THE MODERN LAW REVIEW

Modern Law Review
DOI: 10.1111/1468-2230.12759

Directly Discriminatory Algorithms

Jeremias Adams-Prassl,[*] Reuben Binns[†]
and Aislinn Kelly-Lyth[‡]

Discriminatory bias in algorithmic systems is widely documented. How should the law respond? A broad consensus suggests approaching the issue principally through the lens of indirect discrimination, focusing on algorithmic systems' impact. In this article, we set out to challenge this analysis, arguing that while indirect discrimination law has an important role to play, a narrow focus on this regime in the context of machine learning algorithms is both normatively undesirable and legally flawed. We illustrate how certain forms of algorithmic bias in frequently deployed algorithms might constitute direct discrimination, and explore the ramifications—both in practical terms, and the broader challenges automated decision-making systems pose to the conceptual apparatus of anti-discrimination law.

INTRODUCTION

Algorithmic decision-making systems (ADMS) discriminate: no automated system is completely free of bias. The discriminatory impact of ADMS has been documented in areas ranging from grade allocation and benefits decisions to the criminal justice system.[1] Private operators have, if anything, been even more enthusiastic in their embrace of ADMS, with predictably dire consequences; from banks persistently and systematically rejecting credit applications made by customers from certain ethnic groups to hiring systems automatically rejecting female candidates for engineering positions.[2]

[*] Professor of Law, Magdalen College, University of Oxford.

[†] Associate Professor of Human Centred Computing, Department of Computer Science, University of Oxford.

[‡] Researcher, Bonavero Institute of Human Rights, University of Oxford. The authors acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 947806), and are grateful to Robin Allen QC, Shreya Atrey, Catherine Barnard, Mark Bell, Hugh Collins, Jinghe Fan, Sandy Fredman, Philipp Hacker, Deborah Hellman, Tarun Khaitan, Dee Masters, Dinah Rose QC, Sandra Wachter and Raphaële Xenidis, as well as participants at the Oxford Algorithms at Work Reading Group, the UT Austin iSchool Research Colloquium, the Oxford Ethics in AI Research Seminar, the Lorentz Centre Workshop on Fairness in Algorithmic Decision Making and the Oxford Business Law Workshop, as well as the anonymous reviewers, for feedback and discussion. The usual disclaimers apply.

[1] 'A-levels and GCSEs: How did the exam algorithm work?' *BBC News* 20 August 2020 at https://perma.cc/RCF7-4A9L; Amnesty International, *Xenophobic Machines* (2021); and J. Larson, S. Mattu, L. Kirchner and J. Angwin, 'How We Analyzed the COMPAS Recidivism Algorithm' *ProPublica* 23 May 2016 at https://perma.cc/4QR7-485S.

[2] See N. Campisi, 'From Inherent Racial Bias to Incorrect Data – The Problems with Current Credit Scoring Models' *Forbes Advisor* 26 February 2021 at https://perma.cc/U6K5-UGR7;

The technical causes of machine bias are varied and complex. A vast literature explores different facets of the problem, from proxy discrimination to tainted training data; potential technical solutions to de-bias ADMS; and the inherent trade-offs (if any) between accuracy and bias.[3]

In addition to this technical scrutiny, litigants are increasingly turning to the courts to challenge algorithmic discrimination across a wide range of regulatory domains, from judicial review of public-sector ADMS to employment law.[4] A growing academic literature suggests that most cases of algorithmic bias will best be addressed through the lens of indirect discrimination.[5] Even a facially neutral provision, criterion, or practice (PCP) will be *prima facie* unlawful where it puts people with a protected characteristic at a 'particular disadvantage'. By characterising algorithms as PCPs, the focus shifts from the operation of an ADMS to its impact: are there disparities in its effects on groups sharing a protected characteristic?

Under EU and UK anti-discrimination law, this neatly sidesteps difficult questions of causation and avoids the need for technical explanations of ADMS' underlying mechanisms – but at significant cost. Indirect discrimination will only be unlawful if use of the PCP is not a proportionate means of achieving a legitimate aim.[6] In other words, if the use of a biased ADMS can be justified, then in legal terms no indirect discrimination has occurred.

Financial Conduct Authority, 'Pricing practices in the retail general insurance sector: Household Insurance' Thematic Review TR18/4 (October 2018) para 4.21; T.B. Gillis and J.L. Spiess, 'Big Data and Discrimination' (2019) 86 *The University of Chicago Law Review* 459; and J. Dastin, 'Amazon scraps secret AI recruiting tool that showed bias against women' *Reuters* 11 October 2018 at https://perma.cc/328A-UJFM.

[3] See for example D. Pedreshi, S. Ruggieri and F. Turini, 'Discrimination-aware data mining' (2008) *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 560; C. Dwork and others, 'Fairness through Awareness' (2012) *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* 214; J. Buolamwini and T. Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) 81 *Proceedings of Machine Learning Research* 77; J. Kleinberg, S. Mullainathan and M. Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores' (2017) 67 *Innovations in Theoretical Computer Science* 1, 23. We define 'proxy' here to refer to a feature which is correlated with a protected characteristic.

[4] *R (on the application of Bridges)* v *Chief Constable of South Wales* [2020] EWCA Civ 1058, [2020] 1 WLR 5037; C. Vallance, 'Legal action over alleged Uber facial verification bias' *BBC News* 8 October 2021 at https://perma.cc/TE4M-AMRH; 'Home Office drops "racist" algorithm from visa decisions' *BBC News* 4 August 2020 at https://perma.cc/EG22-4SAT. On the potential for system-level challenges, see A. Adams-Prassl and J. Adams-Prassl, 'Systemic Unfairness, Access to Justice and Futility: A Framework' (2020) 40 OJLS 561.

[5] While our discussion in this paper responds to the legal literature on algorithmic bias as a general phenomenon, we recognise that practitioners' approaches to specific cases will be fact-sensitive. See, for example, Joint Opinion of R. Allen QC and D. Masters in the Matter of Automated Data Processing in Government Decision Making 7 September 2019 at https://perma.cc/M2GU-D8HS, considering a number of case studies.

[6] See Council Directive 2000/43/EC of 29 June 2000 implementing the equal treatment of persons irrespective of racial or ethnic origin [2000] OJ L180/22 (the Racial Equality Directive), art 2(2)(b); Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) [2006] OJ L204/23 (the Recast Directive), art 2(1)(b); Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services [2004] OJ L373/37 (the Gender Access Directive), art 2(b); and Council Directive

Where similarly situated people with different protected characteristics receive different treatment, on the other hand, the law's approach is (in theory) more straightforward: the use of an ADMS which treats individuals less favourably on grounds of, or because of, a protected characteristic, will constitute direct discrimination. In many cases, this renders deployment of the ADMS unlawful.[7]

In this paper, we set out to challenge the persistent assumption that algorithmic decision-making systems will only be caught by the prohibition on direct discrimination in a small set of cases, such as the deployment of an automated system to camouflage intentional discrimination, or where protected characteristics are explicitly coded into an ADMS.[8] In scrutinising two paradigmatic cases of algorithmic discrimination, we demonstrate how a much broader range of ADMS may well treat individuals differently *on grounds of* a protected characteristic – and should thus fall into the scope of direct discrimination. This is not to say that *all* forms of algorithmic bias should be understood as unlawful direct discrimination. Just as there is no one technical cause of algorithmic bias, there cannot be a uniform legal answer.

Discussion proceeds as follows. The next section dissects the default assumption that algorithmic discrimination will usually fall within the scope of indirect discrimination. Originating in the US doctrine of disparate treatment, a near-exclusive focus on indirect discrimination raises the practical problem of self-justifying feedback loops, and runs counter to the principles underpinning the distinction between direct and indirect discrimination in UK and EU law.

2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation, OJ L 303 (2000) (the Framework Directive), art 2(2)(b)(i). Note that under the European Convention on Human Rights, objective justification applies to both direct and indirect discrimination: *Burden* v *the United Kingdom* (2008) 47 EHRR 38 at [60].

[7] UK discrimination law is largely standardised across domains including employment, provision of services, and education. In all cases, direct discrimination is generally not objectively justifiable, save in the case of age: Equality Act 2010, s 13(2). In the employment context, there is a narrow opportunity to justify direct discrimination for a 'genuine occupational requirement', ie where satisfaction of the criterion is strictly necessary to perform the role: Equality Act 2010, Sched 9. At EU level, the approach is less harmonised. Objective justification for direct discrimination remains possible in some specific contexts: see, for example, Gender Access Directive, art 4(5). Nonetheless, recourse to the objective justification framework is still barred in many cases of direct discrimination, including where discrimination is on grounds of sex in the employment context (Framework Directive, art 4(1); Recast Directive, art 14(2)) or on racial grounds in any regulated domain (Racial Equality Directive, art 4). Following the United Kingdom's exit from the European Union, UK courts should continue to have regard to developments in the EU equality *acquis*: European Union (Withdrawal) Act 2018, s 6.

[8] See, for example, F. Zuiderveen Borgesius, 'Price Discrimination, Algorithmic Decision-Making, and European Non-Discrimination Law' (2020) 31 *European Business Law Review* 401, 409-411; P. Hacker, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law' (2018) 55 *Common Market Law Review* 1143, 1151-1152; S. Wachter, B. Mittelstadt and C. Russell, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) 41 *Computer Law & Security Review* 105567, 19-20; J. Gerards and R. Xenidis, *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law* (Brussels: European Commission, 2020), 67-73; A. Kelly-Lyth, 'Challenging Biased Hiring Algorithms' (2021) 41 OJLS 899, 906. See also Decision no 216/2017 of the National Non-Discrimination and Equality Tribunal of Finland, issued 21 March 2018 at https://perma.cc/ZKS8-SFNJ, where factors including gender and age had been labelled as inputs in a credit scoring system.

Although the doctrine has a significant role to play in regulating algorithmic bias, the presumption against direct discrimination therefore raises practical and normative concerns. We develop this argument by reference to two paradigmatic examples of direct algorithmic discrimination: a hiring algorithm developed by Amazon to select the best applicants for software engineering positions, which scored female applicants more poorly than their equivalently qualified male counterparts; and Gender Shades, pioneering research which showed that major commercial facial recognition systems were much worse at recognising darker-skinned females than lighter-skinned males.

The third section argues that the law reflects this broader lens. Proxy discrimination – a common cause of algorithmic bias – can constitute direct discrimination in a number of clear-cut cases, such as Amazon's hiring algorithm. This conclusion entails serious implications for operators of 'black box' ADMS. Other drivers of machine bias, on the other hand, pose more difficulty for the received understanding of direct discrimination, particularly in English law: as the fourth section explains by reference to the Gender Shades case, the courts' approach to the statutory notion of differential treatment 'because of' a protected characteristic is rooted in the concept of human discrimination, and cannot deal with all relevant cases of algorithmic bias. This, in turn, raises a series of fundamental questions for the conceptual apparatus of discrimination law, developed around unobservable mental processes: how should we think about the notion of direct discrimination as a standard applied to an automated system, rather than a human decision maker?

THE DEFAULT ASSUMPTION: INDIRECTLY DISCRIMINATORY ALGORITHMS

Algorithmic bias has been the subject of extensive discussion in computer science for over a decade.[9] Legal analyses soon followed, beginning with Barocas and Selbst's seminal 2016 paper exploring regulatory responses to *Big Data's Disparate Impact*.[10] 'Finding a solution to big data's disparate impact', they argued, 'will require more than best efforts to stamp out prejudice and bias; it will require a wholesale reexamination of the meanings of "discrimination" and "fairness."'.[11] From the perspective of US anti-discrimination law, however, 'the best doctrinal hope for data mining's victims [seemed] to lie in disparate impact doctrine'.[12]

[9] B. Friedman and H. Nissenbaum, 'Bias in computer systems' (1996) 14 *ACM Transactions on Information Systems (TOIS)* 330; Pedreshi and others, n 3 above; Dwork and others, n 3 above.
[10] S. Barocas and A.D. Selbst, 'Big Data's Disparate Impact' (2016) 104 *California Law Review* 671.
[11] *ibid*, 672.
[12] *ibid*.

