of Manchester. Available at: https://documents.manchester.ac.uk/display.aspx?DocID=64125 (accessed: 16 May 2023), pp. 8, 14, 17.

Oswald, M., Grace, J., Urwin, S., and Barnes, G. (2018) Algorithmic risk assessment policing models: Lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, 27(2), 223–250.

R v Gough [1993] 2 WLR 883.

Richardson, R., Schultz, J., and Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *N.Y.U. L. REV. ONLINE*, 94, 192–233.

Rieke, A., & Bogen, M. (2018). *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias.* Upturn. Available at: https://www.upturn.org/work/help-wanted/ (Accessed 3 July 2023), p. 9.

Sanchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives. *FAT '20 Conference, Barcelona, Spain, 27–30 January*, 458–468.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review*, 3(123), 735–790.