The broad(er) scope of direct discrimination

Under US law, this narrow lens makes sense: the utility of the disparate treatment regime is limited by the need to demonstrate discriminatory intent or explicit classification. As such, it is only apt to capture a narrow set of algorithmic discrimination cases, such as targeted proxies selected to mask discriminatory intent. As most algorithmic discrimination is unintentional, on the other hand, the 'disparate impact doctrine should be better suited to finding liability for discrimination'.[13] A rich literature on algorithmic discrimination has proceeded largely along these lines: algorithmic discrimination is acknowledged to have a disparate *impact*, while cases of disparate *treatment* are seen as few and far between.[14] In EU law, this has translated into a general focus on indirect discrimination,[15] with the prohibition on direct discrimination assumed 'likely to be less important'.[16] Where scholars have engaged with the potentially broader scope of direct discrimination in the algorithmic context, that discussion has generally been brief,[17] or has focused on associative discrimination where algorithmic systems are biased against individuals associated with – or mistakenly perceived to have – a protected characteristic.[18]

While indirect discrimination has an important role to play in combatting many forms of algorithmic discrimination, upon closer inspection, the prevailing presumption against direct discrimination proves unfounded. The influence

[13] *ibid*, 701.

[14] Including P.T. Kim, 'Data-Driven Discrimination at Work' (2017) 58 *William & Mary Law Review* 857; C.S. Yang and W. Dobbie, 'Equal Protection Under Algorithms: A New Statistical and Legal Framework' (2020) 119 *Michigan Law Review* 291; D. Hellman, 'Measuring Algorithmic Fairness' (2020) 106 *Virginia Law Review* 811; J.R. Bent, 'Is Algorithmic Affirmative Action Legal?' (2020) 108 *The Georgetown Law Journal* 803.

[15] Hacker, n 8 above; F.J. Zuiderveen Borgesius, 'Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence' (2020) 24 *The International Journal of Human Rights* 1572; R. Xenidis and L. Senden, 'EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination' in U. Bernitz and others (eds), *General Principles of EU Law and the EU Digital Order* (Alphen aan den Rijn: Wolters Kluwer, 2020); P. Hacker, E. Wiedemann and M. Zehlike, 'Towards a Flexible Framework for Algorithmic Fairness' in R.H. Reussner and others (eds), *INFORMATIK 2020* (Bonn: Gesellschaft für Informatik, 2021); Wachter, Mittelstadt and Russell, n 8 above; J. Simons, S. Adams Bhatti and A. Weller, 'Machine Learning and the Meaning of Equal Treatment' (2021) *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* 956; N. Collado-Rogriguez and U. Kohl, '"All Data is Credit Data": Personalised Consumer Credit Score and Anti-Discrimination Law' in U. Kohl and J. Eisler, *Data-Driven Personalisation in Markets, Politics and Law* (Cambridge: Cambridge University Press, 2021). On challenges in other jurisdictions, see C. Abungu, 'Algorithmic Decision-Making and Discrimination in Developing Countries' (2022) 13 *Journal of Law, Technology & the Internet*.

[16] See, for example, Wachter, Mittelstadt and Russell, *ibid*, 44-45; Zuiderveen Borgesius, *ibid*, 1576-1578.

[17] See, for example, R. Allen and D. Masters, 'Artificial Intelligence: The Right to Protection from Discrimination Caused by Algorithms, Machine Learning and Automated Decision-Making' (2020) 20 *ERA Forum* 585, 592, briefly explaining that ADMS can be inherently discriminatory.

[18] R. Xenidis, 'Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience' (2020) 27 *Maastricht Journal of European and Comparative Law* 736, 749, considering the scope of discrimination by association; S. Wachter, 'Affinity Profiling and Discrimination By Association in Online Behavioral Advertising' (2021) 35 *Berkeley Technology Law Journal* 367, 394-398, but noting at 393 that 'indirect discrimination provides a promising alternative route' given the various difficulties with bringing an (associative) direct discrimination claim.

of the US literature is, perhaps, unsurprising: the history of the two frameworks is tightly interwoven.[19] *Griggs* v *Duke Power Co*,[20] the seminal SCOTUS decision on disparate impact, played a central role in the CJEU's judgment in *Jenkins* v *Kingsgate*,[21] a conceptual founding stone in the development of indirect discrimination. Today, however, US and European discrimination law cannot be elided as easily as frequently assumed: the scope of direct discrimination is significantly wider than that of disparate treatment.[22]

While the underlying structure of direct and indirect discrimination can be broadly mapped onto the US doctrines of disparate treatment and disparate impact, there are fundamental differences for present purposes. Crucially, EU law focuses on the *reasons* or *grounds* for a decision,[23] rather than the purported discriminator's intention or motive, which is central to the US analysis.[24] The House of Lords similarly held as early as 1989 that 'the intention or motive of the defendant to discriminate … is not a necessary condition of liability' for the purposes of direct discrimination.[25] Since *unintentional* discrimination can therefore be 'direct' under European law,[26] behaviour which does not amount to disparate treatment in the US may well constitute direct discrimination in Europe: '[t]he dividing line between direct and indirect discrimination is emphatically not to be determined by some sort of *mens rea* on the part of one or more individual discriminators.'[27]

[19] For further details, see C. Tobler, *Indirect Discrimination: a case study into the development of the legal concept of indirect discrimination under EC law* (Oxford: Intersentia, 2005); D. Pannick, *Sex Discrimination Law* (Oxford: Clarendon Press, 1985).

[20] *Griggs* v *Duke Power Co* 401 US 424 (1971) (*Griggs*).

[21] Case 96/80 *Jenkins* v *Kingsgate (Clothing Productions) Ltd* ECLI:EU:C:1981:80.

[22] See Xenidis and Senden, n 15 above, 171, recognising that US notions of 'motive' and 'intent' are not relevant in EU law but still concluding that the relevance of direct discrimination is 'likely to be less important than that of indirect discrimination'. For arguments underpinning the European Union's divergent approach, see R. Allen, 'Equal treatment, social advantages and obstacles: in search of coherence in freedom and dignity' in E. Guild (ed), *The legal framework and social consequences of free movement of persons in the European Union* (Zuidpoolsingel: Kluwer Law, 1999).

[23] See for example Case C-83/14 *CHEZ Razpredelenie Bulgaria AD* v *Komisia za zashtita ot diskriminatsia* ECLI:EU:C:2013:48 (*CHEZ*) at [95]-[96]: 'if it is apparent that a measure which gives rise to a difference in treatment has been introduced for reasons relating to racial or ethnic origin, that measure must be classified as "direct discrimination" … By contrast, indirect discrimination on the grounds of racial or ethnic origin does not require the measure at issue to be based on reasons of that type.'

[24] See for example *Kentucky Retirement Systems* v *EEOC* 554 US 135 (2008), holding that an employer had not discriminated where differences in treatment were not 'actually motivated' by age, but rather by pension status; see also *Washington* v *Davis* 426 US 229 (1976), 240; *Ash* v *Tyson Foods, Inc* 546 US 454 (2006), 456. Disparate treatment can also take the form of the overt application of different criteria to different groups. R.A. Primus, 'The Future of Disparate Impact' (2010) 108 *Michigan Law Review* 1341, fn 56 thus explains that disparate treatment encompasses 'form-based' and 'motive-based' discrimination, whereas disparate impact is essentially an 'impact-based' definition of discrimination. On the other hand, a practice may have a disparate impact despite being 'neutral in terms of intent': *Griggs* n 20 above, 430.

[25] *R* v *Birmingham City Council, ex p Equal Opportunities Commission* [1989] AC 1155, 1194 (*Birmingham*).

[26] European Union Agency for Fundamental Rights, *Handbook on European non-discrimination law* (Luxembourg: Publications Office of the European Union, 2018) 239-242.

[27] *Patmalniece* v *Secretary of State for Work and Pensions* [2011] UKSC 11, [2011] 1 WLR 783 (*Patmalniece*) at [70] per Lord Walker. The difference between the European approach and the US

Hacker argues that these differences will only result in a narrow set of 'borderline case[s]' being classified as direct discrimination: where the decision maker's own biases have informed a machine learning model, for example, 'it seems more convincing to view this as a case of direct discrimination'.[28] Drawing on the US literature, however, the main technical drivers of bias – from proxy discrimination to incorrect labelling or uneven sampling – are characterised as instances of 'accidental discrimination … [and thus] not covered' by the prohibition of direct discrimination.[29] Ultimately, he contends, '[i]n machine learning contexts, direct discrimination will be rather rare … indirect discrimination is the most relevant type of discrimination'.[30]

Highlighting EU discrimination law's failure to grapple with discrimination in a data-driven world, from intersectionality to emerging novel axes of inequality,[31] Xenidis similarly suggests that direct discrimination cannot capture cases where less favourable treatment is caused by correlations between an ADMS' features and protected characteristics. Echoing Hacker, she concludes that it would be 'difficult for algorithmic proxy discrimination to be considered as direct discrimination'; instead, 'algorithmic proxy discrimination could be captured through the doctrine of indirect discrimination', despite the drawbacks of the ensuing 'larger pool of justifications'.[32]

Xenidis' concern about the pool of justifications is similar to that expressed by Barocas and Selbst in the context of US employment anti-discrimination norms:'[u]nless there is a reasonably practical way to demonstrate that [an ADMS's] discoveries are spurious, [the law] would appear to bless its use, even though the correlations it discovers will often reflect historic patterns of prejudice, others' discrimination against members of protected groups, or flaws in the underlying data.'[33] Given the ready availability of 'business necessity' as a defence, they conclude that many instances of discriminatory ADMS will thus be legal, despite the disparate impact.

approach can be seen in the respective courts' treatment of pregnancy discrimination. In Case C-177/88 *Elisabeth Johanna Pacifica Dekker* v *Stichting Vormingscentrum voor Jong Volwassenen (VJV-Centrum) Plus* ECLI:EU:C:1990:383 (*Dekker*), the CJEU held that discrimination on the basis of pregnancy constituted direct discrimination on grounds of sex because only women can be refused employment on that ground. By contrast, the Supreme Court of the United States held in *General Electric Co* v *Gilbert et al* 429 U.S. 125 (1976) 149 (*Gilbert*) that exclusion of pregnancy from a disability benefits payment plan was not disparate treatment, because there was no evidence that the employer had used pregnancy as a 'pretext' to discriminate against women. Justice Stevens, dissenting, took an approach which was very similar to that subsequently adopted by the CJEU (*ibid*, 162). Congress amended Title VII to include pregnancy in the definition of sex just two years after *Gilbert*, see Public Law 95-555, 92 Stat 2076 (codified at 42 U.S.C. § 2000e(k) (2006)).

[28] Hacker, n 8 above, 151; cf Barocas and Selbst, n 10 above 698-701, highlighting the 'surprising unanimity' within US scholarly literature that 'the law does not adequately address unconscious disparate treatment', and proceeding to analyse the issue on the basis that intentionality is required for disparate treatment.

[29] Hacker, *ibid*, 1151-1153. These examples are strikingly similar to those provided by Barocas and Selbst, *ibid*.

[30] Hacker, *ibid*, 1153.

[31] Xenidis, n 18 above; see also Wachter, n 18 above, 414-415, on new axes of discrimination.

[32] Xenidis, *ibid*, 747.

[33] Barocas and Selbst, n 10 above, 701.

Self-justifying feedback loops

This startling conclusion has clear parallels in much of European anti-discrimination law. The crucial difference between direct and indirect discrimination is the fact that the latter can, in many instances, be justified as a proportionate means of achieving a legitimate aim.[34] In simple cases, the rationale is obvious: a prison might be able to justify having a rule which bans members of staff from carrying bladed items, for example, even if that rule puts some Sikh members of staff at a particular disadvantage because they wear the kirpan.[35] In the algorithmic context, the route to justification might seem relatively straightforward: if an ADMS is deployed with a view to achieving a legitimate aim, and its predictive accuracy is high, then a finding of proportionality might appear to follow.[36] This can quickly create self-justifying feedback loops, especially in circumstances where predictions accurately track subsequent outcomes: the very mechanism which discriminates proffers the data required to justify its use.

Take the case of an ADMS used to identify geographical areas in which there is a high crime risk.[37] It is trained using records held by the relevant police force – tainted data, given the impact of factors ranging from implicit bias to unequal stop and search rates between ethnic groups on historic policing decisions.[38] The training data can therefore not provide a representative reflection of crimes committed: areas with larger Asian and black populations are over-represented;[39] the ADMS will disproportionately flag them as 'high risk' accordingly.

Once trained, the system can be deployed to facilitate 'predictive policing', such that areas flagged as 'high risk' receive additional police attention. Even if crime is evenly distributed across the population, more frequent patrols in areas with larger Asian and black populations will lead to higher arrest rates among these groups when compared with the population at large. New arrests in turn are documented by the police and fed back into the predictive policing algorithm, thus supporting the original prediction. This is not a hypothetical concern: computer scientists have documented the dangers of 'runaway feedback loops in predictive policing' when a model becomes increasingly confident in its predictions about crime distribution, even though those predictions are not reflective of reality.[40]

If analysed as a case of indirect discrimination, the predictive policing ADMS will be characterised as a *prima facie* discriminatory PCP, and the burden is on the police force to objectively justify its use of the tool. If populations which

[34] For caveats, see n 7 above.

[35] *Dhinsa* v *Serco and another* [2011] ET/1315002/09.

[36] This point is explored by Hacker, n 8 above, 1160-1163.

[37] See, for example, the software sold by Geolitica, which is in wide use in the US, see https://perma.cc/ZJ97-J5ZQ. For discussion of algorithmic bias in policing in England and Wales, see A. Babuta and M. Oswald, *Briefing Paper: Data Analytics and Algorithmic Bias in Policing* (London: Royal United Services Institute for Defence and Security Studies, 2019).

[38] See stop and search statistics for England and Wales, 22 February 2021 at https://perma.cc/5VXT-2G9J.

[39] For discussion, see K. Lum and W. Isaac, 'To Predict and Serve?' (2016) 13 *Significance* 14.

[40] *ibid*; see also D. Ensign and others, 'Runaway Feedback Loops in Predictive Policing' (2018) 81 *Proceedings of Machine Learning Research* 1.

are identified as 'high risk' do have higher than average arrest rates, this is likely to succeed. Using an ADMS which appears to accurately predict crime rates in order to effectively deploy limited police resources may well appear to be a proportionate means of achieving a legitimate aim.

This problem is not unique to policing. Indeed, it is difficult to overstate the power of such feedback loops which have long been noted in canonical works in technology law: '[P]rofiles will begin to normalize the population from which the norm is drawn, [such that the model] fits you into a pattern; the pattern is then fed back to you in the form of options set by the pattern; the options reinforce the pattern; the cycle begins again.'[41]

Biased algorithmic predictions affect real-world outcomes in a plethora of contexts, with those outcomes then fed back into ADMS to produce further predictions. In the hiring context, for example, an algorithm trained using historic performance data might disproportionately predict that men will do well in a job. If men are consequently hired, their productivity scores as employees will reinforce the model, leading to more predictions of the same type – even if women would have done just as well. Similarly, if a biased credit scoring algorithm predicts that a particular group will repay their loans, a greater number of loans will be granted to members of that group, and the group will become increasingly overrepresented in loan repayment data.[42]

Scrutinising algorithmic bias (near-) exclusively through the lens of indirect discrimination significantly exacerbates the risk of self-justifying feedback loops.[43] An indirect discrimination lens opens the door for justification in a large proportion of cases, thus potentially legitimising the wide-spread deployment of discriminatory ADMS.

The continued significance of indirect discrimination

The self-justifying feedback loop spectre is not insurmountable. Even within the traditional confines of the indirect discrimination framework, judicial acceptance of a given justification is not a foregone conclusion: the justificatory regime is not static.[44] Courts may find respondents' aims – such as mere cost

[41] L. Lessig, *Code: and Other Laws of Cyberspace, Version 2.0* (New York, NY: Basic Books, 2nd rev ed, 2006) 220.

[42] On the issue of highly personalised credit scores pre-empting and thus reinforcing outcomes, see Collado-Rogriguez and Kohl, n 15 above at 132-133.

[43] We do not suggest that all cases of algorithmic feedback loops are necessarily instances of direct discrimination. The question is whether the algorithmic outputs are different *because of* race. For example, if the police decided to patrol an area because it has a majority-black population (even if that reasoning was subconscious), and the algorithm has merely learned to automate those implicit biases, then on our analysis an individual who lives in that area and thus faces a higher chance of being unfairly arrested may have suffered direct discrimination (regardless of their own ethnicity): *CHEZ* n 23 above. See below, under the heading 'Imperfect proxies: subjective direct discrimination', on this form of algorithmic direct discrimination.

[44] For a discussion of indirect discrimination justification in the algorithmic context, see Wachter, n 18 above, 407-413.

savings – to be illegitimate.[45] Even when deployment of an ADMS serves legitimate aims, the ensuing proportionality enquiry is a necessarily contextual question. One bright-line rule is that if the same aim can be achieved through equally satisfactory but less discriminatory means, then that alternative must be used.[46] Furthermore, where a practice is classified as indirect discrimination but is found to operate in practice in a manner *akin* to direct discrimination, courts will apply a 'stringent standard of scrutiny'.[47]

Closer scrutiny of the mechanics of self-justifying feedback loops also raises a potential further objection: if an algorithm recycles biased data to generate biased outcomes, then the algorithm's 'predictive accuracy' may be entirely incapable of supporting a justification. Its predictive accuracy is only high because the world is predictably biased. A foundation for this argument can be found in *Enderby* v *Frenchay Health Authority*[48] (*Enderby*), where an employer was paying its majority female speech therapists less than its majority male pharmacists. The health authority sought to justify this difference on the basis that the two professional groups had separate collective bargaining processes, and that those bargaining processes had reached different outcomes.

This justification was swiftly rejected by the Court of Justice.[49] In his Opinion, Advocate General Lenz explained that if the Court were to accept the employer's justification, then it would afford the decision-maker 'a legal argument for maintaining the status quo'.[50] Although the collective bargaining processes were not individually discriminatory, the differing pay rates for the two groups stemmed from 'historical and social reasons', which could not be understood as being 'unconnected' from sex.[51] The same concern would apply with equal force to reliance on self-justifying feedback loops: the algorithm is

[45] See Case C-243/95 *Hill* v *Revenue Commissioners* ECLI:EU:C:1998:298; *Cross* v *British Airways Plc* [2005] EAT/0572/04/TM.

[46] *CHEZ* n 23 above at [128]. One difficulty in the predictive policing context is that where the base rates are genuinely different, then reducing the algorithm's disparate impact may reduce its predictive accuracy: Kleinberg, Mullainathan and Raghavan, n 3 above. Hacker, n 8 above, 1162, also suggests that where a machine learning algorithm has 'significant predictive accuracy, its effectiveness will likely surpass any alternative ways of decision making, particularly those based on human decision making unaided by algorithmic computing power.'

[47] *Secretary of State for Defence* v *Elias* [2006] EWCA Civ 1293, [2006] WLR 3213 (*Elias*) at [161]; see also [153] and [158]-[163]. Note that the facts of *Elias* seem to be more in line with the concept of direct discrimination, but that the caselaw on inherently discriminatory proxies for national origin is somewhat of an anomaly on this point: see *Patmalniece* n 27 above at [69], and discussion at n 91 below.

[48] C-127/92 *Enderby* v *Frenchay Health Authority* [1993] ECLI:EU:C:1993:859. *Enderby* was an equal pay case, but equal pay caselaw is highly significant in the context of indirect discrimination claims: see the similarities between Equality Act 2010, s 19 and s 69(1)(b) and (2). In equal pay cases, the employer must identify a 'material factor' for a pay differential; if there is evidence that the factor is in some way 'tainted' by indirect sex discrimination, then the objective justification scheme applies: see *Glasgow City Council and others* v *Marshall and others* [2000] UKHL 5, [2000] IRLR 272.

[49] *Enderby ibid* at [20]-[23].

[50] C-127/92 *Enderby* v *Frenchay Health Authority* ECLI:EU:C:1993:313, Opinion of AG Lenz at [49]. UK courts have also held that a justification cannot itself be tainted by protected characteristics: *Mandla* v *Dowell Lee* [1983] 2 AC 548, 566; *Orphanos* v *Queen Mary College* [1985] AC 761, 772. See however *R (on the application of E)* v *JFS Governing Body* [2009] UKSC 15, [2010] 2 AC 728 (*JFS*) at [206] per Lord Hope, dissenting.

[51] Opinion of AG Lenz, *ibid*.

perpetuating and exacerbating a discriminatory status quo, and its outcomes reflect social facts – such as over-policing of certain communities – which are not unconnected from ethnicity. If judges are alive to this issue, then it is entirely possible that they would take as firm a line as the Court did in *Enderby*.[52]

A normative concern

The presumption that algorithmic bias will generally only give rise to indirect discrimination is furthermore difficult to square with the principled distinction between direct and indirect discrimination developed by the courts.

Drawing this line has given rise to significant debate in the literature.[53] Two distinguishing factors can nonetheless be identified as particularly salient in the case law. First, the prohibition on direct discrimination is intended to achieve formal equality, whereas the rules on indirect discrimination seek to advance substantive equality. Secondly, direct discrimination is reason-focussed, whereas indirect discrimination is effects-focussed. When applied to canonical examples of algorithmic bias, the underlying indirect discrimination model provides a surprisingly poor fit, since biased ADMS exhibit characteristics of direct discrimination in multiple dimensions.

The case law describes direct discrimination as seeking to ensure that like candidates are treated alike, regardless of their protected characteristics. The prohibition on direct discrimination therefore performs an exclusionary function: it excludes certain protected characteristics from the range of permissible reasons a decision-maker may legitimately rely upon in order to treat one individual less favourably than another.[54] This is formal equality.[55] Indirect discrimination tackles a broader class of reasons for decision-making, namely a provision, criterion, or practice which is applied to everyone but has an unjustifiably disproportionate impact on a group with protected characteristics. Indirect discrimination law therefore looks beyond formal equality, towards a more substantive equality of results.

Because the two regimes have different end goals, they operate in different ways. While direct discrimination focuses on the *reason* for a harm, indirect discrimination focuses on the disparate *effects* of facially neutral criteria.[56] In

[52] This, of course, depends on whether the workings of the algorithm are cognisable, a significant practical issue which falls beyond the scope of this paper.

[53] See S. Fredman 'Direct and Indirect Discrimination: Is There Still a Divide?' in H. Collins and T. Khaitan (eds), *Foundations of Indirect Discrimination Law* (Oxford: Hart Publishing, 2018) 39-46; J.M. Finnis, 'Directly discriminatory decisions: a missed opportunity' (2010) 126 LQR 491. The Canadian Supreme Court has abolished the distinction altogether: *British Columbia (Public Service Employee Relations Commission)* v *BCGEU* [1999] 3 SCR 3. However, although the line has become increasingly blurry (see, for example, Joined Cases C-804/18 and C-341/19 *IX* v *WABE eV* and *MH Müller Handels GmbH* v *MJ* ECLI:EU:C:2021:594 (*WABE*) and discussion below), the distinction remains fundamentally important to the European legal framework.

[54] Case C-303/06 *S. Coleman* v *Attridge Law and Steve Law* ECLI:EU:C:2008:61 Opinion of AG Poiares Maduro at [18].

[55] *JFS* n 50 above at [56].

[56] *Essop* v *Home Office* [2017] UKSC 27, [2017] 1 WLR 1343 (*Essop*) at [1]; *Nagarajan* v *London Regional Transport* [2000] 1 AC 501 (*Nagarajan*); *McFarlane* v *Relate* [2010] EWCA Civ 880,

some cases, this might entail a focus on mental processes. As Lord Philips put it in *JFS*: 'A fat black man goes into a shop to make a purchase. The shop-keeper says "I do not serve people like you". To appraise his conduct it is necessary to know … [w]as it the fact that the man was fat or the fact that he was black?'[57]

However, it is not *necessary* for the protected characteristic to feature in the decision-maker's mental processes: direct discrimination will also occur if the decision-maker uses a criterion which is inherently discriminatory.[58] When Eastleigh Borough Council offered free swimming pool access to pensioners, it did not intend to discriminate against Mr James. Nonetheless, because parliament had set different pensionable ages for men and women, the Court held that pensionable age had become a 'shorthand expression which refers to the age of 60 in a woman and to the age of 65 in a man'.[59] Discriminating by pension age was thus effectively equivalent to discriminating by sex; the case was one of direct sex discrimination. Similarly, a rule which excludes pregnancy from a health insurance plan is inherently directly discriminatory against women, even if the decision is made for purely financial reasons.[60] The courts have thus 'sever[ed] the notion of direct discrimination from its moral anchor in fault and responsibility of the perpetrator'.[61] Direct discrimination does not require any moral wrongdoing.[62]

Applying these normative distinctions to algorithmic bias, we can see that direct discrimination provides a more pragmatic starting point than is frequently recognised. There are two main mechanisms driving bias in machine learning: proxy discrimination (which arises because of tainted target variables), and sampling bias (which arises from unrepresentative training data sets).[63]

Take Amazon's infamous hiring algorithm as a clear example of the former. The company set out to select the best applicants for software engineering positions. To do so it used a machine learning algorithm which looked for patterns in the historic applicant data. In the past, there had been more successful male software engineering applicants – so some of the correlations between applicant traits and the likelihood of success were linked to sex, rather than aptitude. Amazon's system quickly learned to penalise applications from graduates of two all-women's colleges.[64] In normative terms, penalising the alumnae of women's colleges leads to formal inequality between men and women: equally

[2010] IRLR 872 at [18] per Laws LJ; C. Campbell and D. Smith, 'The Grounding Requirement for Direct Discrimination' (2020) 136 LQR 258; cf T. Khaitan, *A Theory of Discrimination Law* (Oxford: OUP, 2015) ch 6, arguing that all of discrimination law is effects-oriented.

[57] *JFS* n 50 above at [21].

[58] *James* v *Eastleigh Borough Council* [1990] UKHL 6, [1990] 2 AC 751 (*James*); *JFS ibid*; *WABE* n 53 above.

[59] *James ibid*, 764C.

[60] *Dekker* n 27 above.

[61] Fredman, n 53 above, 39-46, commenting on *James* n 58 above and *JFS* n 50 above.

[62] *JFS ibid* at [124] per Lord Kerr: '[i]t is plain that the Chief Rabbi and the governors of JFS are entirely free from any moral blame.' This has been challenged in the literature. See, for example, Finnis, n 53 above, arguing that the idea of inherently discriminatory criteria muddies the waters of discrimination law.

[63] There are also other mechanisms, such as model bias, which might overlap and interact with each other. For a fuller technical discussion, see S. Mitchell and others, 'Algorithmic fairness: Choices, assumptions, and definitions' (2021) 8 *Annual Review of Statistics and its Application* 141.

[64] Dastin, n 2 above.

well-qualified women received lower scores than similarly situated men because of a proxy for their sex.

Turning to the second main mechanism of bias, unrepresentative sampling data can have similarly discriminatory outcomes, albeit through distinct underlying mechanisms. In Gender Shades,[65] Buolamwini and Gebru demonstrated that major commercial gender classification systems were much worse at recognising darker-skinned females than lighter-skinned males, with maximum error rates of 34.7 per cent for the former group and 0.8 per cent for the latter.[66] Error rate differences were also significant along lines of race or gender alone.[67] This was driven by the fact that there were fewer black women in the training data than white men. In 'learning' to recognise faces, the algorithm's underexposure to black female faces meant that its performance was poorer in that area. Deployment of such software can lead to less favourable treatment in regulated contexts. Uber, for example, requires its drivers to complete identity verification checks before driving, and drivers for whom the software does not work can face being barred from the platform.[68]

If a black woman is disadvantaged because a facial recognition algorithm fails correctly to classify her in circumstances where it would recognise a similarly situated white person (in similar lighting conditions, for example), then at least from a theoretical perspective this seems to go beyond the concept of indirect discrimination.[69] Similarly situated individuals are not being treated alike, and the reasons for the system's imbalance *are* race and gender: accuracy is low because few black women are included in the training data. In other words, in both cases, a normative analysis along the distinguishing lines identified by the courts points towards direct discrimination: *formal inequality* has arisen, and race and gender are among the *reasons* for that inequality.

Despite this strong normative case, the legal classification of such cases is far from straightforward, not least because the concepts of direct and indirect discrimination are not hermetically sealed: the 'distinction … is hard to draw on a conceptual level'.[70] What is clear, however, is that there are cases of indirect proxy discrimination and sampling bias which do intuitively go beyond the scope of indirect discrimination. The question, then, is whether the legal literature has been too narrow in its discussion of algorithmic bias. Do cases like

[65] Buolamwini and Gebru, n 3 above.

[66] *ibid*. Subsequent improvements significantly reduced these discrepancies: see I.D. Raji and J. Buolamwini, 'Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products' (2019) *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* 429.

[67] Buolamwini and Gebru, *ibid*. On the treatment of intersectional discrimination under EU law, see European Union Agency for Fundamental Rights, n 26 above, 59-63.

[68] Vallance, n 4 above.

[69] In legal terms, however, it is possible to argue that the algorithm is a PCP which puts certain groups at a particular disadvantage: see 'Challenge Uber's discriminatory use of facial recognition systems' CrowdJustice, posted by The App Drivers and Couriers Union, undated at https://perma.cc/MK9C-ZAAV, indicating that an indirect discrimination case has been brought against Uber, challenging its use of facial recognition technology.

[70] H. Collins and T. Khaitan, 'Indirect Discrimination Law: Controversies and Critical Questions' in Collins and Khaitan, n 53 above, 18-19. For discussion of difficulties drawing this line in the algorithmic context, see Wachter, n 18 above, focussing on associative discrimination as a means of closing the gap.

the Amazon hiring algorithm and the facial recognition technologies discussed by Gender Shades fit within the *legal* definition of direct discrimination?

DIRECT ALGORITHMIC DISCRIMINATION

Automated direct discrimination can arise without any proxies: a surprising number of systems have been shown to rely directly on protected characteristics in assigning study places,[71] generating credit scores,[72] or screening for benefit fraud.[73] From a legal perspective, these cases are clear-cut instances of direct discrimination. Algorithmic discrimination, however, is not always so straightforward. What if bias has arisen because salient features in the model are correlated with protected characteristics, as seen in the Amazon case? The law has long grappled with situations where discrimination is clothed by closely correlated factors: it is unlawful to exclude civil partners from renting B&B rooms available to married couples, for example. As Lady Hale explained in *Lee* v *Ashers*, direct discrimination occurs whenever the 'criterion used as the reason for less favourable treatment is not the protected characteristic itself but some proxy for it.'[74]

In this section, we examine different instances of biased automated decision-making to demonstrate how at least some cases of proxy discrimination do, in fact, fit into existing categories of direct discrimination.[75] Two sub-types of direct discrimination are identified in the case law: decisions made using an *inherently discriminatory* criterion, and decisions made through *subjectively discriminatory* mental processes.[76] Inherent discrimination occurs when a criterion used by a decision-maker is inextricably linked to a protected characteristic. Subjectively discriminatory decision-making arises when a person's protected

[71] Commission for Racial Equality, *Medical School Admissions: Report of a formal investigation into St George's Hospital Medical School* (London: CRE, 1988), finding that both 'non-Caucasian' and 'female' carried negative weightings.

[72] Decision no 216/2017 of the National Non-Discrimination and Equality Tribunal of Finland, n 8 above.

[73] The Dutch Tax and Customs Administration used an algorithm to detect benefits fraud which included nationality as a risk factor: see Amnesty International, n 1 above. Racial profiling allegations did not result in legal action, but the scandal did result in the fall of the cabinet in 2021. Nationality is a protected characteristic under UK law. For another example of nationality being used in ADMS, see 'Home Office drops "racist" algorithm from visa decisions' n 4 above.

[74] *Lee* v *Ashers Baking Co* [2018] UKSC 49, [2018] 3 WLR 1294 at [25].

[75] Xenidis does recognise that 'some proxies have been accepted as falling within the scope of given protected grounds by the Court of Justice', but notes that '[t]he degree of overlap required between a proxy and a given protected group to give rise to discrimination is unclear', n 18 above, 746. Beyond this absence of clarity, the reason given for the exclusion of *algorithmic* proxy discrimination is that the definition of direct discrimination 'involves a causality link between a given treatment and a protected ground, while inferential analytics rely on correlations', *ibid*, 747. There is no need for a causal link between the *selection* of the proxy and the protected characteristic, however: the proxies in *James* n 58 above and *Dekker* n 27 above were not applied because of their inherently discriminatory nature, but were *so strongly correlated* to protected grounds that the affected individuals were still held to have suffered less favourable treatment 'because of' their protected characteristics. This is the significance of intention being irrelevant under EU law.

[76] *JFS* n 50 above at [78] per Lord Phillips.

characteristic influences the decision-maker's conscious or subconscious mental processes, such that a different outcome is reached.[77]

Perfect proxies: inherently discriminatory direct discrimination

We have already encountered the case of *James* v *Eastleigh BC*, in which a rule based on the statutory retirement age was held to be directly discriminatory on grounds of sex. The Council's reason for adopting the policy was irrelevant to the analysis; direct discrimination followed from the mere application of this inherently discriminatory criterion.[78] Subsequent cases have confirmed that if there is 'no doubt as to the factual criteria' forming the basis for the decision, then there is no need to examine the mental processes of the alleged discriminator.[79] On this basis, there are at least three technical ways in which ADMS can effect direct proxy discrimination.

Trained Proxy

The most straightforward example of proxy discrimination occurs where an inherently discriminatory criterion is directly coded into the algorithm. Take, for example, a bank automating its mortgage application assessments. If the bank can identify from past repayment data that marital status is correlated with likelihood of repayment, it might include it as a feature in the algorithm.[80] If 'marriage' is defined to exclude civil partnerships, then a same-sex couple could receive a score that is lower than that of a similarly situated heterosexual couple. If the couple in a civil partnership is consequently denied a mortgage, direct discrimination has occurred: marital status has been held to be inextricably linked to sexual orientation if marriage is open only to heterosexual couples.[81] An inherently discriminatory criterion cannot be any less discriminatory merely because it is applied by a computer system, rather than by a human through a paper process.

[77] *Nagarajan* n 56 above; *JFS* n 50 above at [64]. While the CJEU has not delineated these two 'types' of direct discrimination as clearly as the UK Supreme Court, its caselaw proceeds broadly on the same lines. For examples of the Court considering subjectively discriminatory decision-making, see *CHEZ* n 23 above; C-54/07 *Centrum voor gelijkheid van kansen en voor racismebestrijding* v *Firma Feryn NV* ECLI:EU:C:2008:397 (*Firma Feryn*). For examples of inherently discriminatory criteria, see C-267/06 *Maruko* v *Versorgungsanstalt der deutschen Bühnen* ECLI:EU:C:2008:179 (*Maruko*); Case C–356/09 *Pensionsversicherungsanstalt* v *Kleist* ECLI:EU:C:2010:703 (*Kleist*); *Dekker* n 27 above.

[78] *James* n 58 above, 669F-H.

[79] *JFS* n 50 above at [23].

[80] Note that marital status is a protected characteristic in UK (Equality Act 2010, s 4), but protection is not comprehensive across the EU.

[81] *Maruko* n 77 above; *Bull* v *Hall* [2013] UKSC 73, [2013] 1 WLR 3741 at [16]-[30]. Note that the claim would not succeed if the couple's relationship was akin to that of an unmarried heterosexual couple, rather than a married couple; or if civil partnership were available to same-sex couples, and the respondent's definition of 'marriage' included such partnerships: in these cases, less favourable treatment is not occurring on grounds of sexual orientation, but on some other grounds.

Learnt Proxy

Machine learning algorithms can also draw their own correlations between datapoints. Given that intentionality is irrelevant when establishing direct discrimination, the outcome should be no different if the algorithm, rather than the human trainer, created the indissociable proxy. The discrimination stems from the application of the criterion, not from the mental processes of the decision-maker.[82]

Consider, for example, the Amazon recruitment algorithm discussed above, which learned to penalise graduates of two all-women's colleges. Sex and sex-selective education are inextricably linked. In *R* v *Birmingham City Council* (*Birmingham*), a local authority provided 360 grammar school places for girls, and 540 for boys.[83] As a result, girls receiving a borderline test mark in the grammar school entrance exam had a substantially smaller chance of obtaining grammar school education than boys with comparable marks. In essence, those wishing to attend a girl's grammar school had to get a higher grade than those wishing to go to a boy's grammar school. The House of Lords held that it was 'because of their sex that the girls in question receive[d] less favourable treatment than the boys', and as such, direct discrimination had occurred '[w]hatever may have been the intention or motive of the council'.[84] Holding graduates from two women's colleges to a higher standard, as the Amazon algorithm did, is similarly direct discrimination on grounds of sex, regardless of the decision-maker's intentions.[85] Both criteria – applying to a girls' grammar school and graduating from a women's college – are inherently directly discriminatory.

Latent Variable Proxy

In the previous two examples, the criterion applied by the algorithm was readily comprehensible in human analysis: we can understand why direct discrimination might arise from differential treatment of marriage and civil partnerships, or women's colleges and other colleges. Algorithms, however, are infinitely better than humans at amassing data and analysing it for correlations. As such, while the majority of inherently discriminatory rules created by humans will also be intuitively recognisable as discriminatory, the same may not be true of inherently discriminatory rules created by algorithms.

[82] In *James* n 58 above, 781F-H, per Lord Lowry (dissenting) considered, and dismissed, the relevance of foreseeability of the discriminatory effect of an inherently discriminatory criterion: 'foreseeability … adopted by analogy with the criminal law … is not the appropriate test'. Lord Lowry adopted a purely subjective interpretation of the statute; the other judges did not discuss foreseeability or knowledge as necessary criteria. Foreseeability has not been discussed in more recent caselaw, and mental processes have been held irrelevant when assessing inherent discrimination, see *JFS* n 50 above at [23]. Note, however, the discussion at text to n 171 below. For a discussion of 'reckless' reliance on tainted data, see T.Z. Zarsky, 'An Analytic Challenge: Discrimination Theory in the Age of Predictive Analytics' (2018) 14 *I/S: A Journal of Law and Policy for the Information Society* 11, 24-25.

[83] *Birmingham* n 25 above.

[84] *ibid*, 1194C-D per Lord Goff; see also *R (on the application of Coll)* v *Secretary of State for Justice* [2017] UKSC 40, [2017] 1 WLR 2093 (*Coll*).

[85] See also *JFS* n 50 above at [89] per Lord Mance: 'an organisation which admitted all men but only women graduates would be engaged in direct discrimination on the grounds of sex.'

Davies and Douglas, for example, explain how a machine learning tool could 'learn a perfect proxy for race – a combination of variables that fully captures the correlation between race and measured recidivism, such that including race over and above this combination would have no effect on risk classifications'.[86] If that combination is applied when predicting the likelihood of recidivism in the context of a sentencing decision, 'we might legitimately wonder whether direct discrimination had been avoided, or avoided in any more than name'.[87]

The legal answer is straightforward: direct discrimination has *not* been avoided. We have already established that a criterion can be inherently discriminatory even if it bears no facial relationship to a protected characteristic.[88] The only question is whether the criterion is inextricably linked to a protected characteristic, and a one-to-one correspondence between a failure to satisfy the requirement and the relevant protected characteristic will be dispositive,[89] or at least highly persuasive,[90] in answering this question in the affirmative.

The fact that the algorithm posited by Davies and Douglas involves the application of multiple criteria, rather than just one inherently discriminatory criterion, is irrelevant.[91] Imagine, for example, a formerly boys-only secondary school which began admitting girls in 2010. Now imagine that an employer will only hire a job applicant if they are (i) a graduate of that school and (ii) were born before the year 1990. Neither of those criteria is inherently discriminatory on grounds of sex when taken individually: some of the school's graduates are women, and so are about half of the population born before 1990. But no woman meets both criteria when applied together: the set of cumulative criteria constitutes an inherently discriminatory rule.

Returning to the hypothetical posited by Davies and Douglas, we can therefore conclude that where an algorithm identifies a (possibly very large) set of factors which, when applied together, constitute a perfect proxy for a protected characteristic, then direct discrimination will arise – even if no human would have realised that this result would occur.[92]

We pause at this juncture to note a potential limitation which arises from the UK caselaw. Courts have generally held that a criterion will only be in-

[86] B. Davies and T. Douglas, 'Learning to Discriminate: The Perfect Proxy Problem in Artificially Intelligent Criminal Sentencing' in J. Ryberg and J.V. Roberts (eds), *Principled Sentencing and Artificial Intelligence* (Oxford: OUP, 2022).

[87] *ibid*.

[88] Such as pensionable age and sex: *James* n 58 above; and *Kleist* n 77 above.

[89] In UK law.

[90] In EU law.

[91] For the reverse position, see Case C-73/08 *Nicolas Bressol and Others and Céline Chaverot and Others* v *Gouvernement de la Communauté française* ECLI:EU:C:2010:181 (*Bressol*), in which individuals were required to meet multiple requirements, one of which could be met by all nationals but only some non-nationals. The Court considered the conditions as a whole and analysed the case under the indirect discrimination framework. In contrast, the Advocate General's Opinion identified this criterion as directly discriminatory: *ibid* at [64]-[76] Opinion of AG Sharpston. The Court's judgment in *Bressol* subsequently attracted strong criticism from the UK Supreme Court in *Patmalniece* n 27 above at [32]-[33] per Lord Hope, at [63]-[64] per Lord Walker. Note that this line of caselaw specifically relates to EU law precluding discrimination against nationals from different Member States, which is distinct from the law relating to the supply of employment, goods, and services: *Patmalniece ibid* at [83].

[92] On the relevance of foreseeability, see n 82 above.

herently discriminatory where there is an 'exact correspondence' between the group disadvantaged by its application, and a group defined by the protected characteristic.[93] The meaning of this distinction is not entirely clear. It seems that if 100 per cent of those affected by a given criterion also have a particular protected characteristic, then the criterion is inherently discriminatory, but if the proportion is only 99 per cent, then the application of the criterion is potentially justifiable indirect discrimination: hardly a principled distinction.[94]

The three examples analysed above fit within the narrow bounds of the UK courts' standard: all job applicants disadvantaged by the downgrading of women's colleges are women; all loan applicants affected by the differential treatment of marriage and civil partnership are in same-sex relationships;[95] and in Davies and Douglas' hypothetical case, all individuals affected by the precise combination of factors forming the perfect proxy are of a particular race. In each case, one can imagine a claimant who – had they not been of a particular gender, sexual orientation, or ethnicity – would have been spared a disadvantage. In practice, however, there will be many cases in which ADMS mark down an individual on the basis of multiple factors which are *imperfectly* correlated with a particular protected characteristic. Will the 'one-to-one' requirement be fatal to a finding of inherent discrimination in such cases? We suggest not.

First, even on its own terms, the UK's 'perfect proxy' requirement may not be as rigid as it seems. Two points demonstrate this. First, in the *Birmingham* case, only a subset of girls was affected, namely those who wished to attend a grammar school in Birmingham; Campbell and Smith similarly argue that in *James* there was no exact correlation between the adverse treatment (having to pay) and the protected group (men), because some women (ie those below pensionable age) did have to pay to enter the swimming pool.[96] In other words, the disadvantaged group did have some women in it; it was only comprised entirely of men *within the 60–65 age category*.

A similar point was noted by Langstaff P in *Chief Constable of West Midlands Police and others* v *Harrod and others*: '[W]here it is clear that a *large cohort* simply was excluded by application of an age-related criterion, applied as a threshold provision, that seems … to constitute direct discrimination against either the group excluded by falling below the threshold, or the group of others (above the threshold) who are defined by not being so excluded.'[97]

Second, some of these girls did meet the higher threshold and were admitted to the school. The UK Supreme Court has since recognised that 'it cannot be a

[93] *Essop* n 56 above at [17]; *Lee* v *Ashers* n 74 above at [25]; *Patmalniece* n 27 above at [29].

[94] Collins and Khaitan, n 70 above, 20.

[95] This will be the case if civil partnership is not open to opposite-sex couples, n 81 above. It is unclear whether the UK Supreme Court would find that civil partnership is inextricably linked to sexual orientation following The Civil Partnership (Opposite-sex Couples) Regulations 2019, SI 2019/1458, as the 'one-to-one correlation' standard can no longer be met.

[96] Campbell and Smith, n 56 above, 269-270. See also *Patmalniece* n 27 above at [64] per Lord Walker. For treatment of subsets at the EU level, see Case C-16/19 *VL* v *Szpital Kliniczny im. Dra J. Babińskiego Samodzielny Publiczny Zakład Opieki Zdrowotnej w Krakowie* ECLI:EU:C:2021:64.

[97] *Chief Constable of West Midlands Police and others* v *Harrod and others* [2015] UKEAT/0189/14/DA at [49] (emphasis supplied), on a compulsory retirement scheme based on length of service and age discrimination.

requirement of direct discrimination that all the people who share a particular protected characteristic must suffer the less favourable treatment complained of'.[98]

These two points are relevant to the Amazon case, in which (i) a subset of women were disadvantaged, namely those who attended one of two sex-selective colleges and wished to work at Amazon; and (ii) some of those women may well have overcome the disadvantage of having their applications penalised, and still been offered a job at Amazon. In neither case do these factors change the analysis. In sum, the 'exact correspondence' might be satisfied if a *subset* of individuals within a protected group suffers a *higher risk* of receiving less favourable treatment. This is a less exacting test than appeared at first blush.[99]

Second, the CJEU has taken a less rigid approach to inherent discrimination than the UK courts.[100] Consider, for example, the recent joined cases of *WABE and MH Müller Handel*, in which the Court considered rules adopted by two employers.[101] Both rules prohibited employees from wearing visible political, philosophical, or religious signs at work. The rules primarily affected Muslim women, and the question was whether such a rule constituted direct discrimination on grounds of religion. The Court held that 'neutrality' rules do not constitute direct discrimination as long as they are applied to all political, philosophical, and religious beliefs in a general and undifferentiated way.[102] However, the Court went on to consider the fact that one rule in issue only applied to *conspicuous, large-sized* signs.[103] Here, the Court's approach was different. Holding that such a rule might be 'inextricably linked to one or more specific religions or beliefs',[104] the Court determined that it was 'liable to constitute direct discrimination on the grounds of religion or belief'.[105] This was the case even though the rule would also prohibit the expression of *political* views with conspicuous signs; in other words, there was not *necessarily* a 'one-to-one' correlation between those affected by the rule and those holding religious beliefs.

It is true that the CJEU's more flexible approach might give rise to counterintuitive results, with the Court finding no direct discrimination in circumstances where the relevant criterion does appear to be inherently discriminatory.[106] Even in *WABE*, it is not entirely clear why a rule against *all* visible

[98] *Coll* n 84 above at [29]-[31]. Note that in *Coll* the criterion *was* the protected characteristic, rather than a proxy for it, so there was no need to consider whether there was an 'exact correspondence' between the criterion and the protected characteristic before assessing the disadvantageous effect of the criterion, see *ibid* at [28]-[29].

[99] The fact that this description begins to sound similar to that of indirect discrimination indicates how difficult it is to draw a bright-line distinction between the concepts. We return to this challenge in the section below headed 'The complexity of traceability' and in the Conclusion.

[100] Or perhaps more analytical: see *Patmalniece* n 27 above, in which Lord Hope and Lord Walker concluded that the CJEU, in *Bressol* n 91 above, must have regarded AG Sharpston's excellent Opinion as 'too analytical'.

[101] *WABE* n 53 above.

[102] *ibid* at [55].

[103] *ibid* at [71]-[78].

[104] *ibid* at [73].

[105] *ibid* at [78]. Even if direct discrimination were not established, the Court held that the indirect discrimination would be unjustifiable: *ibid* at [74].

[106] See C-79/99 *Schnorbus* v *Land Hessen* ECLI:EU:C:2000:676; Case C-457/17 *Heiko Jonny Maniero* v *Studienstiftung des deutschen Volkes eV* ECLI:EU:C:2018:912; and Case C-668/15 *Jyske*

manifestations of belief is not indissociable from religion, while a rule against only *conspicuous* manifestations is.[107] For present purposes, however, the case law consistently demonstrates that a 'one-to-one' requirement is not as lethal as it might seem when considering whether criteria applied by an algorithm are inherently discriminatory.

In any event, finally, even under current UK jurisprudence, inherent discrimination is not the only route to challenging algorithmic bias. There is a further recognised category of direct discrimination, *viz* subjective discrimination, for which the 'one-to-one' test has never been relevant, and which will therefore cover at least some cases of imperfect proxy discrimination.[108]

Imperfect proxies: subjective direct discrimination

Some decisions are discriminatory because the protected characteristic played a part in the decision-maker's mental processes. There is no strict legal test for establishing whether this is the case, and the protected characteristic does not have to be the only or even the main cause of the result complained of; it is enough that it was *a* cause.[109] Moreover, the subjective mental processes which constitute direct discrimination need not be conscious:[110] '[t]here are … cases in which the ostensible criteria is something [other than a protected characteristic] – usually, in job applications, that elusive quality known as "merit"'. In such cases, 'the discriminator may … unconsciously be making his selections on the basis of race or sex. He may not realise that he is doing so, but that is what he is in fact doing.'[111]

In certain instances, a machine learning algorithm can operate as an *automated version* of this kind of unconscious human bias. In order to avoid an inconsistent application of the law in these cases, such scenarios should therefore be conceptualised as situations of subjective direct discrimination. Take the example of a recruiter who, on receiving a job application from a woman, subconsciously takes a dimmer view of it. He does not object to anything specific in the application, but multiple indicators of gender cumulatively affect his overall impression. The recruiter might, for example, be more impressed by a candidate who

Finans A/S v *Ligebehandlingsnævnet* ECLI:EU:C:2017:278; and see Lady Hale expressing 'surprise' about the CJEU's approach in *Patmalniece* n 27 above at [90].

[107] E. Howard, 'Headscarf-wearing employees and the CJEU: what employers can and cannot do' (2021) 22 *ERA Forum* 687.

[108] Perfect proxies (or the protected characteristic itself) could be applied with the intention of producing a discriminatory effect, in which case subjective direct discrimination will also arise. However, such proxies are likely to be litigated as inherent direct discrimination in the first instance, as the criterion will therefore be clear: see *JFS* n 50 above.

[109] *O'Neill* v *Governors of St Thomas More Roman Catholic Voluntary Aided Upper School* [1996] EAT/1180/94; EHRC Employment Statutory Code of Practice (2011), para 3.11.

[110] *Nagarajan* n 56 above.

[111] *JFS* n 50 above at [64]. The fact that the scope of direct discrimination covers implicit bias has not been recognised as explicitly at the EU level, but the Court has held, for example, that an employer's failure to explain its decision to hire a particular candidate could be a relevant factor when determining a direct discrimination case brought by another candidate: Case C-415/10 *Galina Meister* v *Speech Design Carrier Systems GmbH* ECLI:EU:C:2012:217.

mentions that he is the captain of a rugby team (a majority-male sport) than one who mentions that she is the captain of a netball team (a sport played mainly by women). Similarly, he could be more impressed by applications containing strong statements with active verbs; statements which are more frequently deployed by men.[112] In short, the employer is influenced by his subconscious belief that men are a better 'fit' for the relevant role. A female applicant who misses out on the job to a similarly qualified man would, in these circumstances, have a claim for direct discrimination.[113]

Now translate this into the algorithmic context. In many cases, algorithms are making judgements which could previously only have been made by humans.[114] In the same vein, they are also learning to discriminate like humans. Returning once more to the example of the Amazon algorithm, in addition to penalising the names of women's colleges, it also learned to *mark up* applicants with indicators of masculinity – such as those using more active verbs.[115] In other words, the algorithm learned to prefer applications which exhibited male traits. It preferred these traits not because they were genuinely indicative of software engineering skills, but because they were indicative of masculinity: the algorithm had learned to copy the implicit biases of past human decision-makers.[116] The legal position cannot be any different because unfavourable treatment is meted out by an algorithm, rather than a human: where a process is automated, and the ADMS exhibits what would – in a human context – be termed implicit bias, then the category of 'subconscious discrimination' should be expanded to cover it.[117]

[112] C. Leaper and R.D. Robnett, 'Women Are More Likely Than Men to Use Tentative Language, Aren't They? A Meta-Analysis Testing for Gender Differences and Moderators' (2011) 35 *Psychology of Women Quarterly* 129, finding a small but statistically significant difference between the tentativeness of language used by women, as compared with men.

[113] This is a form of direct discrimination, see *King* v *The Great Britain-China Centre* [1991] EWCA Civ 16, [1991] IRLR 513 (*King*) at [36].

[114] These include identifying potential fraudulent benefits claims, assessing job applications, and analysing the risk of recidivism in bail applications: see S. Marsh and N. McIntyre, 'Nearly half of councils in Great Britain use algorithms to help make claims decisions' *The Guardian*, 28 October 2020 at https://perma.cc/B5MB-79VF; J.B. Fuller and others, *Hidden Workers: Untapped Talent* (Accenture / Harvard Business School, 2021) 20-25; A.L. Park, 'Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing' *UCLA Law Review – Law Meets World* 19 February 2019 at https://perma.cc/9M6D-YMFV. In the EU and UK, significant decisions require human involvement: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (GDPR), art 22.

[115] Dastin, n 2 above; see also R. Steed and A. Caliskan, 'Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases' (2021) *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 701.

[116] Amazon has not disclosed whether its algorithm's bias stemmed solely from skewed sampling data (since most software engineers are men) or also from historically biased decision-making, and indeed drawing this distinction would be difficult as it would require identification of past implicitly biased decision-making. Since we know that implicit biased labelling is a common problem, we use the example here for clarity of explication.

[117] Note, also, the relationship between this incremental expansion and the exclusion of reliance on a justification which is itself tainted by a protected characteristic, discussed above in the section headed 'The continued significance of indirect discrimination'. On either reading, the use of the ADMS is unlawful.

The risks of black box bias

Whether analysed as inherently discriminatory criteria or automated subjective discrimination, the implications of our argument thus far are stark: a series of automated decision-making systems, which have hitherto been seen as indirectly discriminatory and therefore potentially justifiable, may in fact be directly discriminatory and thus unlawful in the EU and the UK.[118] A *prima facie* finding of direct discrimination affords the decision-maker very little scope for justification. In practical terms, if a claimant can provide facts from which the court or tribunal might presume that direct discrimination has occurred, then the burden of proof shifts to the respondent to provide an adequate non-discriminatory explanation for its actions.[119] In other words, rather than showing that the use of the algorithm can be objectively justified (as would be the case with indirect discrimination),[120] the decision-maker will have to show that the unfavourable algorithmic output was *not* because of a protected characteristic. An inability to provide such proof will result in a finding in favour of the claimant.

Whilst formally limited to EU and UK law, the practical reverberations of this conclusion are likely to be felt globally: whilst many ADMS are sold by US-based developers for deployment in Europe, previous work has highlighted the need for vendors to consider differences between the US and EU discrimination frameworks in designing algorithmic management tools.[121] As increasing numbers of biased black-box systems are found to fall foul of the prohibition on direct discrimination in the EU, a 'Brussels effect' may well lead to fundamental changes in terms of which systems are placed on the global market.[122]

What facts will suffice to raise the presumption of direct discrimination is a contested question.[123] Simulating a counterfactual to show that a similarly situated individual without the protected characteristic would have received more favourable treatment – showing that an otherwise identical male applicant would have received a higher score from the Amazon algorithm, for example – may not be sufficient.[124] A *prima facie* case is not established merely by demonstrating that a claimant was treated less favourably than an actual or

[118] See for example ongoing litigation against Uber, n 69 above.

[119] Racial Equality Directive, art 8(1); Gender Equality Directive, art 19(1); Employment Directive, art 10(1); Equality Act 2010, s 136. See also *Firma Feryn* n 77 above at [32]: it is for the decision-maker to 'adduce evidence that it has not breached the principle of equal treatment'. The two-stage approach was recently affirmed by the Supreme Court for direct discrimination cases: *Efobi* v *Royal Mail Group Ltd* [2021] UKSC 33, [2021] 1 WLR 3863.

[120] Note the caveats at n 7 above.

[121] Kelly-Lyth, n 8 above, 901; see also R. Allen and D. Masters, *Regulating for an Equal AI: A New Role for Equality Bodies* (Brussels: Equinet, 2020) 30-31.

[122] A. Bradford, *The Brussels Effect: How the European Union Rules the World* (Oxford: OUP, 2020).

[123] See *IDS Employment Law Handbooks: Vol 4 – Discrimination at Work* (London: Sweet & Maxwell, 2022) para 33.2: 'complaints of direct discrimination … [are] the area in which the burden of proof tends to be the most hotly contested and, incidentally, where the most confusion reigns'. EU law provides for national autonomy on the standard required to prove a *prima facie* case, and the UK framework is particularly complex (on which see *ibid*, para 33.29).

[124] On counterfactual explanations for ADMS, see S. Wachter, B. Mittelstadt and C. Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31 *Harvard Journal of Law & Technology* 842.

hypothetical comparator: something 'more' is needed.[125] In cases of human implicit bias, that 'something more' can be statistical evidence suggesting consistent implicit bias.[126] If it 'establish[es] a discernible pattern in the treatment of a particular group … it may give rise to an inference of discrimination against the group', such that it is 'reasonable to infer that the complainant, as a member of the group, has himself been treated less favourably' on grounds of a protected characteristic.[127] Further complications arise from a combination of technical complexity and algorithmic opacity, leading to potential difficulties both in terms of analytical capacity and evidence gathering.[128]

These observations notwithstanding, where an analogy can be drawn between human implicit bias and learned algorithmic bias,[129] it seems likely that a claimant will successfully establish a *prime facie* case of ADMS-driven direct discrimination if they can show (i) that the algorithm gives a more favourable output for a similarly situated individual without the claimant's protected characteristic (for example an identical job application with indicators of masculinity, rather than femininity); and (ii) that the algorithm's outputs are consistently skewed in favour of a particular protected characteristic (such that, for example, women on average receive lower scores than men).[130]

A potential finding of direct discrimination thus significantly raises the stakes for decision-makers seeking to deploy potentially biased 'black box' ADMS. If a claimant establishes a *prima facie* case of direct discrimination, the respondent generally cannot rely on the algorithmic scores' apparently high predictive value

[125] *Madarassy* v *Nomura International plc* [2007] EWCA Civ 33, [2007] IRLR 246 at [51], [54] and [56].

[126] *Home Office (UK Visas and Immigration)* v *Kuranchie* EAT [2017] UKEAT/0202/16/BA.

[127] *West Midlands Passenger Transport Executive* v *Singh* [1988] 1 WLR 730, 735E: '[i]f there is evidence of a high percentage rate of failure to achieve promotion at particular levels by members of a particular racial group, this may indicate that the real reason for refusal is a conscious or unconscious racial attitude which involves stereotyped assumptions about members of that group'. This was particularly so in cases of potential subjective bias, 736B. See also *Rihal* v *Ealing LBC* [2004] EWCA Civ 623, [2004] IRLR 642 (*Rihal*), in which the Court of Appeal held that an employment tribunal was entitled to draw an inference of racial stereotyping where a less experienced white employee was promoted ahead of the claimant and there was a lack of diversity within the respondent's senior management.

[128] On recent efforts towards capacity building for judicial systems dealing with AI, see 'Artificial Intelligence and The Rule of Law: Course Overview' The National Judicial College at https://perma.cc/CK2M-AA2Y. On the barriers to evidence-gathering necessary to raise a *prima facie* case of direct discrimination, see Kelly-Lyth, n 8 above, 918-921. Recent legislative proposals at the EU level present an opportunity to address challenges of explainability and evidential access: see Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act) and amending certain Union legislative acts COM/2021/206 final; Proposal for a Directive of the European Parliament and of the Council on Improving working conditions in platform work COM/2021/762 final. Neither proposed instrument currently provides for transparency around equality metrics which would be required to ground a *prima facie* case of direct discrimination, but inspiration could be drawn in developing the proposed measures from US developments including the proposed Algorithmic Accountability Act of 2022 (HR 6580) § 5(1)(E)(Iii) and A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools (Local Laws of the City of New York 2021 No 144).

[129] We return to this discussion in the section below headed 'The complexity of traceability'.

[130] It may, however, be near-impossible for the claimant to obtain this evidence in practice: see Kelly-Lyth, n 8 above, 918-921.

in order to justify the use.[131] Instead, they will have to show that the decision was *not because of a protected characteristic* – which might be near-impossible if the workings of the algorithm are inscrutable.[132]

MACHINE BIAS' CHALLENGE TO RECEIVED APPROACHES

Discussion thus far has focussed on the categories developed in current jurisprudence considering the scope of direct discrimination. The courts have developed a complex taxonomy, despite the fact that the statutory language of direct discrimination is 'comparatively simple: it is treating one person less favourably than you would treat another person, *because of* a particular protected characteristic that the former has.'[133] We have shown that existing interpretations of this phrase cover many cases of algorithmic discrimination, but these interpretations were developed in a context in which only humans discriminate, and where traceability of the factors impacting a decision is therefore limited. In the algorithmic context, on the other hand, it is – at least in principle – more straightforward to identify the factors which give rise to bias: even if an ADMS operates as a black box, key elements such as labelling choices or the composition of a training data set can be recorded and evaluated. The question, then, is whether the courts' approach to categorising direct discrimination *adequately* deals with the mechanisms through which protected characteristics shape algorithmic outcomes.

Broadening the scope of direct discrimination

Are old interpretations sufficient? In the section above headed, 'The default assumption: indirectly discriminatory algorithms', we outlined two cases which – on a principled reading of the legislation – should constitute direct discrimination: Amazon's biased hiring algorithm, and the Gender Shades investigation of bias in facial recognition systems. As demonstrated in the previous section, the Amazon case fits comfortably within the courts' current interpretation of direct discrimination. Tackling Gender Shades under the existing framework, on the other hand, potentially raises a number of additional hurdles.

It is difficult to argue that biased facial recognition tools apply any inherently discriminatory criterion. The 'inherently discriminatory criterion' cases usually follow an 'if/then' approach, framed around a particular policy: if not of pensionable age, then no free entry to swimming pool;[134] if mother not born Jewish, then no entry to school;[135] if wearing a large and conspicuous sign of

[131] Note the caveats at n 7 above.
[132] Allen and Masters, n 5 above at [139]-[143].
[133] *Essop* n 56 above at [1]. Note that this is a simplification of 'on grounds of', which still appears in EU legislation; there was no intention to change the meaning: Explanatory Notes to the Equality Act 2010, para 61.
[134] *James* n 58 above.
[135] *JFS* n 50 above.

religious belief, then disciplinary action taken by employer;[136] if a graduate of a women's college, then assign a lower score.[137] The 'if' part of each rule is a proxy for a protected characteristic, and the 'then' part is the relevant disadvantage. It is difficult to conceptualise biased facial recognition tools in this way. The closest framing would be: 'if a darker-skinned face, then less likely to be recognised'.[138] The case *could* be shoe-horned into the 'inherently discriminatory criterion' box in this way, but it is not a comfortable fit.[139] The operation of a facial recognition model does not involve the kind of explicit 'if-then' statement described above, but rather places faces within high-dimensional mathematical spaces which don't directly correspond to human concepts like 'dark skin', let alone socially constructed categories like 'race': the framing describes an outcome, not a policy.[140]

Nor is the algorithm's poor performance when recognising black women an example of implicit bias (that is, akin to the automation of biased mental processes). It would therefore be difficult to analyse even within an expanded conceptualisation of subjective discrimination. The problematic bias is caused by the imbalanced demographics of the training data, and would likely still occur for a minority group even if the training data were representative of the overall population.

The existing 'two categories' approach to direct discrimination thus fails to capture the Gender Shades case. If, as we argued above in 'A normative concern', it fits within the normative bounds of the concept, how do we deal with this novel situation in legal terms? By returning to 'drink from the pure waters of the statute'.[141] Although the court in *JFS* sought to simplify the law, its approach was limited to human discrimination. A rethinking of the definition is required in order to map the statutory language onto computer science approaches for explaining algorithmic outputs.[142]

Judicial interpretations thus far have excluded certain computer science approaches. For example, the courts have rejected a general 'but for' test for direct discrimination, so direct discrimination does not necessarily arise where the protected characteristic is merely part of the circumstances surrounding the less favourable treatment, but not an activating 'cause'.[143] For this reason, it is

[136] *WABE* n 53 above.

[137] As with the Amazon algorithm, see Dastin, n 2 above.

[138] Simons and others, n 15 above. It is not necessary for every member of the protected group to suffer the actual detriment; 'disadvantage' can occur if one group is given fewer opportunities to attain something than another, for example, even if some individuals in the former group do nonetheless attain that thing: *Coll* n 84 above at [30].

[139] See, however, Allen and Masters, n 17 above, 592.

[140] See, by analogy, the Court of Appeal in *Taiwo* v *Olaigbe* [2014] EWCA Civ 279, [2014] 1 WLR 3636 (*Taiwo*), rejecting the criterial approach where the claimant employees' status as migrant domestic workers had affected their employers' treatment of them.

[141] *Owens* v *Wealden District Council* [2005] UKEAT/0186/05/CK at [17]; *Windle* v *Arada & Anor* [2014] UKEAT/0339/13/RN at [2].

[142] The judicial interpretation of 'because of' must therefore evolve: D. Feldman, D. Bailey and L. Norbury (eds), *Bennion, Bailey and Norbury on Statutory Interpretation* (London: LexisNexis Butterworths, 2020) section 14.1, previously approved in *Moorthy* v *HMRC* [2018] EWCA Civ 847, [2018] 3 All ER 1062 at [58]. See also *Taiwo* n 140 above at [40].

[143] *Amnesty International* v *Ahmed* [2009] UKEAT/0047/08 [37]; *Martin* v *Lancehawk Ltd (t/a European Telecom Solutions)* [2004] UKEAT/0525/03 (*Martin*). In *Martin*, for example, the EAT held

probably inadequate simply to simulate counterfactual situations in which the protected characteristic is changed, for example. Despite setting these limits, the courts have expressly avoided creating a generalisable definition of 'because of'. We return to the challenges this has created, below. For present purposes, Lord Nicholls' suggestion in *Nagarajan* that we should ask whether the protected characteristic had a '*significant influence*' on the outcome provides a helpful starting point for identifying direct discrimination in the algorithmic context.[144]

The 'significant influence' enquiry is closely mirrored in computer science discussions on how best to identify the salient features in a model which contribute towards a particular output.[145] SHAP methods, for example, borrow from game theory to measure the contribution of different features used in an algorithmic decision as if they are players in a coalition in a game.[146] A SHAP explanation of the Amazon hiring algorithm would list the features used (for example previous employment, SAT scores, attendance at an all-women's college) and estimate how important they were to the overall score for that candidate. Similarly, a LIME-based explanation might highlight the portions of text which positively or negatively correlate with a hiring decision.[147]

Whether protected characteristics had a 'significant influence' on a decision thus becomes an empirical question. Both methods have their limitations and are the subject of ongoing research and debate within computer science; they nonetheless illustrate the kinds of computational approaches that might be needed to ground claims under the proposed new category of direct algorithmic discrimination. Sophisticated statistical methods may furthermore not even be necessary in all cases, including scenarios where the impact of a protected characteristic on the performance of a model is clearly demonstrable in practice. In the wake of the Gender Shades study, for example, commercial facial recognition companies made efforts to collect more racially diverse training sets, which effectively eliminated the disparity in treatment.[148] In such cases, it seems intuitive to attribute the cause of the initial disparity to race and gender.

that the dismissal of a female employee by her male boss after a sexual relationship between them ended was not direct sex discrimination. The affair, and therefore the dismissal, would not have occurred but for her sex, but her sex was not the 'cause' of the dismissal.

[144] *Nagarajan* n 56 above, 512H-513B per Lord Nicholls. Emphasis supplied.

[145] C. Molnar, C. Giuseppe and B. Bernd, 'Interpretable machine learning–a brief history, state-of-the-art and challenges' in I. Koprinska and others (eds), *ECML PKDD 2020 Workshops* (Cham: Springer, 2020).

[146] See for example W. Kruskal, 'Relative importance by averaging over orderings' (1987) 41 *The American Statistician* 6; S.M. Lundberg and L. Su-In, 'A unified approach to interpreting model predictions' (2017) 30 *Advances in neural information processing systems*.

[147] M.T. Ribeiro, S. Singh and C. Guestrin, '"Why should I trust you?" Explaining the predictions of any classifier' (2016) *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* 1135. While such methods do purport to measure the contributions of different features, they can be unstable, misleading, and prone to misinterpretation, thus requiring an exploration of further methods to support the legal analysis of whether a feature has 'significant influence'. D. Alvarez-Melis and T.S. Jaakkola, 'On the robustness of interpretability methods' (2018) *arXiv preprint* arXiv:1806.08049; D. Slack, S. Hilgard, E. Jia, S. Singh and H. Lakkaraju. 'Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods' (2020) *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 180.

[148] Raji and Buolamwini, n 66 above.

A stricter standard for ADMS?

By returning to the statutory phrase, and thereby going beyond the two established categories of direct discrimination, are we creating a stricter standard for algorithms than humans? In theory, no: we are just recognising that the categories developed to deal with the opacity of the human mind are not applicable in an algorithmic context.[149] Although the factors behind biased outcomes may be more traceable, the standards should be similarly strict.

Thus far, the courts have dealt with the opacity of the human mind by distinguishing between two types of cases: (i) the 'obvious cases', where the 'criterion applied [is] not in doubt'; and (ii) the 'other cases', where the protected characteristic features (consciously or unconsciously) in the decision-maker's mental processes.[150] Cases of type (i) are dealt with under the 'inherent discrimination' framework; cases of type (ii) are considered subjective discrimination.

Like the human mind in case type (ii), a machine learning algorithm can act as a black box: the specific criteria applied by the algorithm may not be identifiable.[151] However, *unlike* human bias, the *causes* of algorithmic bias may well be cognisable. In the Gender Shades case, for example, one might say that a black woman was misidentified *because* she was a black woman. However, the traceability of that bias also means that we could alternatively analyse the situation as misidentification *because* few black women were included in the data used to train the algorithm. In the human context, the latter reason would hold no (legal) water: the discrimination framework does not, in theory, permit rationalisation of implicit biases.[152] A manager who refuses to allow a black female employee into the workplace because he is 'not as good' at recognising black women would presumably be held to be directly discriminating against that employee.[153]

The complexity of traceability

In practice, however, biased human decision-makers can engage in an *ex post facto* rationalisation of decisions.[154] Indeed, whilst implicit bias has been classed as unjustifiable direct discrimination since the 1980s,[155] it remains rife to this

[149] The approaches were developed to deal with the two common mechanisms by which human discrimination occurs, but the ultimate question is always what the ground of the treatment complained of was: *Ahmed* v *Amnesty International* [2009] UKEAT/0447/08 at [32].

[150] *JFS* n 50 above at [64].

[151] W. Samek and others (eds), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Cham: SpringerLink, 2019).

[152] This is despite the fact that humans are better at recognising faces of people with whom they share an ethnicity, an effect which can be reduced by exposure to faces of people with other ethnicities in infancy: M. Heron-Delaney and others, 'Perceptual Training Prevents the Emergence of the Other Race Effect during Infancy' (2011) 6 *PLOS ONE* e19858.

[153] Compare Uber's use of a facial recognition system, which drivers must use at the start of their shifts: Vallance, n 4 above.

[154] See discussion in A. Carolina, A. Vieira and A. Graser, 'Taming the Biased Black Box? On the Potential Role of Behavioural Realism in Anti-Discrimination Policy' (2015) 35 OJLS 121.

[155] *Nagarajan* n 56 above.

day: in Europe, a similarly qualified job applicant with an immigrant background will have to send 'about thirty percent more applications than majority applicants to have a similar likelihood of receiving a positive response from employers'.[156] Compelling macro evidence of this pervasive (yet theoretically unlawful) bias can be contrasted with the striking scarcity of judicial findings of implicitly biased decision-making in individual cases.[157] The problem, as the Court of Appeal has long recognised, is that it is 'unusual to find direct evidence of racial discrimination' because '[f]ew [decision-makers] will be prepared to admit such discrimination even to themselves'.[158] By identifying another 'reason' for his decision, the purported discriminator may discharge the burden of disproving that it was not taken on discriminatory grounds.

Automated decisions are more examinable, thus making existing law more enforceable.[159] Traceability, however, is a dangerous blessing. In one sense, it is an opportunity: unlike humans, ADMS can provide a quantifiable account of the role played by a protected characteristic. On the other hand, it forces us to contend with a fundamental, thorny question: how close must the connection between the protected characteristic and the outcome be for the outcome to be 'because of' the protected characteristic? In circumstances where machine learning algorithms are trained on real world data, protected characteristics will frequently have played *some* upstream role. In other words, an approach which assesses all upstream factors to ask whether an outcome is 'because of' a protected characteristic could quickly become over-determinative, identifying a broad range of factors which arise at multiple stages of the process.

We noted above that the courts have avoided adopting a generalisable definition of 'because of'. In *Nagarajan*, Lord Nicholls stated that '[n]o one phrase is obviously preferable to all others … in the application of this legislation legalistic phrases, as well as subtle distinctions, are better avoided so far as possible.'[160] He later suggested that while 'causation is a legal conclusion', the 'reason why a person acted as he did' in the context of a direct discrimination claim 'is a question of fact'.[161] The courts' refusal to adopt a 'legalistic' definition of causation in the discrimination context has led to problematic ambiguity.[162] The

[156] N.V. Demireva and M. Mangiantin (eds), *GEMM Project in Focus 2015-2018* (GEMM Project 2018) 20-21.

[157] For some rare exceptions, see *North West Thames Regional Health Authority* v *Noone* [1988] IRLR 530; *Rihal* n 127 above; *Geller* v *Yeshurun Hebrew Congregation* [2016] UKEAT/0190/15/JOJ (finding that the tribunal had failed to properly consider the possibility of subconscious discrimination); and the Irish case of *Micheline Sheehy Skeffington* v *National University of Ireland, Galway* [2014] DEC-E2014-078.

[158] *King* n 113 above at [36].

[159] We recognise that a challenge arises where the human alternative is more discriminatory than the algorithmic decision-making, but the law is more enforceable against the algorithm: see Kelly-Lyth, n 8 above, 908. This practical challenge is nonetheless the consequence of applying the law *equally* to both systems.

[160] *Nagarajan* n 56 above, 513A.

[161] *Chief Constable of West Yorkshire Police* v *Khan* [2001] UKHL 48, [2001] 1 WLR 1947 at [29].

[162] Compare, for example, *Seide* v *Gillette Industries* [1980] IRLR 427; *Rees* v *Apollo Watch Repairs Plc* [1996] ICR 466 at [13], on which cf *Indigo Design Build and Management Ltd* v *Martinez* [2014] UKEAT/0020/14/DM; *O'Neill* v *Governors of St Thomas More Roman Catholic Voluntary Aided Upper School* [1996] IRLR 372, stating the causation is 'not simply a matter of a factual, scientific or historical explanation of a sequence of events', but that the question is the 'real

statutory language of direct discrimination law is ambitious: in certain areas, less favourable treatment cannot be accorded 'because of a protected characteristic'. This ambition, however, is not always reflected in the caselaw.

Take the recent Supreme Court decision in *Lee* v *Ashers*.[163] In that case, a bakery refused to provide a cake with the words 'support gay marriage' iced on top because the owners 'oppose[d] same sex marriage'.[164] Intuitively, it seems that this refusal of service was, at least in a general sense, because of (the bakery owners' views on) sexual orientation. But the Supreme Court held that the direct discrimination framework did not apply. The same treatment would have been accorded to a heterosexual customer seeking the same cake;[165] support for gay marriage is not a perfect proxy for sexual orientation;[166] and associative discrimination does not cover all cases in which 'the reason for the less favourable treatment has something to do with the sexual orientation of some people' – there must be 'a closer connection than that'.[167] The Court arrived at this conclusion while also noting that it would be 'unwise … to attempt to define the closeness of the association which justifies' a finding of direct discrimination.[168]

We see a similar unresolved tension when comparing the ambition of the legislative language to the treatment of decisions which are, in some sense, 'because of' a protected characteristic, but where the decision-maker is not *aware* of that causal relationship.[169] For example, if a decision-maker receives 'tainted information' from an external source – such as a job reference written by a biased former employer, or a poor mark awarded by a discriminatory examiner – then no liability arises if that information is taken at face value, resulting in less favourable treatment.[170] It is difficult to square this position with *James*. In both

and effective cause' or 'effective and predominant cause' of the treatment; *Titterington* v *Ensor* [2003] UKEAT/0052/02/3110 at [15]-[16], finding 'much force' in the submission that direct discrimination occurs if a protected characteristic is a 'cause in the overall pattern', and that the question is whether 'as part of the matrix, sex or race or both could have been a cause of the less favourable treatment'; *Taiwo* n 140 above at [40]-[57]; and *Hall* v *Chief Constable of West Yorkshire* [2015] IRLR 893 on the Equality Act 2010, s 15 (and note also that the existence of s 15 itself indicates a limit to causation under s 13). For similar concerns in the US context, see M.J. Katz, 'The Fundamental Incoherence of Title VII: Making Sense of Causation in Disparate Treatment Law' (2005–2006) 94 *Georgetown Law Journal* 489, 493, suggesting that an inability to define the meaning of 'because of' in legal terms has 'hampered the normative debate over the appropriate standard of causation in disparate treatment law'.

[163] *Lee* v *Ashers* n 74 above.

[164] *ibid* at [22].

[165] *ibid* at [23].

[166] *ibid* at [25].

[167] *ibid* at [33].

[168] *ibid* at [34]. The absence of a justificatory regime has arguably led to increased strictness in defining the scope of causation. On justification and *Lee* v *Ashers*, see H. Collins, 'A missing layer of the cake with the controversial icing' *UK Labour Law Blog* 4 March 2019 at https://perma.cc/VL5D-WTGL.

[169] See for example the cases in which an absence of knowledge has been a successful defence to claims of direct discrimination: *Crouch* v *Mills Group Ltd and anor* ET Case No.1804817/06; *McClintock* v *Department for Constitutional Affairs* [2008] IRLR 29; *Patel* v *Lloyds Pharmacy Ltd* [2013] UKEAT/0418/12/ZT; cf *Urso* v *Department for Work and Pensions* [2016] UKEAT/0045/16/DA. It seems likely that such arguments would not succeed in 'criterial' cases; for example, if a respondent applied the 'pensionable age' criterion in *James* without being aware of its inherently discriminatory nature: see n 82 above.

[170] *CLFIS (UK) Ltd* v *Reynolds* [2015] EWCA Civ 439, [2015] IRLR 562.

cases, the decision-maker uses information supplied by a third party. In *James*, the information was the category of people who might require free swimming pool entry given their likely employment status. In the biased reference case, it is whether an applicant will perform well in a job. In both instances, the classification – whether the person is of pensionable age; whether the applicant would make a good employee – is different because of the claimant's protected characteristic. The clearest distinguishing feature between the cases appears to be the decision-maker's *knowledge* of the relationship between the information and the protected characteristic. Indeed, this much is clear from the Court of Appeal's discussion of a decision-maker unknowingly using 'tainted information' to make a decision: 'We are… concerned … [with a case in which] an act which is detrimental to a claimant is done by an employee who is innocent of any discriminatory motivation but who has been influenced by information … supplied … [by someone who is] discriminatory … It would be quite unjust for [the individual decision-maker] … to be liable … where he was personally innocent of any discriminatory motivation.'[171] Yet Underhill LJ's consideration of 'innocence' appears to run contrary to the courts' assertions that questions of morality are irrelevant to the scope of direct discrimination.[172]

Analysed thus, ADMS challenge us to (re-)open fundamental questions about the appropriate scope of discrimination law. First, how far upstream must a protected characteristic be before it loses its causal relationship to the treatment, for the purposes of direct discrimination? For example, there is now significant empirical evidence that women are less frequently shown Facebook advertisements for certain well-paid jobs when compared with similarly qualified men.[173] Is this less favourable treatment because of sex, or because the relevant positions are more frequently held by men?[174]

Second, even if we do accept that less favourable algorithmic treatment has occurred because of a protected characteristic, are courts institutionally best

[171] *ibid* at [34], [36] per Underhill LJ.

[172] See n 62 above. See also Khaitan, n 56 above, 184, suggesting in the context of inherently discriminatory criteria that 'it is likely that direct discrimination will usually entail at least a reckless disregard for the risk that the victim could suffer because of her group membership … Difficulties arise in those rare cases where there is no blame whatsoever – because these cases exist … It is for this reason that, subject to regulatory concerns, it may be wise for the law to allow some possibility of justifying direct discrimination.'

[173] B. Imana, A. Korolova and J. Heidemann, 'Auditing for Discrimination in Algorithms Delivering Job Ads' (2021) *Proceedings of the Web Conference 2021 (WWW '21)* 3767; 'How Facebook's Ad Targeting May Be in Breach of UK Equality and Data Protection Laws' *Global Witness* 9 September 2021 at https://perma.cc/Y8BG-57GH; on the difficulties of identifying whether an advert has been shown to someone because of their protected characteristic or because of their interest in activities which are linked to individuals with that characteristic, see Wachter, n 18 above, 400-401. Although advertisements are excluded from the Gender Access Directive, employment advertisements would be covered because the Recast Directive operates as a *lex specialis*: see Wachter, *ibid*, 391.

[174] Imana, Korolova and Heidemann, *ibid*, 3774, found that jobs requiring similar qualifications reached different audiences depending on the gender balance of the company. For example, the authors found that Facebook distribution of an advertisement for a delivery driver at Domino's Pizza was skewed male, whereas the distribution of an advertisement for a delivery driver at Instacart was skewed female. The qualification requirements for the jobs were the same, but while Domino's Pizza had a delivery driver workforce that was 98 per cent male, Instacart's delivery drivers were more than 50 per cent female.

suited to protect victims against that treatment? Or are black-box algorithmic outcomes merely 'tainted information', the 'innocent' use of which cannot constitute direct discrimination?

The problem is that direct discrimination law focuses on a 'discrete moment … of decision-making'[175] – but it is difficult to establish whether discrete decisions are *because of* factors like gender or race when *all* decisions are made in the context of a fundamentally unequal world.[176] When ADMS are being used, this difficulty is particularly sharp, as machine learning algorithms are trained on data drawn from that unequal world.

CONCLUSION

The widespread deployment of automated decision-making systems is increasingly facing legal challenges under UK and EU discrimination law. Such challenges, we suggest, should frequently be framed as directly discriminatory algorithms: even on a conservative reading of the current law, many more instances of automated bias fall within the scope of the near-absolute prohibition of differential treatment *because of* a protected characteristic than has previously been assumed.

Legal debates in mainstream algorithmic fairness research have primarily been rooted in US discrimination law, and its categories of disparate treatment and impact. Translated into the European context, this has led to a focus on indirect discrimination, rather than direct discrimination, as the primary lens through which to challenge biased ADMS. While indirect discrimination law occupies an important place in the claimant's toolkit, there are practical and normative reasons to reject a default assumption that it is the *only* tool available in many cases of algorithmic bias. Challenging widely accepted assumptions in the existing literature, we have shown that in legal terms, some paradigmatic cases of ADMS bias do fit comfortably within the definition of direct discrimination provided by the EU and UK courts. This poses particular problems for biased black-box ADMS: it will be near-impossible for regulated decision-makers to justify the use of such systems.

Focusing on indirect discrimination alone would result in a narrowing of protection against discrimination in Europe: automated implicitly biased decisions, or decisions based on inherently discriminatory criteria, might escape the net of unjustifiable direct discrimination, which has been woven over the course of decades. As discrimination evolves, so must the law, reinterpreting the statutory language to ensure that it captures algorithmic direct discrimination which fits within both the statutory language and established normative bounds, but does not fit within judicially defined categories. The traceability of ADMS opens up new possibilities in establishing when protected

[175] R. Abebe and others, 'Roles for Computing in Social Change', (2020) *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT∗ '20) 252.

[176] See D. Hellman, 'Big Data and Compounding Injustice' (2022) *Journal of Moral Philosophy* (forthcoming).

characteristics have had 'significant influence' in discriminatory decisions, effectively rendering them unjustifiable.

Algorithmic fairness research has thus laid bare the limits of the law, both in terms of existing categories and broader questions as to the fundamental structures of discrimination law. The scope of the phrase 'because of' constitutes the very distinction between direct and indirect discrimination – and thus the boundaries of legal conduct. At the same time, however, ADMS 'cannot be treated as individual models, but must be assessed as a linear series of interdependent models'. The biases under scrutiny will not only 'exist as a function of data or model choice, but in the epistemological roots of the system.'[177]

As Abebe et al have noted,

> Much computational research on fairness is built on frameworks borrowed from discrimination law … perhaps most crucially, the belief that fairness can be achieved by simply altering how we assess people at discrete moments of decision-making (for example hiring, lending, etc). At best, discrimination law is an incomplete mechanism to remedy historic injustices and deeply entrenched structures of oppression … exposing the limits of algorithmic notions of fairness has exposed the limits of the underlying legal and philosophical notions of discrimination on which this work has built.[178]

This is as true in Europe as it is in the US: direct discrimination protects against more than intentional discrimination, but its scope is far from clear.[179] If discrimination law is an inadequate mechanism to address deeply entrenched structures of oppression, how far *should* it try to go? How proximate must a protected characteristic be to render a decision unjustifiably discriminatory? Empirical evidence of bias in automated decision-making will require us to revisit many of the fundamental challenges to the conceptual apparatus of discrimination law – from academic debates about the discipline's normative foundations to courts' reluctance to approach causation through a 'legalistic' lens in the equality context.

[177] M. Sloane, E. Moss and R. Chowdhury, 'A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for Auditability' (2021) *ACM CHI Virtual Conference on Human Factors in Computing Systems* (CHI '21) 2.

[178] Abebe et al, n 175 above.

[179] On the limitations of US discrimination law and concepts of causation, see T.B. Gillis, 'The Input Fallacy' (2022) 106 *Minnesota Law Review* 1175, 1220-1221, suggesting that while disparate treatment and disparate impact have 'centred on the question of … causal effect', machine learning 'is a world of correlation and not causation'; and proposing a shift to outcome-focused analysis for both disparate treatment and disparate impact.